



UNIVERSIDAD AUTÓNOMA DE  
**CHIHUAHUA**

### TRABAJO DE TESIS

La academia de Sistemas Computacionales en Hardware, de la carrera de Ingeniero en Sistemas Computacionales en Hardware en su sesión celebrada el 13 de agosto de 2020, conoció la solicitud presentada por el pasante C. RAYDESEL ARIEL SÁNCHEZ MONTES, con número de matrículas 310830, a elaborar trabajo de tesis denominado “RECONOCIMIENTO DE EMOCIONES HUMANAS POR VOZ USANDO MFCC Y FIRMAS DE ENTROPÍA ESPECTRAL MULTIBANDA”; y con ello cumplir con el artículo 114 del Reglamento Interno de la Facultad. El trabajo abarcará de manera general los siguientes aspectos:

- Introducción
- Marco Teórico
- Metodología
- Experimentos
- Conclusión
- Referencias

Habiendo revisado el expediente del pasante **Sánchez Montes**, se tomaron los siguientes acuerdos:

1. Se acepta el tema propuesto
2. Se designa como Director del trabajo al C. Catedrático de esta Facultad: **DR. ALAIN MANZO MARTÍNEZ**.
3. El comité revisor estará integrado por los Catedráticos de esta Facultad:

**DR. FERNANDO MARTÍNEZ REYES.**

**DR. LUIS FERNANDO GAXIOLA ORDUÑO.**

4. Se le pide incluir en la primera página del proyecto, copia de este registro.

Atentamente

“naturam subiecit aliis”

Director del trabajo

DR. ALAIN MANZO MARTÍNEZ

Revisor del trabajo

DR. FERNANDO MARTÍNEZ REYES

Revisor del trabajo

DR. LUIS FERNANDO GAXIOLA ORDUÑO

Director de la Facultad



UNIVERSIDAD AUTÓNOMA DE  
CHIHUAHUA

Chihuahua, Chih. 02 de marzo de 2021

**C. Raydesel Ariel Sánchez Montes (Mat. 310830)**

Pasante del programa educativo de Ingeniería en Sistemas Computacionales en Hardware  
Presente.-

De acuerdo al reglamento interno de la Facultad de Ingeniería y habiendo cumplido con todas las indicaciones que la comisión revisora realizó al material con respecto a la elaboración de la tesis denominada "**RECONOCIMIENTO DE EMOCIONES HUMANAS POR VOZ USANDO MFCC Y FIRMAS DE ENTROPÍA ESPECTRAL MULTIBANDA**" dirigida por el Dr. Alain Manzo Martínez, revisada por el Dr. Fernando Martínez Reyes y el Dr. Luis Fernando Gaxiola Orduño; la Facultad de Ingeniería concede autorización para que proceda la impresión de la misma.

Se le pide incluir en la segunda página de la tesis, copia de este registro.

Se extiende la presente para los fines que al interesado le convenga.

**ATENTAMENTE,**  
"naturam subiecit aliis"

**M.I. Ana Lucía Corral Flores**  
Coordinadora los Programas Académicos en  
Ingeniería en Sistemas y Computación

**UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA**  
**FACULTAD DE INGENIERÍA**

---



**RECONOCIMIENTO DE EMOCIONES HUMANAS POR VOZ USANDO  
MFCC Y FIRMAS DE ENTROPÍA ESPECTRAL MULTIBANDA**

**POR:**

**RAYDESEL ARIEL SÁNCHEZ MONTES**

**TESIS**

**PRESENTADA COMO REQUISITO PARA OBTENER EL GRADO**

**DE:**

**INGENIERO EN SISTEMAS COMPUTACIONALES EN HARDWARE**

**COMITÉ**

**Dr. Alain Manzo Martínez**

**Dr. Fernando Martínez Reyes**

**Dr. Luis Fernando Gaxiola Orduño**

# Agradecimientos

La vida es aquello que te va sucediendo mientras te empeñas en hacer otros planes, decía John Lennon. Siempre he pensado que se llega a algún punto por la casualidad de cruzarse con personas que, siendo o no conscientes de ello, influyen extraordinariamente en los caminos que un individuo elige.

A usted Alain, mi director de Tesis, tengo mucho que agradecer, por darme la oportunidad de participar en este trabajo de investigación (gracias por esta experiencia, ¡tremenda!, por enorme y sin duda valiosa para mí). Muchas gracias por sus consejos y por su valiosísima ayuda en la elaboración de este documento. Debo agradecer también la ayuda prestada a los compañeros, que me proporcionaron los medios y los recursos necesarios para elaborar este trabajo. A los amigos, que en algún momento escucharon mis elucubraciones y “soportaron” teorías que no entraban seguramente dentro de sus temas favoritos de debate. A todos aquellos a quienes, seguramente, no preste toda la atención que debía; por estar enfrascado en exceso en estos y otros temas. Esta tesis no habría podido finalizarse sin la paciencia y la colaboración de todas las personas que han arrimado el hombro y animado cuando más lo necesitaba, mostrándome su cariño incondicional. Por último, agradezco especialmente a mi familia, que son conocedores de la evolución de mi vida y del esfuerzo realizado; sin su valioso apoyo, de ningún modo habría podido hacerlo.

# ÍNDICE

CAPITULO 1 INTRODUCCIÓN.....	1
1.1    Introducción.....	1
1.2    Mecanismo de producción del habla humana.....	2
1.3    Clasificación de la señal de voz .....	3
1.4    Desafíos del reconocimiento automático del habla .....	3
1.5    Revisión del estado del arte.....	4
1.5.1 Noise Robust Speaker Identification Using RASTA–MFCC Feature with Quadrilateral Filter Bank Structure. ....	4
1.5.2 Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods.....	4
1.5.3 Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition. .....	5
1.5.4 Speech Emotion Recognition Using Fourier Parameters.....	5
1.5.5 Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo. ....	5
1.5.6 Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models.....	6
1.5.7 Investigation of the Relation between Emotional State and Acoustic Parameters in the Context of Language. ....	6
1.5.8 Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique....	7
1.5.9 Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace.....	7
1.5.10 The Effect of Noise on Emotion Perception in an Unknown Language. ....	7
1.5.11 A novel feature selection method for speech emotion recognition. ....	7
1.5.12 Wavelet packet analysis for speaker-independent emotion recognition. ....	8
1.6    Planteamiento del problema .....	10
1.7    Hipótesis.....	11
1.8    Objetivos .....	12
1.8.1 Objetivo general.....	12
1.8.2 Objetivos particulares .....	12
1.9    Justificación.....	13
CAPITULO 2 MARCO TEÓRICO .....	14
2.1    Introducción.....	14
2.2    Arquitectura del sistema de reconocimiento automático del habla. ....	15
2.3    Emociones.....	16
2.4    Base de datos .....	18

2.4.1 Berlin Database of Emotional Speech (EMODB) .....	21
2.4.2 EMOVO una base de datos de habla emocional italiana .....	22
2.5 Preprocesamiento.....	23
2.5.1 Eliminación de ruido ambiental o de fondo:.....	23
2.6 Extracción de características.....	26
2.6.1 Coeficientes cepstrales de frecuencia Mel .....	27
2.6.2 Entropía espectral y entropía de Shannon .....	29
2.6.3 Firma de entropía espectral multibanda .....	29
2.6.4 Transformada discreta de Fourier .....	31
2.6.5 RASTA-MFCC .....	31
2.7 Tipos de clasificadores .....	34
2.7.1 MLP .....	35
2.7.2 KNN .....	36
2.7.3 SVM .....	37
2.7.4 Métricas .....	38
CAPITULO 3 METODOLOGIA .....	40
3.1 Análisis y división de datos .....	40
3.1.1 Inspección auditiva .....	40
3.1.2 Inspección visual .....	40
3.1.3 Propiedades de los archivos de audio.....	42
3.1.4 Distribuciones de clases .....	43
3.1.5 Dividir el conjunto de datos .....	44
3.2 Preprocesamiento.....	45
3.2.1 Stereo.....	45
3.2.2 Preemphasis.....	46
3.2.3 Framing .....	46
3.2.4 Autocorrelación .....	47
3.2.5 Windowing .....	48
3.3 Extracción de firmas espectrales .....	49
3.3.1 MFCC.....	50
3.3.2 RASTA-MFCC .....	51
3.3.3 Entropy signature.....	52
3.3.4 MSES .....	53
3.3.5 Extracción de características para cada archivo .....	54
3.3.6 Convertir los datos y las etiquetas .....	55
3.4 Clasificadores .....	56

3.4.1 MLP .....	56
3.4.2 SVM .....	60
3.4.3 KNN .....	63
CAPITULO 4 EXPERIMENTOS.....	65
4.1     Experimentos con autocorrelación de la base de datos EMODB .....	65
4.2     Experimentos con autocorrelación de la base de datos EMOVO .....	74
4.3     Mejores resultados de los experimentos con autocorrelación .....	83
4.4     Experimentos sin autocorrelación de la base de datos EMODB.....	87
4.5     Experimentos sin autocorrelación de la base de datos EMOVO .....	96
4.6     Mejores resultados de los experimentos sin autocorrelación.....	105
4.7     Mejor resultado EMODB.....	109
4.8     Mejor resultado EMOVO.....	109
4.9     Comparación con los mejores resultados en el estado del arte.....	110
CAPITULO 5 CONCLUSIÓN.....	111
5.1     Conclusión.....	111
REFERENCIAS.....	113

## ÍNDICE DE FIGURAS

Figura 1.1: Modelo de producción del habla humana.....	2
Figura 1.2: Una visión general de los sistemas de reconocimiento de emociones del habla.....	10
Figura 2.1: Arquitectura del sistema de reconocimiento automático del habla .....	15
Figura 2.2: Modelo tridimensional continuo de las emociones. Valencia - Activación – Dominación.	16
Figura 2.3: Etapas de preprocesamiento .....	23
Figura 2.4: Efectos del preprocesamiento de la señal de voz.....	25
Figura 2.5: Estructura del banco de filtros triangulares .....	28
Figura 2.6: Estructura del banco de filtros Gaussianos.....	28
Figura 2.7: Diagrama a bloques para extraer MFCC, RASTA-MFCC y ARMA-MFCC.....	32
Figura 2.8: Diagrama a bloques de RASTA .....	33
Figura 2.9: Activación de un perceptrón o neurona artificial .....	35
Figura 2.10: Red neuronal artificial de tipo perceptrón multicapa .....	36
Figura 2.11: Ruido en clasificación.....	37
Figura 2.12: Un Hiperplano de clasificación {w →, b} para un conjunto de entrenamiento de dos dimensiones .....	37
Figura 3.1: Forma de onda del archivo 03a04Ad.wav de la clase angustia de la base de datos EMODB .....	41
Figura 3.2: Forma de onda del archivo 14a04Ed.wav de la clase disgusto de la base de datos EMODB .....	41
Figura 3.3: Forma de onda del archivo 03a04Lc.wav de la clase aburrimiento de la base de datos EMODB .....	41
Figura 3.4: Forma de onda del archivo 03a04Fd.wav de la clase felicidad de la base de datos EMODB .....	41
Figura 3.5: Forma de onda del archivo 03a04Nc.wav de la clase neutral de la base de datos EMODB .....	41
Figura 3.6: Forma de onda del archivo 03a04Ta.wav de la clase tristeza de la base de datos EMODB .....	41
Figura 3.7: Forma de onda del archivo 03a04Wc.wav de la clase irá de la base de datos EMODB ....	41
Figura 3.8: Distribución de clases de EMODB .....	43
Figura 3.9: Distribución de clases de EMOVO.....	44
Figura 3.10: Metadatos de EMODB .....	44
Figura 3.11: Diagrama a bloques del preprocesamiento de la señal del habla .....	45
Figura 3.12: Grafica de la señal de audio stereo gio-f1-b1.wav .....	45
Figura 3.13: Grafica de la señal de audio monoaural gio-f1-b1.wav .....	45
Figura 3.14: Grafica de la aplicación del filtro de premphasis a la señal de audio gio-f1-b1.wav.....	46
Figura 3.15: Framing, 50% traslape entre frames de 30ms .....	46

Figura 3.16: Aplicación de la función de autocorrelación en frames de la señal de voz para descartar sonidos no vocalizados .....	47
Figura 3.17: Grafica de frame multiplicado por la ventana de Hann.....	48
Figura 3.18: Grafica de la ventana de Hann.....	48
Figura 3.19: Diagrama a bloques con los pasos a seguir para obtener las características MFCC, RASTA-MFCC, Entropy signature y MSSES.....	49
Figura 3.20: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica MFCC, con y sin función de autocorrelación.....	50
Figura 3.21: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica RASTA-MFCC, con y sin función de autocorrelación.....	51
Figura 3.22: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica Entropy Signature, con y sin función de autocorrelación.....	52
Figura 3.23: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica MSSES, con y sin función de autocorrelación.....	53

## ÍNDICE DE TABLAS

Tabla 1.1: Revisión del estado del arte .....	9
Tabla 2.1: Principales avances en la tecnología de reconocimiento del habla.....	14
Tabla 2.2: Conjuntos de emociones básicas propuestos por diferentes autores.....	17
Tabla 2.3: Bases de datos para el reconocimiento de emociones en el habla .....	21
Tabla 2.4: Comparación de técnicas para extracción de características .....	27
Tabla 2.5: Bandas críticas para la escala Bark.....	30
Tabla 2.6: Comparación de varias técnicas de clasificación .....	35
Tabla 2.7: Evaluación de los clasificadores .....	38
Tabla 4.1: Resultados de recall usando el clasificador MLP.....	65
Tabla 4.2: Resultados de recall usando el clasificador KNN.....	66
Tabla 4.3: Resultados de recall usando el clasificador SVM .....	66
Tabla 4.4: Resultados de recall usando el clasificador MLP.....	67
Tabla 4.5: Resultados de recall usando el clasificador KNN.....	67
Tabla 4.6: Resultados de recall usando el clasificador SVM .....	68
Tabla 4.7: Resultados de recall usando el clasificador MLP.....	68
Tabla 4.8: Resultados de recall usando el clasificador KNN.....	69
Tabla 4.9: Resultados de recall usando el clasificador SVM .....	70
Tabla 4.10: Resultados de recall usando el clasificador MLP.....	71
Tabla 4.11: Resultados de recall usando el clasificador KNN .....	72
Tabla 4.12: Resultados de recall usando el clasificador SVM .....	73
Tabla 4.13: Resultados de recall usando el clasificador MLP.....	74
Tabla 4.14: Resultados de recall usando el clasificador KNN .....	74
Tabla 4.15: Resultados de recall usando el clasificador SVM .....	75
Tabla 4.16: Resultados de recall usando el clasificador MLP.....	75
Tabla 4.17: Resultados de recall usando el clasificador KNN .....	76
Tabla 4.18: Resultados de recall usando el clasificador SVM .....	76
Tabla 4.19: Resultados de recall usando el clasificador MLP.....	77
Tabla 4.20: Resultados de recall usando el clasificador KNN .....	78
Tabla 4.21: Resultados de recall usando el clasificador SVM .....	79
Tabla 4.22: Resultados de recall usando el clasificador MLP.....	80
Tabla 4.23: Resultados de recall usando el clasificador KNN .....	81
Tabla 4.24: Resultados de recall usando el clasificador SVM .....	82
Tabla 4.25: Mejores resultados para la firma MFCC.....	83
Tabla 4.26: Mejores resultados para la firma RASTA-MFCC.....	84

Tabla 4.27: Mejores resultados para la firma Entropy Signature .....	85
Tabla 4.28: Mejores resultados para la firma MFCC.....	86
Tabla 4.29: Resultados de recall usando el clasificador MLP.....	87
Tabla 4.30: Resultados de recall usando el clasificador KNN .....	87
Tabla 4.31: Resultados de recall usando el clasificador SVM .....	88
Tabla 4.32: Resultados de recall usando el clasificador MLP.....	88
Tabla 4.33: Resultados de recall usando el clasificador KNN .....	89
Tabla 4.34: Resultados de recall usando el clasificador SVM .....	89
Tabla 4.35: Resultados de recall usando el clasificador MLP.....	90
Tabla 4.36: Resultados de recall usando el clasificador KNN .....	91
Tabla 4.37: Resultados de recall usando el clasificador SVM .....	92
Tabla 4.38: Resultados de recall usando el clasificador MLP.....	93
Tabla 4.39: Resultados de recall usando el clasificador KNN .....	94
Tabla 4.40: Resultados de recall usando el clasificador SVM .....	95
Tabla 4.41: Resultados de recall usando el clasificador MLP.....	96
Tabla 4.42: Resultados de recall usando el clasificador KNN .....	96
Tabla 4.43: Resultados de recall usando el clasificador MLP.....	97
Tabla 4.44: Resultados de recall usando el clasificador MLP.....	97
Tabla 4.45: Resultados de recall usando el clasificador KNN .....	98
Tabla 4.46: Resultados de recall usando el clasificador SVM .....	98
Tabla 4.47: Resultados de recall usando el clasificador MLP.....	99
Tabla 4.48: Resultados de recall usando el clasificador KNN .....	100
Tabla 4.49: Resultados de recall usando el clasificador SVM .....	101
Tabla 4.50: Resultados de recall usando el clasificador MLP.....	102
Tabla 4.51: Resultados de recall usando el clasificador KNN .....	103
Tabla 4.52: Resultados de recall usando el clasificador SVM .....	104
Tabla 4.53: Mejores resultados para la firma MFCC.....	105
Tabla 4.54: Mejores resultados para la firma RASTA-MFCC.....	106
Tabla 4.55: Mejores resultados para la firma Entropy Signature .....	107
Tabla 4.56: Mejores resultados para la firma MSES .....	108
Tabla 4.57: Matriz de confusión del mejor resultado de la base de datos EMODB .....	109
Tabla 4.58: Matriz de confusión del mejor resultado de la base de datos EMOVO .....	109
Tabla 4.59: Mejores resultados en el estado del arte para la base de datos EMODB.....	110
Tabla 4.60: Mejores resultados en el estado del arte para la base de datos EMOVO .....	110

# CAPITULO 1 INTRODUCCIÓN

## 1.1 Introducción

Las emociones son un elemento inherente a los seres humanos. El afecto y la emoción juegan un papel importante en nuestras vidas y están presentes en mucho de lo que hacemos [1]. Como humanos encontramos que el habla es la forma más natural de expresarnos. Dependemos tanto de ella que reconocemos su importancia cuando tenemos que utilizar otras formas de comunicación, como los correos electrónicos o los mensajes de texto. No es sorprendente que los emojis se hayan convertido en algo común en los mensajes de texto, porque estos mensajes de texto podrían ser malinterpretados, y nos gustaría pasar la emoción junto con el texto como lo hacemos en el habla.

Dado que las emociones nos ayudan a entendernos mejor, un resultado natural es extender esta comprensión a los ordenadores [2]. En la actualidad, los sistemas de Interacción Humano Computadora (IHC) tienden a incorporar sistemas de habla y visión, ya que estos medios son los canales más naturales en la comunicación humana. Uno de los objetivos que persiguen los sistemas de IHC es que la interacción en estos escenarios sea bidireccional, para lo cual una máquina debe escuchar el mensaje del usuario y responder de manera natural. Para alcanzar esta forma de interacción la expresión emocional debe ser reconocida y sintetizada [1]. De esta manera, los sistemas de IHC pueden mejorar el rendimiento de los sistemas de reconocimiento automático del habla (RAH) y, por lo tanto, es muy útil para la investigación criminal, la asistencia inteligente, la vigilancia y la detección de eventos potencialmente peligrosos, y los sistemas de atención sanitaria. El reconocimiento de las emociones del habla es particularmente útil en la interacción hombre-máquina [3].

Las siguientes aplicaciones son un ejemplo de cómo se puede aprovechar el conocimiento del estado emocional de los usuarios para tomar decisiones sobre qué acciones debe seguir un sistema

A) Un sistema telefónico de atención automática a clientes que provee asistencia médica a usuarios que llaman pidiendo ayuda [4]. Dichos usuarios podrían presentar diferentes emociones como tensión, miedo, dolor o pánico dependiendo de la enfermedad o de la emergencia que están experimentando. El manejo de una llamada será diferente dependiendo de la clasificación del estado emocional del usuario, dando prioridad a las llamadas más urgentes; dirigiéndose a la persona indicada.

B) Un tutorial interactivo en el que se podría adaptar la carga emocional de la respuesta del sistema buscando motivar y captar el interés dependiendo del estado emocional del alumno [5].

C) Otra aplicación es un Sistema de Respuesta Interactiva por Voz que atiende pacientes con problemas psicológicos [6]. El sistema detecta si hay algún grado de depresión basándose principalmente en características articulatorias de la calidad de voz del paciente. El sistema alerta a un experto humano cuando detecta en el paciente un grado de depresión alarmante.

D) Las aplicaciones del reconocimiento automático de carga emocional en la voz no se limitan únicamente a la IHC. En la interacción humano – humano, puede usarse para monitorear conversaciones entre agentes y clientes en call centers y detectar estados emocionales no deseados [4]. Por ejemplo, un cliente enojado o frustrado o un agente con actitud altanera. De esta manera un inspector de calidad puede tomar decisiones sobre la administración y mejora del personal y de los servicios.

Como muestran estos ejemplos de aplicaciones, mediante el reconocimiento de emociones se puede incrementar el desempeño, la usabilidad y en general la calidad de sistemas de IHC, sistemas de atención a clientes y otros tipos de aplicaciones. Sin embargo, el reconocimiento automático de emociones es un problema complejo, por lo cual ha sido difícil de implementar en aplicaciones reales [1].

Para reconocer las áreas de investigación futuras en los sistemas de RAH, hay que ser consciente de los enfoques actuales, los retos a los que se enfrenta cada uno y las cuestiones que deben abordarse. Por consiguiente, en este documento se examina el mecanismo de producción de habla humana. Se abordan en detalle las diversas técnicas, y modelos de reconocimiento del habla. Se describen los parámetros de rendimiento que miden la precisión del sistema en el reconocimiento de la señal del habla.

## 1.2 Mecanismo de producción del habla humana

El habla es la transformación de los pensamientos en palabras. Los oídos humanos perciben la señal de sonido y hay una conversión de la señal de presión en señal eléctrica. El mensaje es transmitido al cerebro donde se realiza el procesamiento y se toma la decisión apropiada. Si la respuesta va a ser verbal, entonces se transmite al modelo de habla a través del sistema motor del cuerpo humano. La articulación y la formación de mensajes significativos se llevan a cabo por el modelo del habla. El modelo de producción del habla humana se muestra en la Figura 1.1. Aquí, el sonido se produce cuando la presión del aire se aplica al pulmón a través de los músculos y luego esta señal de presión pasa a través del tracto vocal. El tracto vocal tiene pliegues vocales que se caracterizan por tener diferentes frecuencias de resonancia. La apertura y el cierre de las cuerdas vocales producen diferentes palabras o señales de sonido. La señal sonora puede pasar a través de la cavidad oral produciendo sonidos orales o de la cavidad nasal produciendo sonidos nasales, dependiendo del cierre o la apertura del velo respectivamente [7].

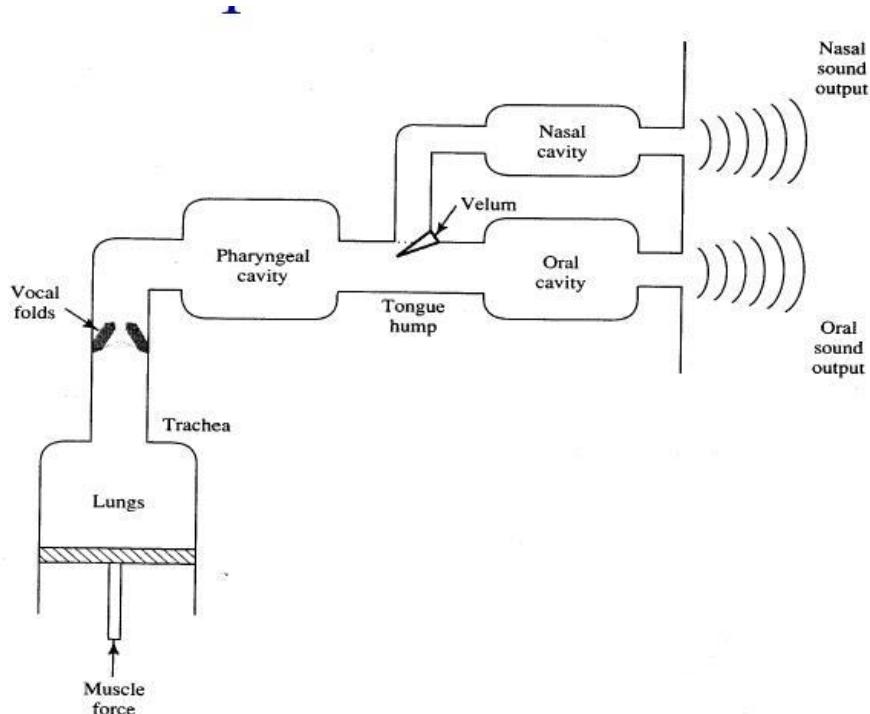


Figura 1.1: Modelo de producción del habla humana [7].

## 1.3 Clasificación de la señal de voz

Hay varias maneras de clasificar la señal del habla. Uno de los métodos se basa en la periodicidad de la señal. Se identifican tres clases más amplias y sus descripciones se discuten a continuación.

- **voz vocalizada:** Si la señal de voz es cuasiperiódica, se dice que la señal es vocal y se produce por la vibración periódica de las cuerdas vocales en el tracto vocal. La señal vocal lleva información importante sobre las expresiones y por lo tanto reconocerla se vuelve importante.
- **voz no vocalizada:** Los sonidos no verbales son señales aperiódicas que se producen como turbulencias debido a la constricción de la señal en algún punto del tracto vocal (especialmente cerca del extremo de la boca). La señal no verbal puede ser el ruido que está presente en las frases habladas reales.
- **Silencio:** El silencio es el período de pausas entre las palabras o frases. Tiene cero información y tiene que ser cuidado mientras se procesa.

Se observa que se realiza un cuidadoso análisis mediante el modelo RAH para determinar las porciones de voz, no voz o silencio de la señal de voz. Este modelo formula algoritmos que entrena a la computadora para realizar el procesamiento de la información y determina cuál es la palabra o la frase pronunciada. También realiza una especie de mapeo de la señal de voz en tiempo continuo a una secuencia de muestras de tiempo discretas [7].

## 1.4 Desafíos del reconocimiento automático del habla

El habla es una señal continua no estacionaria cuya aplicación para el reconocimiento del habla plantea muchos problemas. Algunos de ellos se describen a continuación.

- Los oradores tienen un lenguaje de comunicación diferente. Así que la máquina tiene que ser alimentada con la base de datos de cada idioma que será enorme y esto aumenta el costo de implementación y el tiempo de búsqueda.
- Los humanos interactúan no sólo con palabras sino también con la ayuda de gestos. La necesidad de inculcar los gestos y el reconocimiento de emociones con el habla se convierte en una tarea difícil.
- La grabación de la señal de voz introducirá ruido de fondo que afectará a la precisión del reconocimiento. Por lo tanto, se debe adoptar un proceso de eliminación de ruidos para mejorar el rendimiento del RAH.
- El reconocimiento del habla dará resultados precisos cuando el procesamiento se lleve a cabo en un solo fonema, palabra, frase u oraciones. Pero en tiempo real, el habla continua tiene que ser procesada, lo cual es muy difícil de implementar.
- Las características del tracto vocal varían con la edad, el género, el acento utilizado al hablar, el estado emocional del hablante al grabar, etc. También el lenguaje utilizado para la comunicación será diferente. Por lo tanto, generalizar el modelo de reconocimiento de voz requerirá una enorme base de datos y tiempo de procesamiento [7].

## 1.5 Revisión del estado del arte

### 1.5.1 Noise Robust Speaker Identification Using RASTA–MFCC Feature with Quadrilateral Filter Bank Structure.

Este documento [8] motiva el uso de Relative Spectra–Mel Frequency Cepstral Coefficients (RASTA–MFCC), característica extraída de un nuevo diseño de la estructura del banco de filtros cuadriláteros y de Gaussian Mixture Model–Universal Background Model (GMM–UBM) para mejorar la identificación del orador independiente del texto en un entorno ruidoso.

A diferencia del modelo de red neuronal que requiere la readaptación de toda la base de datos cuando se añade una nueva muestra, el modelo GMM–UBM no requiere la readaptación de toda la base de datos lo que lleva a un procesamiento más fácil y rápido. El RASTA–MFCC suele ser más robusto para un ambiente ruidoso en comparación con el método tradicional de MFCC. MFCC es una característica eficiente para identificar al individuo ya que tiene la capacidad de captar información específica de la persona.

El procesamiento RASTA mejora el rendimiento del reconocedor en presencia de la convolución y el ruido aditivo. Este trabajo combina lo mejor de estos dos procesos para producir la característica de RASTA–MFCC que es robusta al ruido y también propone una nueva estructura Cuadrilátera del banco de filtros que se aproxima a la respuesta de la membrana coclear del oído humano para capturar eficazmente los vectores de características. La estructura propuesta del banco de filtros con la característica RASTA–MFCC y el modelado GMM–UBM para la identificación de los individuos demuestra supremacía sobre los bancos de filtros triangulares y gaussianos y ofrece una precisión de identificación del 97,67% para la base de datos de habla ruidosa de la MEPCO para 50 oradores.

### 1.5.2 Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods.

Las investigaciones sobre el reconocimiento automático de emociones rara vez han abordado la cuestión de la utilización de los recursos. Con el advenimiento de la tecnología de inteligencia ambiental que emplea una variedad de dispositivos de baja potencia y recursos limitados, este tema está ganando cada vez más interés.

Esto es especialmente cierto en el contexto de las tecnologías para la salud y el cuidado de los ancianos, donde las intervenciones pueden basarse en la vigilancia del estado emocional para proporcionar apoyo o alertar a los cuidadores como sea apropiado.

El trabajo presentado por [9] se centra en el reconocimiento de emociones a partir de los datos del habla, en entornos donde es deseable minimizar los requisitos de memoria y computación. La reducción del número de características para la inferencia inductiva es una ruta hacia este objetivo.

En este estudio, se evalúan tres métodos diferentes de selección de características de última generación: Infinite Latent Feature Selection (ILFS), ReliefF y Fisher (puntuación generalizada de Fisher). Realizado en tres conjuntos de datos de reconocimiento de emociones (EmoDB, SAVEE y EMOVO) usando dos conjuntos de características acústicas paralingüísticas estándar (es decir, eGeMAPs y emobase). Los resultados muestran que se puede lograr una precisión similar o mejor utilizando subconjuntos de características sustancialmente más pequeñas que todo el conjunto de características.

### **1.5.3 Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition.**

La señal de voz lleva un mensaje emocional durante su producción. Con el análisis de la relación entre la producción de sonido y la glotis, el trabajo presentado por [10] introduce los rasgos glotales en el reconocimiento de las emociones del habla y ha extraído el rasgo GCZCMT. Se diseñaron dos experimentos, el primero utilizó las bases de datos de habla emocional TYUT y Berlín Emotional Database (EMODB), el propósito de dicho experimento fue investigar la capacidad de reconocimiento de emociones de la característica GCZCMT. Los resultados del experimento mostraron que la GCZCMT es una característica que efectivamente distingue el estado emocional. El segundo experimento consistió en mezclar la base de datos del habla, el propósito de este experimento fue investigar la capacidad de reconocimiento de emociones de la característica del GCZCMT en el lenguaje de la base de datos ross. Los resultados experimentales mostraron que la dependencia de la base de datos de la característica del GCZCMT es mínima, y que dicha característica es más adecuada para el entorno real del lenguaje complejo.

- El documento ha introducido los rasgos glotales en el reconocimiento de las emociones del habla
- Los resultados muestran que el Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT) tiene una característica que distingue efectivamente el estado emocional.
- Se puede ver que el GCZCMT tiene un alto valor práctico.

### **1.5.4 Speech Emotion Recognition Using Fourier Parameters.**

Recientemente se han realizado estudios sobre características de armonía para el reconocimiento de la emoción del habla. En el trabajo presentado por [3] realizaron un estudio utilizando las diferencias de primer y segundo orden de las características de armonía y observaron que también juegan un papel importante en el reconocimiento de las emociones del habla. Por lo tanto, se propuso un nuevo modelo de parámetros de Fourier utilizando el contenido perceptivo de la calidad de la voz y las diferencias de primer y de segundo orden para el reconocimiento de las emociones del habla independiente del hablante. Los resultados experimentales muestran que las características del parámetro de Fourier (FP) propuesto son eficaces para identificar varios estados emocionales en las señales del habla. Mejoran las tasas de reconocimiento con respecto a los métodos que utilizan los coeficientes cepstrales de frecuencia de Mel (MFCC) utilizando la base de datos EMODB, la base de datos en idioma chino (CASIA) y la base de datos de emociones de ancianos chinos (EESDB).

### **1.5.5 Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo.**

En esta investigación [1] se trabajó en el reconocimiento de emociones a partir de la señal de voz enfocándose en bases de datos de emociones espontáneas. El reconocimiento de emociones tiene especial importancia en el área de sistemas de interacción humano - computadora, y en sistemas de interacción humano - humano, ya que permite mejorar la calidad de los servicios prestados por estos sistemas, habilitándolos para tomar decisiones importantes basándose en el estado emocional de los usuarios. Para atacar este problema se pretende explorar la utilización de características acústicas principalmente. Se adoptó el enfoque del modelado continuo de emociones, probando técnicas de computación suave y probabilistas para la clasificación de emociones. Este trabajo aportó en la

comprensión de los elementos del habla que ayudan a reconocer las emociones y en la creación de un método de reconocimiento de patrones basado en el modelo emocional continuo apropiado para emociones espontáneas. Hasta el momento se ha experimentado con varios tipos de características, incluyendo nuevas características utilizadas en otros campos y se ha recurrido a técnicas de selección de atributos para encontrar las más importantes. Así mismo, se han clasificado estados emocionales basándose en 2 modelos psicológicos de las emociones, el modelo continuo y el modelo discreto. Los resultados obtenidos son comparables con los mejores resultados en el estado del arte.

### **1.5.6 Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models.**

Las emociones que exhibe un orador pueden ser detectadas analizando su habla, sus expresiones faciales y sus gestos o combinando estas propiedades. El trabajo presentado en [11] se concentra en determinar el estado emocional a partir de las señales del habla. Se extraen varias características acústicas como la energía, la tasa de cruce cero (ZCR), la frecuencia fundamental, los MFCC, etc., para un análisis de tiempo corto derivados de la señal del habla. Se construyó un vector de características para cada expresión analizando las estadísticas globales (media, mediana, etc.) de las características extraídas en todos los frames. Para seleccionar un subconjunto de características útiles del vector de características candidato completo, se utilizó el método de Sequential Backward Selection (SBS) con validación cruzada de pliegue k. La detección de la emoción en las muestras se realiza clasificando sus respectivos vectores de características en clases, utilizando un modelo de Máquinas de soporte vectorial (SVM) previamente entrenado y un clasificador de Linear Discriminant Analysis (LDA). Este enfoque se probó con dos bases de datos de emociones actualizadas: la Base de datos EMODB y la Base de datos de emociones RML (RED). Para la clasificación de clases múltiples, se logró una precisión del 80% para la EMODB y del 73% para la RED, que son superiores o comparables a los trabajos anteriores de ambas bases de datos.

### **1.5.7 Investigation of the Relation between Emotional State and Acoustic Parameters in the Context of Language.**

El análisis acústico es el método más básico utilizado para el reconocimiento de emociones del habla. Los registros del habla se digitalizan mediante métodos de procesamiento de señales, y varias características acústicas del habla son obtenidas por métodos de análisis acústico. La relación entre las características acústicas y la emoción se ha investigado en muchos estudios. Sin embargo, los estudios se han centrado sobre todo en el éxito del reconocimiento de las emociones o en los efectos de las emociones en los rasgos acústicos. El efecto del lenguaje hablado en el reconocimiento de las emociones del habla se ha investigado en un número limitado. En este estudio [12] se investigó la variabilidad de la relación entre las características acústicas y las emociones según el lenguaje hablado. Para ello, se utilizaron tres emociones (ira, miedo y neutralidad) de tres idiomas hablados diferentes (inglés, alemán e italiano). En estos conjuntos de datos, se investigó estadísticamente el cambio en las características acústicas según el idioma hablado. De acuerdo con los resultados obtenidos, el efecto de la ira en los rasgos acústicos no cambia según el idioma hablado. Por miedo, el cambio en el lenguaje hablado muestra una alta similitud en el italiano y el alemán, pero una baja similitud en el inglés.

### **1.5.8 Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique.**

El reconocimiento de las emociones del habla de corpus abstracto puede ser una técnica de aprendizaje de transferencia útil para construir un sistema de reconocimiento de las emociones del habla robusto aprovechando la información de varios conjuntos de datos del habla: de corpus transversal y de corpus cruzado. Sin embargo, es necesario llevar a cabo más investigaciones para comprender los escenarios operativos efectivos del reconocimiento de emociones del habla entre corpus, especialmente con la utilización de las poderosas técnicas de aprendizaje profundo. En este artículo [13], se utilizaron cinco corpus diferentes de tres idiomas diferentes para investigar el reconocimiento de emociones entre corpus y entre idiomas utilizando Deep Belief Networks (DBNs). Los resultados experimentales demuestran que las DBNs con poder de generalización ofrecen una mayor precisión que un método discriminatorio basado en el Sparse Auto Encoder y SVM. Los resultados también sugieren que el uso de un gran número de idiomas para el entrenamiento y el uso de una pequeña fracción de datos de destino en el entrenamiento pueden aumentar significativamente la precisión en comparación con el uso del mismo idioma para el entrenamiento y las pruebas.

### **1.5.9 Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace.**

Este documento [14] propone un enfoque analítico basado en Kernel Canonical Correlation Analysis (KCCA) para la adaptación del dominio. Para generar instancias emparejadas para el KCCA, mapean los datos de origen y de destino en los componentes principales de origen y de destino. Realizan una adaptación de dominio por parejas entre cuatro corpus emocionales de habla con diferentes idiomas (inglés, alemán, italiano y polaco) para validar el enfoque. Comparan su enfoque con el Auto-Encoder SharedHidden-Layer (SHLA) y los componentes principales basados en el núcleo. En promedio, el enfoque propuesto produce un mayor rendimiento de clasificación.

### **1.5.10 The Effect of Noise on Emotion Perception in an Unknown Language.**

Este estudio [15] investiga la influencia del ruido realista en la percepción de las emociones verbales en un idioma desconocido. Se hace vinculando la percepción de las emociones a características acústicas que se sabe que están correlacionadas con la percepción de las emociones e investigando el efecto del ruido en la percepción de estas características acústicas. Los estudiantes holandeses escucharon frases italianas en cinco emociones y se les pidió que indicaran la emoción que se transmitía en la frase. Las frases se presentaron en condiciones de ruido limpio y con balbuceos. Los resultados mostraron que los participantes eran capaces de reconocer las emociones en idioma desconocido, y continuaron reconociendo por encima de la casualidad incluso en condiciones de escucha bastante malas, indicando que la emoción verbal puede contener características universales. El ruido tuvo un efecto perjudicial similar en la percepción de las diferentes emociones, aunque el impacto en el uso de los parámetros acústicos para las diferentes categorías de emociones fue diferente.

### **1.5.11 A novel feature selection method for speech emotion recognition.**

El reconocimiento de las emociones del habla implica analizar los cambios vocales causados por las emociones con análisis acústico y determinar las características que se utilizarán para el reconocimiento de emociones. El número de características obtenidas por el análisis acústico alcanza

valores muy elevados según el número de parámetros acústicos utilizados y las variaciones estadísticas de estos parámetros. No todas estas características son efectivas para el reconocimiento de emociones; Además, diferentes emociones pueden afectar diferentes características vocales. Por esta razón, los métodos de selección de características se utilizan para aumentar el éxito del reconocimiento emocional y reducir la carga de trabajo con menos funciones. Ahí no hay certeza de que los métodos de selección de características existentes aumenten / disminuyan el éxito del reconocimiento de emociones; algunos de estos métodos aumentan la carga de trabajo total. En este estudio [16], se propone un método basado en los cambios en las emociones sobre las características acústicas. El método de la propuesta se compara con otros métodos más utilizados en la literatura. La comparación se realizó en base a número de éxito en el reconocimiento de funciones y emociones. Según los resultados obtenidos, en la propuesta. El método proporciona una reducción significativa en el número de características, así como también aumenta de la precisión en la clasificación.

### **1.5.12 Wavelet packet analysis for speaker-independent emotion recognition.**

Extraer características efectivas de las señales del habla es esencial para reconocer diferentes emociones. Estudios recientes han demostrado que el análisis wavelet es una técnica útil en el procesamiento de señales. En este estudio [17], se extraen características emocionales de la señal de voz utilizando wavelet packet análisis. Se exploraron y evaluaron estas características a partir de dos bases de datos EMODB y EESDB. Se encuentra que las características extraídas son efectivas para reconocer diversas emociones del habla. Además, en comparación con características comunes como los coeficientes cepstrales de frecuencia Mel (MFCC), estas características pueden mejorar las tasas de reconocimiento en un 14.9% y un 4.3% en EMODB y EESDB, respectivamente.

En la Tabla 1.1, se muestra un resumen del estado del arte enfocado en los resultados obtenidos con las bases de datos EMODB y EMOVO, por ello, se descartan los detalles de otras bases de datos utilizadas por estos trabajos.

Las primeras dos filas de la Tabla 1.1 muestran los detalles para los trabajos [1], [17] que utilizaron la base de datos EMODB, de la tercera a la sexta fila están los trabajos [9], [13], [14], [16] que utilizaron tanto EMODB como EMOVO, de la séptima a la décima fila están los trabajos que no utilizaron la base de datos completa [3], [10]–[12], en la onceava fila está el trabajo [15] el cual utilizo EMOVO y uso como clasificador estudiantes holandeses, y en la última fila está el trabajo [8] el cual no uso ninguno de los conjuntos de datos utilizados en esta tesis pero obtuvo buenos resultados al utilizar la característica RASTA-MFCC.

Paper	Dataset	Features	Classifier	Results
Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo [1].	EMODB	Características acústicas.	SVM	Se obtuvo un puntaje recall de 84.13%
Wavelet packet analysis for speaker-independent emotion recognition [17].	EMODB	Wavelet Packet Coefficient (WPC) con Sequential Floating Forward Search (SFFS).	RSVM	El puntaje recall obtenido es de 79.2%
Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace [14].	EMODB y EMOVO	Se extrajeron 384 características usando openSMILE.	Logístico simple con Kernel Canonical Correlation Analysis (KCCA).	El puntaje recall obtenido para EMODB y EMOVO es de 71.9% y 59.5% para respectivamente.
Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique [13].	EMODB EMOVO	eGemaps feature set.	Deep Belief Networks (DBNs).	El puntaje recall obtenido para EMODB y EMOVO es de 72.38% y 76.22% respectivamente.
A novel feature selection method for speech emotion recognition [16].	EMODB EMOVO	Se extrajeron 1582 características usando OpenSMILE.	SVM	El puntaje recall obtenido para EMODB y EMOVO es de 84.62% y 60.40% respectivamente.
Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods [9].	EMODB EMOVO	Emobase y eGeMAPs con Infinite Latent Feature Selection (ILFS) y puntuación generalizada de Fisher.	SVM	El puntaje recall obtenido para EMODB y EMOVO es de 76.9% y 41.0% respectivamente.
Speech Emotion Recognition Using Fourier Parameters [3].	EMODB Se descarta la clase disgusto, por lo que solo se utilizan 6 clases.	Características del parámetro de Fourier (FP).	SVM	Se obtuvo un puntaje recall de 89%.
Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models [11].	EMODB Se descarta la clase disgusto, por lo que solo se utilizan 6 clases.	Características acústicas.	SVM	se logró un puntaje recall de 81%.
Investigation of the Relation between Emotional State and Acoustic Parameters in the Context of Language [12].	Solo utiliza 3 clases de los conjuntos EMODB y EMOVO (ira, angustia y neutral).	Características acústicas.	Se realizó un análisis estadístico mediante la prueba U de MannWhitney.	De acuerdo con los resultados obtenidos, el efecto de la ira en los rasgos acústicos no cambia según el idioma hablado. Por miedo, el cambio en el lenguaje hablado muestra una alta similitud en el italiano y el alemán.
Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition [10].	3 clases de EMODB (felicidad, ira y neutral).	Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT).	SVM	El puntaje recall obtenido es de 81.83%.
The Effect of Noise on Emotion Perception in an Unknown Language [15].	EMOVO	Características acústicas.	Estudiantes holandeses.	Los resultados indican que las emociones pueden tener características universales.
Noise Robust Speaker Identification Using RASTA-MFCC Feature with Quadrilateral Filter Bank Structure [8].	MEMPCO	RASTA-MFCC	Gaussian Mixture Model-Universal Background Model (GMM-UBM).	ofrece una precisión de identificación del 97,67%.

Tabla 1.1: Revisión del estado del arte.

## 1.6 Planteamiento del problema

El área de reconocimiento de emociones por voz (REV) existe desde hace más de dos décadas y ha sido un área de investigación muy activa en los últimos años, llegando a tener muchas aplicaciones en la interacción hombre-computadora, así como en robots, servicios móviles, centros de llamadas, juegos de ordenador y evaluación psicológica. Aunque tiene muchas aplicaciones, la detección de emociones es una tarea difícil, porque las emociones son subjetivas.

Definimos un sistema REV como una colección de metodologías que procesan y clasifican las señales del habla para detectar las emociones incrustadas en ellas. Podemos separarlas en varias áreas distintas, como se muestra en la Figura 1.2.

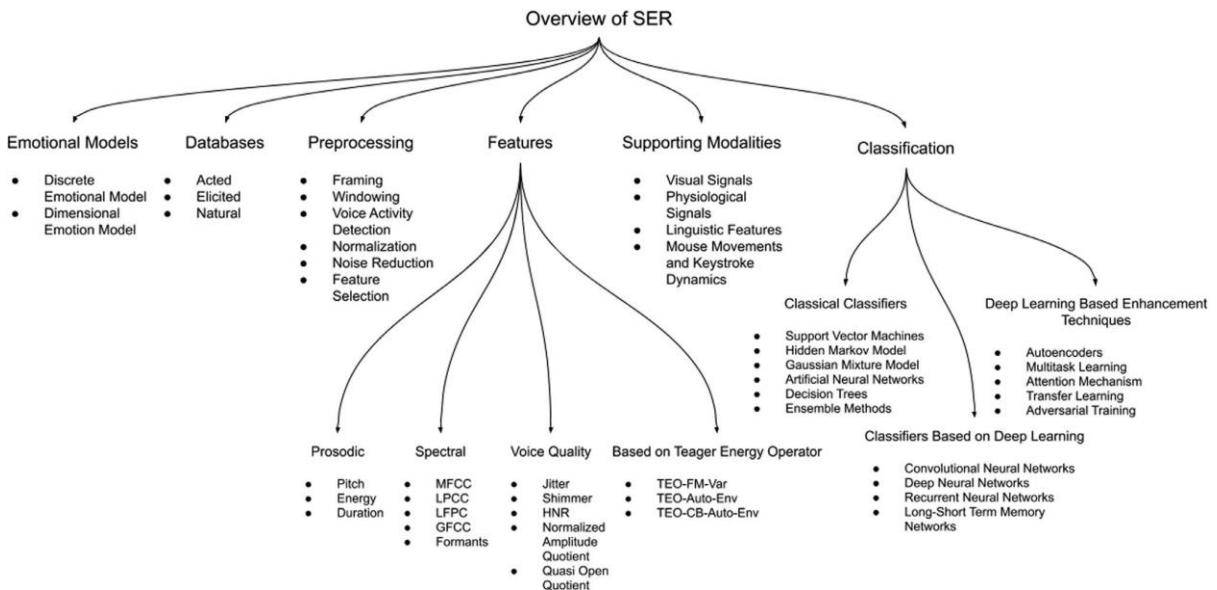


Figura 1.2: Una visión general de los sistemas de reconocimiento de emociones del habla. Los requisitos de reconocimiento fluyen de izquierda a derecha. Las emociones están incrustadas en las bases de datos en el extremo izquierdo, y se extraen en el extremo derecho de la figura [2].

Sería beneficioso entender mejor las emociones para poder mejorar el proceso de clasificación. Existen varios enfoques para modelar las emociones, y sigue siendo un problema abierto; sin embargo, los modelos discretos y los dimensionales se utilizan comúnmente.

Para implementar con éxito un sistema de reconocimiento de emociones del habla, necesitamos definir y modelar las emociones cuidadosamente. Sin embargo, no hay ningún consenso sobre la definición de la emoción, y sigue siendo un problema abierto en la psicología. Las emociones son estados psicológicos enrevesados que se componen de varios componentes, como la experiencia personal, las reacciones fisiológicas, conductuales y comunicativas. Basados en estas definiciones, dos modelos han sido comunes en el reconocimiento de las emociones del habla: el modelo emocional discreto y el modelo emocional dimensional.

La teoría de las emociones discretas se basa en las seis categorías de emociones básicas: tristeza, felicidad, miedo, ira, asco y sorpresa. Otras emociones se obtienen por la combinación de las básicas. La mayoría de los sistemas de REV existentes se centran en estas categorías emocionales básicas. En la vida diaria, la gente usa este modelo para definir sus emociones, por lo que el esquema de etiquetado basado en categorías emocionales es intuitivo. Sin embargo, estas categorías discretas de

emociones no son capaces de definir algunos de los complejos estados emocionales observados en la comunicación diaria.

El modelo emocional dimensional es un modelo alternativo que utiliza un pequeño número de dimensiones latentes para caracterizar emociones como valencia, excitación, control, poder. Estas dimensiones son aspectos definitivos y genéricos de la emoción. En el enfoque dimensional, las emociones no son independientes entre sí, sino que son análogas entre sí de manera sistemática. Uno de los modelos dimensionales más preferidos es un modelo bidimensional que utiliza la excitación, la activación o la excitación en una dimensión, frente a la valencia, la valoración o la evaluación en la otra. La dimensión de valencia describe si una emoción es positiva o negativa, y oscila entre lo desagradable y lo agradable. La dimensión de la excitación desmiente la fuerza de la emoción sentida. Puede ser excitante o apática, y va desde el aburrimiento a la excitación frenética. El modelo tridimensional incluye una dimensión de dominio o poder, que se refiere a la aparente fuerza de la persona que está entre débil y fuerte. Por ejemplo, la tercera dimensión diferencia la ira del miedo considerando la fuerza o la debilidad de la persona, respectivamente.

La representación dimensional tiene varias desventajas. No es lo suficientemente intuitiva y puede ser necesario un entrenamiento especial para etiquetar cada emoción. Además, algunas de las emociones son idénticas, como el miedo y la ira, y algunas emociones como la sorpresa no se pueden categorizar y se encuentran fuera del espacio dimensional, ya que la emoción de la sorpresa puede tener una valencia positiva o negativa según el contexto.

## 1.7 Hipótesis

Dado que el proceso de reconocimiento de emociones humanas por voz es complejo por las diferencias estructurales de la señal, es importante hacer un estudio de los diferentes parámetros involucrados en el proceso de extracción de características. El análisis en el procesamiento de la señal de voz permitirá conseguir una estrategia para elegir los parámetros más adecuados para determinar, por ejemplo, tamaño de frame, tamaño del vector característico y el uso, o no, del filtro de preénfasis. Además, dado que el proceso de reconocimiento de emociones por voz se hace generalmente utilizando sonidos vocalizados y no vocalizados, se plantea la hipótesis de que eliminando los sonidos no vocalizados se incrementará la tasa de clasificación de las diferentes emociones humanas, ya que estos sonidos suelen confundirse con el ruido (zonas de silencio) de la grabación de la señal de voz al poseer características espectrales de baja energía similares.

## **1.8 Objetivos**

### **1.8.1 Objetivo general**

Realizar un estudio computacional para el reconocimiento de emociones humanas a partir del análisis de las señales de voz y los diferentes parámetros utilizados en el método de extracción de los MFCC y las firmas de entropía espectral multi-banda, apoyandose en algoritmos de machine learning.

### **1.8.2 Objetivos particulares**

- Identificar, analizar y caracterizar los retos del problema para estructurar la hipótesis de la tesis.
- Hacer una revisión profunda del estado del arte sobre algoritmos de reconocimiento de emociones humanas.
- Implementar los algoritmos propuestos para extraer los descriptores de las señales de voz que caracterizan a cada tipo de emoción humana.
- Caracterizar las bases de datos a utilizar en la sección de experimentos, en específico, Berlin Emotional Database (EMODB) y EMOVO database.
- Clasificar las emociones presentadas en las bases de datos utilizando diferentes tamaños de frame.
- Clasificar las emociones presentadas en las bases de datos utilizando diferentes tamaños del vector característico.
- Clasificar las emociones presentadas en las bases de datos con y sin el proceso de preénfasis.
- Clasificar las emociones presentadas en las bases de datos utilizando la función de autocorrelación para eliminar segmentos no vocalizados en la señal de voz.
- Reportar los resultados obtenidos y compararlos con los del estado del arte.

## 1.9 Justificación

La demanda de la interacción hombre-máquina aumenta día a día. Por lo tanto, los desafíos del reconocimiento automático del habla (RAH) deben ser abordados con un algoritmo eficiente de reconocimiento de voz. El conocimiento de la producción del habla humana, la naturaleza de la señal del habla y la síntesis del habla mediante la máquina, las técnicas actuales de reconocimiento del habla disponibles y sus lagunas, así como para abordarlas, se hace necesario para poder destacar las áreas de mejora.

Aunque hay muchos avances en los sistemas de reconocimiento de las emociones del habla, todavía hay varios obstáculos que deben eliminarse para que el reconocimiento tenga éxito.

Uno de los problemas más importantes es la generación del conjunto de datos que se utiliza para el proceso de aprendizaje. La mayoría de los conjuntos de datos utilizados para el reconocimiento de emociones de la voz (REV) se actúan en salas especiales de silencio. Sin embargo, los datos de la vida real son ruidosos y tienen características mucho más diferentes que los otros. Aunque también se dispone de conjuntos de datos naturales, son menos numerosos. Hay problemas legales y éticos para registrar y utilizar las emociones naturales. La mayoría de las expresiones en los conjuntos de datos naturales se toman de programas de entrevistas, grabaciones de centros de llamadas y casos similares en los que las partes involucradas son informadas de la grabación. Estos conjuntos de datos no contienen todas las emociones y pueden no reflejar las emociones que se sienten. Además, hay problemas durante el etiquetado de las declaraciones. Hay anotadores humanos que etiquetan los datos del discurso después de que se registran las declaraciones. La emoción real sentida por el orador y las emociones percibidas por los anotadores humanos puede mostrar diferencias. Incluso los índices de reconocimiento de los anotadores humanos no superan el 90%. A favor de los humanos, sin embargo, creemos que también dependemos del contenido y el contexto del discurso que estamos evaluando.

También hay efectos culturales y de lenguaje en la REV. Hay varios estudios disponibles que trabajan en la REV en varios idiomas. Sin embargo, los resultados muestran que los sistemas y características actuales utilizadas no son suficientes para ello. La entonación de las emociones en el habla entre varios idiomas puede mostrar diferencias.

Un desafío que se pasa por alto es el caso de las señales de habla múltiples, donde el sistema de REV tiene que decidir en qué señal centrarse. Aunque se puede manejar mediante un algoritmo de separación del habla en la etapa de preprocesamiento, los sistemas actuales no se dan cuenta de este problema [2].

## CAPITULO 2 MARCO TEÓRICO

### 2.1 Introducción

Los investigadores y científicos han contribuido mucho en la propuesta de nuevas técnicas que ayudarán a aumentar la precisión del reconocimiento del habla. Vijayalakshmi et al., [18] destaca la arquitectura del sistema RAH utilizando diferentes enfoques. Li Deng y Xiao Li [19] discutieron los paradigmas de Machine Learning (ML) que están motivados por aplicaciones RAH. Se introducen las técnicas de polinización cruzada de ML y RAH para obtener mejores resultados. Pandey et al., [20] discutieron varias bases de datos generadas para el reconocimiento de idiomas indios. Swati et al., [21] explicaron varias técnicas de extracción y clasificación de características del habla y compararon las diversas técnicas existentes para el reconocimiento de voz. Itoh et al., [22] proporciona las métricas de medición del rendimiento que se utilizan normalmente en las técnicas de reconocimiento de voz. Se ve que la Word Error Rate (WER) es una métrica de medición de precisión ampliamente utilizada para RAH. Considerando los recientes avances en el reconocimiento de voz, no sería sorprendente predecir que para 2050 el cincuenta por ciento de las búsquedas serán búsquedas de voz. La Tabla 2.1 muestra el crecimiento de la tecnología de reconocimiento de voz a lo largo de los años.

Año	Avances importantes
1784	Wolfgang inventó la Máquina de velocidad acústico mecánica
1879	Thomas Edison inventa la primera máquina de dictado.
1952	Los laboratorios Bell presentan a Audrey capaz de reconocer números por voz.
1962	IBM Shoebox entiende dieciséis palabras en inglés.
1971	Harpy puede comprender 1011 palabras y algunas frases.
1986	IBM Tangora predice los próximos fonemas en el discurso.
2006	La NSA es capaz de aislar las palabras clave en el discurso.
2008	Google lleva el reconocimiento de voz a los dispositivos móviles.
2011	Apple lanza Siri para facilitar el asistente de voz.

Tabla 2.1: Principales avances en la tecnología de reconocimiento del habla [7].

## 2.2 Arquitectura del sistema de reconocimiento automático del habla.

El reconocimiento del habla se lleva a cabo en dos etapas, a saber, la etapa de entrenamiento y la etapa de prueba. En la etapa de entrenamiento, la entrada al sistema es la señal del habla obtenida de una base de datos. La muestra de la base de datos se procesa previamente y se extraen las características significativas aplicando diversas técnicas de extracción de características. En la etapa de prueba, se desconoce la muestra de prueba y se realiza el análisis acústico. Para analizar las etapas que intervienen en la arquitectura del sistema RAH, es muy importante conocer las bases de datos que sirven de entrada al sistema RAH [7]. La Figura 2.1 muestra la arquitectura del sistema RAH.

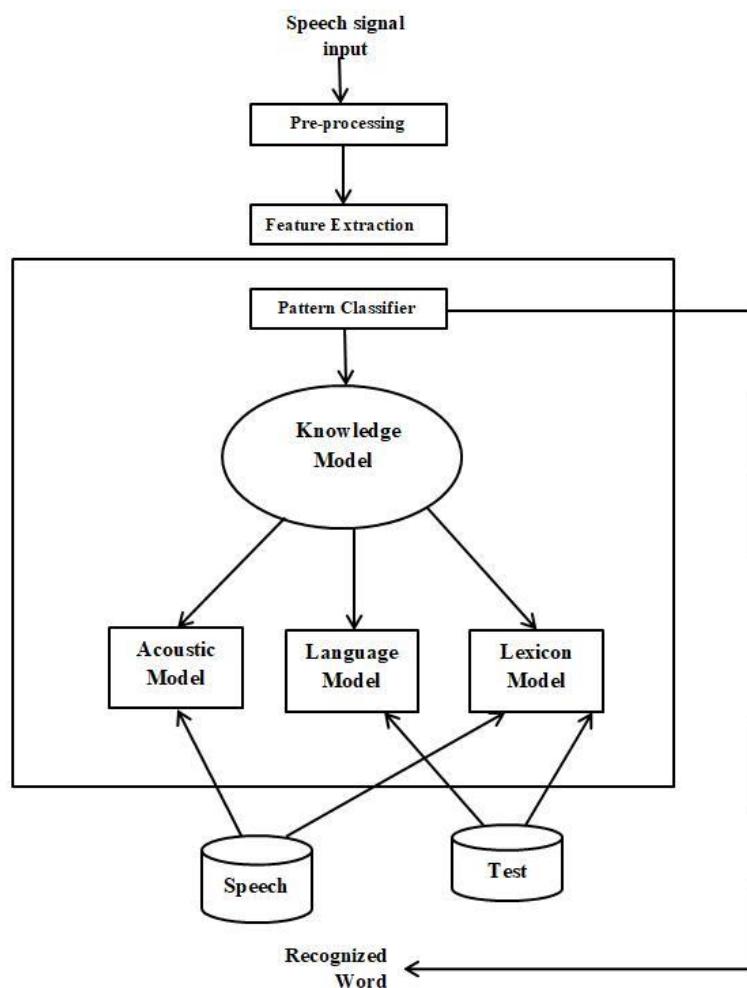


Figura 2.1: Arquitectura del sistema de reconocimiento automático del habla [7].

Los diversos pasos que intervienen en el RAH son la base de datos, el preprocesamiento, la extracción de características y la clasificación.

## 2.3 Emociones

Para implementar con éxito un sistema de reconocimiento de emociones de voz, necesitamos definir y modelar las emociones con cuidado. Sin embargo, no existe un consenso sobre la definición de emoción, y todavía es un problema abierto en psicología. Según Plutchik, más de noventa definiciones de emociones fueron propuestos en el siglo XX [23]. Las Emociones son estados psicológicos complicados que se componen de varios componentes aspectos tales como la experiencia personal, fisiológica, conductual y reacciones comunicativas. Con base en estas definiciones, dos modelos han sido comunes en el reconocimiento de emociones del habla: modelo emocional discreto y modelo emocional dimensional.

La teoría de la emoción discreta se basa en las seis categorías de emociones básicas; tristeza, felicidad, miedo, ira, disgusto y sorpresa, como se describe por [24], [25]. Estas emociones innatas y culturalmente independientes se experimentan durante un breve período de tiempo [26]. Otras emociones se obtienen mediante la combinación de las básicas. La mayoría de los sistemas REV existentes se enfocan en estas categorías emocionales. En la vida diaria, la gente usa este modelo para definir sus observaciones emocionales, de ahí el esquema de etiquetado basado en categorías son intuitivos. No obstante, estas categorías discretas de emociones no son capaces de definir algunos de los complejos estados emocionales observados en la comunicación espontánea del día a día.

El modelo emocional dimensional es un modelo alternativo que utiliza un pequeño número de dimensiones latentes para caracterizar emociones como valencia, excitación, control, poder [27], [28]. Estas dimensiones son aspectos definitivos y genéricos de emoción. En el enfoque dimensional, las emociones no son independientes el uno del otro; en cambio, son análogos entre sí de manera sistemática. Uno de los modelos dimensionales más preferidos es un modelo bidimensional que usa excitación, activación o excitación en una dimensión, versus valencia, valoración o evaluación en la otra. La dimensión de valencia describe si una emoción es positiva o negativa, y oscila entre desagradable y agradable. La dimensión de excitación refina la fuerza de la emoción sentida. Puede estar excitado o apático, y va desde el aburrimiento hasta la excitación frenética [29]. El modelo tridimensional incluye una dimensión de dominio o poder, que se refiere a la fuerza aparente de la persona que se encuentra entre lo débil y lo fuerte. Por ejemplo, la tercera dimensión diferencia la ira de miedo al considerar la fuerza o la debilidad de la persona, respectivamente [30].

Existen varias desventajas para la representación dimensional. No es lo suficientemente intuitivo y puede ser necesario un entrenamiento especial para etiquetar cada emoción [31]. Además, algunas de las emociones son idénticas, como el miedo y la ira, y algunas emociones como la sorpresa no se pueden categorizar y se encuentran fuera del espacio dimensional ya que la emoción sorpresa puede tener valencia positiva o negativa dependiendo del contexto.

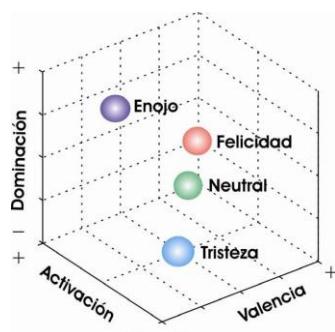


Figura 2.2: Modelo tridimensional continuo de las emociones. Valencia - Activación – Dominación [1].

Autor	Emociones Básicas
Plutchik	Aceptación, enfado, anticipación, disgusto, alegría, miedo, tristeza, sorpresa
Ekman, Friesen, Ellsworth	Ira, asco, miedo, alegría, tristeza, sorpresa
Gray	Rabia, terror, ansiedad, alegría
Izard	Ira, desprecio, disgusto, angustia, miedo, culpa, interés, alegría, vergüenza, sorpresa
James	Miedo, dolor, amor, rabia
Mowrer	Dolor, placer
Oatley and Johnson-Laird	Ira, dolor, ansiedad, felicidad, tristeza
Paksepp	Esperanza, miedo, rabia, pánico
Tomkins	Ira, interés, desprecio, disgusto, angustia, miedo, alegría, vergüenza, sorpresa
Watson	Miedo, amor, rabia
Weiner and Graham	Felicidad, tristeza

Tabla 2.2: Conjuntos de emociones básicas propuestos por diferentes autores [1].

## 2.4 Base de datos

Las bases de datos son una parte esencial del reconocimiento de las emociones del habla, ya que el proceso de clasificación se basa en los datos etiquetados. La calidad de los datos afecta al éxito del proceso de reconocimiento. Los datos incompletos, de baja calidad o defectuosos pueden dar lugar a predicciones incorrectas; por lo tanto, los datos deben cuidarse, diseñarse y recopilarse en su totalidad.

Las bases de datos para el reconocimiento de la emoción del habla pueden investigarse en tres partes:

- Bases de datos de emociones del habla actuadas (simuladas)
- Bases de datos de emociones del habla provocadas (inducidas)
- Bases de datos de emociones del habla natural

Las bases de datos de discursos actuados son grabadas por actores profesionales o semiprofesionales en estudios insonorizados. Es relativamente más fácil crear una base de datos de ese tipo en comparación con los demás métodos; sin embargo, los investigadores afirman que el discurso actuado no puede transmitir adecuadamente las emociones de la vida real, e incluso puede ser exagerado. Esto reduce las tasas de reconocimiento de las emociones de la vida real.

Las bases de datos de habla provocada se crean colocando a los oradores en una situación emocional simulada que puede estimular diversas emociones. Aunque las emociones no son totalmente provocadas, están cerca de las reales.

Las bases de datos del habla natural se obtienen principalmente de programas de entrevistas, grabaciones de centros de llamadas, charlas de radio y fuentes similares. A veces, estos discursos del mundo real se conocen como discursos espontáneos. Es más difícil obtener los datos ya que surgen problemas éticos y legales al procesarlos y distribuirlos.

Una vez que se decide el método de creación de una base de datos, se consideran otros diseños, como la edad y el género. La mayoría de las bases de datos contienen oradores adultos, pero también existen bases de datos de niños y ancianos. Otras consideraciones incluyen la repetición de declaraciones con diferentes actores, diferentes emociones y diferentes géneros.

Por ejemplo, el conjunto de datos de Berlín, de uso común, contiene siete emociones pronunciadas por diez actores profesionales, mitad hombres y mitad mujeres. Cada afirmación se repite con diferentes actores y diferentes emociones [2].

Para lograr un reconocimiento del habla preciso, las bases de datos utilizadas deben recopilarse con precisión y su alcance debe ser lo suficientemente amplio como para dar cobertura a todas las unidades acústico-fonéticas del habla. Hay dos razones principales para desarrollar una base de datos del habla. Una es la utilización de la base de datos en las esferas de investigación que abarcan los fonemas, la acústica, el léxico y las expresiones del lenguaje; y la otra es la diferenciación de los oradores en función de la edad, el sexo, el entorno, etc., al diseñar una base de datos del habla, deben tenerse en cuenta los siguientes factores.

- **Diccionario:** El vocabulario utilizado en el entrenamiento debe ser de la misma clase para el reconocimiento. Debe tener cobertura de todos los fonemas de la clase de habla en cuestión para obtener resultados satisfactorios.

- **Número de sesiones:** Esto da los detalles de las entradas hechas por un orador en particular en un determinado período de tiempo. También se puede determinar simultáneamente el tiempo de grabación de cada orador. Es deseable que el tiempo de grabación sea menor y más eficiente.
- **Aspectos técnicos:** Las grabaciones se hacen en tiempo real y es muy posible que se le añada ruido, ya sea por el entorno de grabación o por el equipo técnico utilizado para la grabación del discurso.
- **Población de individuos participantes:** Cuanto mayor sea la población de individuos participantes, mejor será la cobertura de todas las unidades de habla y mayor será la tasa de reconocimiento. Sin embargo, las grandes bases de datos requerirán más almacenamiento. Por lo tanto, tiene que haber un equilibrio entre el tamaño de la base de datos y el número de grabaciones del habla.
- **Uso previsto:** En base a la aplicación, se debe obtener el corpus de voz.
- **Variabilidad intra hablante:** La base de datos debe tener las grabaciones del hablante en diferentes emociones, expresiones, acento, etc., lo que es la variabilidad intra hablante.

La recopilación de la base de datos es un reto, ya que hay muchos idiomas y variaciones entre un mismo idioma según la cultura y la región geográfica del hablante. La mayoría de las veces no existe una base de datos estándar disponible y, por lo tanto, las bases de datos se diseñan a medida que surge la necesidad. A continuación, se presenta un panorama general de algunas de las bases de datos, para el reconocimiento de emociones, existentes [7].

Bases de datos	Idioma	Tamaño	Tipo de acceso	Emociones	Tipo	Modalidades
Berlin Emotional Database (EMODB)	alemán	7 emociones x 10 oradores (5 hombres, 5 mujeres) x 10 declaraciones.	Acceso abierto	Ira, aburrimiento, disgusto, miedo, felicidad, tristeza, neutral	Actuada	Audio
Chinese Emotional Speech Corpus (CASIA)	mandarín	6 emociones x 4 oradores (2 hombres, 2 mujeres) x 500 declaraciones (300 textos paralelos, 200 no paralelos).	Comercialmente disponible	Sorpresa, felicidad, tristeza, ira, miedo, neutral	Actuada	Audio
The Interactive EmotionalDyadic Motion CaptureDatabase (IEMOCAP)	Inglés	10 oradores (5 hombres, 5 mujeres), 1150 declaraciones.	Disponible con licencia	Alegría, enfado, tristeza, frustración, neutral	Actuada	Audiovisual
Surrey Audio-Visual Expressed Emotion (SAVEE)	Inglés	14 oradores (hombres) x 120 declaraciones.	Gratis	Ira, asco, miedo, felicidad, tristeza, sorpresa, neutral, común	Actuada	Audiovisual
Toronto Emotional SpeechDatabase (TESS)	Inglés	2 oradores (mujeres), 2800 declaraciones.	Gratis	Ira, repugnancia, miedo neutral, felicidad, tristeza, agradable, sorpresa	Actuada	Audio
Beihang University Database of Emotional Speech (BHUES)	mandarín	5 oradores (2 hombres, 3 mujeres), 323 declaraciones.		Ira, felicidad, miedo, asco, sorpresa	Actuada	Audio
Chinese Annotated Spontaneous Speech corpus (CASS)	mandarín	7 oradores (2 hombres, 5 mujeres), 6 horas de charla.	Comercialmente disponible	Ira, miedo, felicidad, tristeza, sorpresa, neutral	Natural	Audio

Bases de datos	Idioma	Tamaño	Tipo de acceso	Emociones	Tipo	Modalidades
Chinese Natural Emotional Audio-Visual Database (CHEAVD)	mandarín	238 oradores (niño a anciano), 140 min de segmentos emocionales de películas y programas de televisión.	Uso gratuito para investigaci ón	Ira, ansiedad, disgusto, alegría, neutral, tristeza, sorpresa y preocupación	Actuada natural	Audiovisual
Danish Emotional SpeechDatabase (DES)	danés	4 oradores (2 hombres, 2 mujeres), 10 min de charla.	Gratis	Neutral, sorpresa, enfado, felicidad, tristeza	Actuada	Audio
Chinese Elderly Emotional Speech Database (EESDB)	mandarín	16 oradores (8 hombres, 8 mujeres), 400 declaraciones de teleplay.	Libre para investigar	Ira, disgusto, miedo, felicidad, neutralidad, tristeza, sorpresa	Actuada	Audio
Electromagnetic Articulography Databas (EMA)	Inglés	3 oradores (1 hombre, 2 mujeres), 14 oraciones para hombres y 10 oraciones para mujeres.	Libre para investigar	Ira, felicidad, tristeza, neutral	Actuada	Audio y datos de movimiento articulatorio
Italian Emotional Speech Database (EMOVO)	italiano	6 oradores (3 hombres, 3 mujeres) x 14 frases x 7 emociones = 588 declaraciones.	Gratis	Asco, felicidad, miedo, rabia, sorpresa, tristeza, neutral	Actuada	Audio
eINTERFACE'05 Audio-Visual Emotion Database	Inglés	42 oradores (34 hombres, 8 mujeres) de 14 nacionalidades, 1116 secuencias de video.	Gratis	Ira, asco, miedo, felicidad, tristeza, sorpresa	Provocada	Audiovisual
Keio University Japanese Emotional Speech Database (Keio-ESD)	Japonés	71 oradores (hombre) 940 expresiones.	Gratis	Ira, felicidad, repugnancia, degradación, gracioso, preocupado, gentil, alivio, indignación, vergüenza , etc. (47 emociones)	Actuada	Audio
LDC Emotional Speech Database	Inglés	7 oradores (4 hombres, 3 mujeres), 470 declaraciones.	Comercial mente disponible	ira caliente, ira fría, disgusto, miedo, desprecio, felicidad, tristeza, neutral, pánico, orgullo, desesperación, júbilo, interés, vergüenza, aburrimiento	Actuada	Audio
RECOLA Speech Database	Francés	46 oradores (19 hombres, 27 mujeres) 7 horas de habla.	Gratis	5 comportamientos sociales (acuerdo, dominio, compromiso, rendimiento, compenetración); excitación y valencia	Natural	Audiovisual
SAMAINÉ Database	Inglés griego hebreo	150 oradores, 959 charlas.	Gratis	valencia, activación, poder, expectativa e intensidad	Natural	Audiovisual
Speech Under Simulated and Actual Stress Database (SUSAS)	Inglés	32 oradores (19 hombres, 13 mujeres) 16,000 declaraciones, también incluye habla de los pilotos del helicóptero apache.	Comercial mente disponible	Cuatro estados del habla bajo estrés: Neutral, enojado, ruidoso y lombardo	Actuado natural	Audio
Vera Am Mittag Database (VAM)	alemán	47 oradores del programa de entrevistas, 947 declaraciones	Gratis	valencia, activación y dominación	Natural	Audiovisual

Bases de datos	Idioma	Tamaño	Tipo de acceso	Emociones	Tipo	Modalidades
FAU Aibo Emotion Corpus	alemán	51 niños hablando con el perro robot Aibo 9 horas de discurso	Comercialmente disponible	Ira, aburrimiento, enfático, intrigado, alegre, maternal, neutral, reprimenda, descanso, sorprendido, susceptible	Natural	Audio
TUM AVIC Database	Inglés	21 oradores (11 hombres, 10 mujeres) 3091 declaraciones	Gratis	5 niveles de interés; 5 no lingüístico vocalizaciones (respiración, consentimiento, basura, vacilación, risa)	Natural	Audiovisual
AFEW Databas	Inglés	330 oradores, 1426 declaraciones de películas y programas de televisión	Gratis	Ira, asco, sorpresa, miedo, felicidad, neutral, tristeza	Natural	Audiovisual
Turkish Emotional Speech Database (TURES)	turco	582 oradores (394 hombres, 188 mujeres) de películas, 5100 declaraciones	Libre para investigar	Alegría, sorpresa, tristeza, enfado, miedo, neutral, valencia, activación, dominio	Actuada	Audio
BAUM-1 Speech Database	turco	31 oradores (18 hombres, 13 mujeres) 288 declaraciones, 1222 videoclip espontáneo	Libre para investigar	Alegría, enfado, tristeza, asco, miedo, sorpresa, molestia, aburrimiento, desprecio inseguro, pensativo, concentración, interés	Actuado natural	Audiovisual

Tabla 2.3:

Hay varios conjuntos de datos que se utilizan para el reconocimiento de emociones. Esta tabla contiene los más destacados, junto con conjuntos de datos únicos para varios idiomas y casos especiales, como los que contienen expresiones de ancianos y niños [2].

#### 2.4.1 Berlin Database of Emotional Speech (EMODB)

EMODB es una base de datos en alemán [32]. Está formada por una o dos sesiones de grabación de habla con emociones interpretadas por diez actores (cinco hombres y cinco mujeres). Cada sesión incluye diez frases (cinco cortas y cinco largas, obteniendo un total de 24,5 minutos), interpretando seis emociones (alegría, enfado, tristeza, aburrimiento, asco y miedo) y voz interpretada según el estado neutro. EMODB dispone de 24,5 minutos de voz distribuidos de manera no homogénea entre los distintos actores y las distintas emociones. Esta base de datos está etiquetada fonéticamente, pero no prosódicamente.

## **2.4.2 EMOVO una base de datos de habla emocional italiana**

Es un corpus emocional, llamado EMOVO, aplicable al idioma italiano [33]. Es una base de datos construida a partir de las voces de hasta 6 actores que interpretaron 14 frases simulando 6 estados emocionales (disgusto, miedo, ira, alegría, sorpresa, tristeza) además del estado neutro. Estas emociones son las conocidas Seis Grandes que se encuentran en la mayor parte de la literatura relacionada con el discurso emocional. Las grabaciones se hicieron con equipo profesional en los laboratorios de la Fondazione Ugo Bordoni.

## 2.5 Preprocesamiento

La señal de voz, siendo continua tiene que ser digitalizada usando convertidor analógica-digital y dada como entrada al procesamiento del sistema. El primer paso es el preprocesamiento de la señal de voz e involucra varias etapas como se muestra en la Figura 2.3.

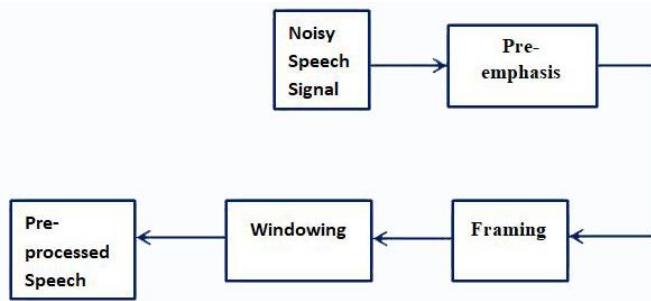


Figura 2.3: Etapas de preprocesamiento [7].

### 2.5.1 Eliminación de ruido ambiental o de fondo:

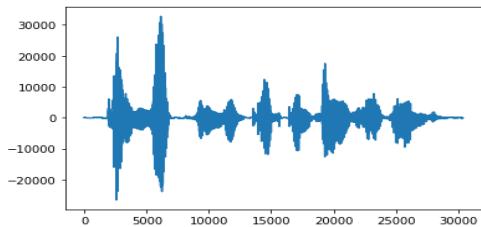
El ruido de fondo es una señal no deseada y tiene que ser suprimido para reconocer el discurso real. Por lo tanto, los filtros adecuados están diseñados para eliminar interferencias o ruidos no deseados presentes en la señal.

- **Pre-éñfasis:** El bloque de pre-éñfasis resalta las altas frecuencias que normalmente son de poca energía. El pre-éñfasis que iguala la inclinación espectral del habla se da en la Ecuación (2.1) con el factor de pre-éñfasis  $\alpha$  con un valor de 0,97. Donde  $s(n)$  es el n-ésimo instante de la señal de voz,  $s(n - 1)$  es el n - 1° instante de la señal de voz,  $\hat{s}(n)$  es el n-ésimo instante de la señal pre-enfatizada.

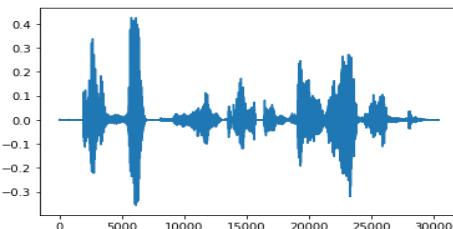
$$\hat{s}(n) = s(n) - \alpha s(n - 1) \quad (2.1)$$

- **Detección de actividad de voz:** Los detectores están diseñados para seleccionar en la señal las secciones sordas y eliminarlas usando algunos de los parámetros clásicos como la tasa de cruce por cero, la energía de la función de señal y autocorrelación.
- **Framing:** el habla es una señal casi estacionaria y por lo tanto, el procesamiento del habla debe realizarse en un análisis de tiempo corto el cual consiste en extraer de la señal de voz frames de longitud fija (número de muestras en el frame) que se superponen entre sí. Para lograr esto, la señal de voz normalmente se divide en frames de 20-30 ms y superposición de 30-50% de cada frame con frame adyacentes.
- **Ventanas:** los frames de la señal de voz son multiplicados con una función ventana para minimizar el fenómeno de Gibbs que se ocasiona en el espectro de frecuencia por haber truncado la señal de voz.

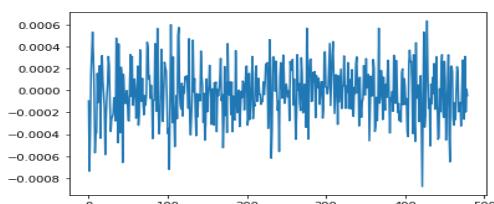
La figura 2.4 muestra el preprocesamiento que se hace a la señal de voz en cada uno de los bloques de la figura 2.3. La figura(a) es la forma de onda original. La figura(b) es la salida de la etapa de preénfasis después de resaltar los componentes de alta frecuencia. La figura(c) muestra la forma de onda obtenida después de realizar el framing en la señal. Aquí se eligen frames de 30 ms y superposición entre frames del 50%. La figura(d) muestra la representación gráfica de la ventana de Hann. La figura(e) muestra el efecto de aplicar una ventana de Hann en uno de los frames. Al final de la etapa de preprocesamiento, la señal de voz se encuentra en la forma adecuada para extraer características útiles y, por lo tanto, ayudar a un reconocimiento preciso.



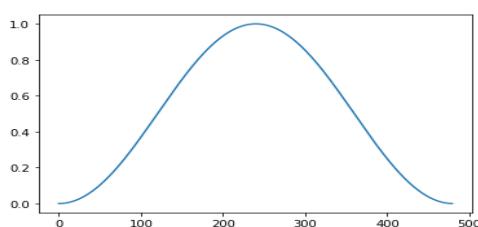
*(a): Señal del habla original*



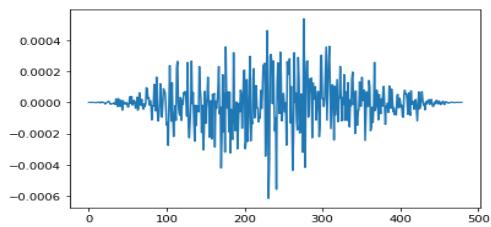
*(b): Preénfasis*



*(c): Framing, 50% traslape entre frames de 30ms*



*(d): Ventana de Hann*



(e): Salida del bloque de ventana

Figura 2.4: Efectos del preprocessamiento de la señal de voz.

## 2.6 Extracción de características

El objetivo de la extracción de características es extraer la información significativa de la señal de voz preprocesada. Las técnicas de extracción de características convierten la señal preprocesada en un conjunto de vectores de características y estos vectores caracterizan la naturaleza del habla. La Tabla 2.4 muestra la comparación de algunas de las técnicas de extracción de características.

Método	Banco de filtros	Técnica utilizada	Méritos	Deméritos
MFCC (Mel-Frequency Cepstral Coefficients)	Se utiliza el banco de filtros Mel para espectro de potencia.	Para la producción de coeficientes cepstrales, MFCC utiliza compresión de amplitud logarítmica y operación IDFT.	<ul style="list-style-type: none"> <li>Útil para varios oradores y múltiples lenguas.</li> <li>Confiable para vocabulario de moderado a gran tamaño.</li> <li>Es fácil de implementar.</li> </ul>	El ruido no se suprime.
PLP (Perception Linear Prediction)	Se utiliza escala de Bark que tiene forma de filtro trapezoidal.	Para estimar la sensibilidad del volumen auditivo dependiente de la frecuencia, PLP utiliza preénfasis de igual volumen y cepstrum se calcula mediante los coeficientes de predicción lineal (LP).	<ul style="list-style-type: none"> <li>Descartar información irrelevante del discurso y así mejorar el reconocimiento de voz.</li> <li>Velocidad.</li> </ul>	Da menos tasa de reconocimiento que MFCC y RASTA.
PNCC (Power-Normalized Cepstral Coefficients)	Los filtros de tono gamma son utilizados para simular el comportamiento de la cóclea.	PNCC incluye potencia de tiempo medio y eliminación de sesgos que se utiliza para aumentar la robustez. Se calcula utilizando la relación entre la media aritmética y geométrica para estimar la reducción en la calidad de habla causada por ruido.	<ul style="list-style-type: none"> <li>Tasa de precisión más alta en comparación a MFCC Y RASTA.</li> </ul>	La implementación es compleja.
LPC (linear Prediction Coding)	LPC Usa método de autocorrelación autorregresivo modelado para encontrar los coeficientes de filtro.	Obtiene los coeficientes de predicción lineal minimizando el error de predicción en el error de mínimos cuadrados.	<ul style="list-style-type: none"> <li>Requiere pocos recursos.</li> <li>Fácil implementación.</li> </ul>	<ul style="list-style-type: none"> <li>Incapaz de distinguir palabras con los mismos sonidos y vocales.</li> <li>Útil solo para un solo hablante y un solo idioma.</li> <li>Es confiable solo para pequeños tamaños de vocabulario.</li> </ul>

Método	Banco de filtros	Técnica utilizada	Méritos	Deméritos
RASTA (Relative Spectral Filtering)	Los filtros pasa banda son desplegados en el dominio del espectro logarítmico.	El filtro RASTA pasará por banda el coeficiente de todas las características debido a que el ruido se convierte en la parte aditiva. En pasa banda filtra los espectros resultantes, el ruido se suprime y el espectro se suaviza.	<ul style="list-style-type: none"> <li>• Útil para varios oradores y varios idiomas.</li> <li>• Confiable para tamaño moderado de vocabulario.</li> </ul>	Requiere implementación de moderada a difícil.
Spectral Entropy (Hermansky)	Se utiliza el banco de filtros rectangulares.	Para convertir el espectro en una función de distribución de probabilidad (PDF), los componentes de frecuencia individuales del espectro se separan y se dividen por la suma de todos los componentes. Esto asegura que el área PDF sea uno y pueda usarse para calcular la entropía.	La entropía espectral multibanda funciona muy bien con ruido de banda ancha aditivo y con niveles bajos de SNR.	Solo trabaja bien en reconocimiento de voz y no en otro tipo de problemas.
MSES (Multiband Spectral Entropy Signature)	Se utiliza el banco de filtros Mel para espectro de potencia.	Para calcular MSES, se utiliza la entropía de un proceso aleatorio. Se asume que la parte real e imaginaria de la transformada de Fourier discreta se comportan como dos variables aleatorias con media de cero.	La entropía espectral multibanda funciona muy bien con ruido de banda ancha aditivo y con niveles bajos de SNR. También se ha utilizado con éxito en diversas aplicaciones como en reconocimiento de música, reconocimiento de spots publicitarios, sonidos ambientales etc.	No supera a RASTA.

Tabla 2.4: Comparación de técnicas para extracción de características.

### 2.6.1 Coeficientes cepstrales de frecuencia Mel

MFCC son características basadas en espectros a corto plazo y su éxito ha sido debido a su capacidad para representar el espectro de amplitud en forma compacta. MFCC se basa en la escala de frecuencia no lineal de la percepción auditiva humana. que utilizan dos tipos de filtros, filtros espaciados linealmente y filtros espaciados logarítmicamente. La señal se expresa en la escala de frecuencia de Mel para capturar las características más importantes de un audio [34].

Para calcular MFCC, la señal de audio se divide en períodos de tiempo corto para extraer de cada uno un vector de características con coeficientes  $L$ . En este trabajo se calcula la Transformada de Fourier de Tiempo Corto para cada frame, que viene dada por (2.2), para  $k = 0, 1, \dots, N - 1$ , donde  $k$  corresponde a la frecuencia  $f(k) = kf_s/N$  y  $f_s$  es la frecuencia de muestreo en hercios. Aquí,  $x(n)$  denota un marco de longitud  $N$  y  $w(n)$  es la función de ventana de Hann, que viene dada por  $w(n) = 0.5 + 0.5\cos(2\pi n/N)$ .

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-\frac{i2\pi kn}{N}} \quad (2.2)$$

El proceso continúa escalando el espectro de magnitud  $|X(k)|$  tanto en frecuencia como en magnitud. Primero, la frecuencia se escala utilizando el banco de filtros de Mel  $H(k, m)$  y luego se toma el logaritmo usando (2.3),

$$X'(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|H(k, m)\right) \quad (2.3)$$

para  $m = 1, 2, \dots, M$ , donde  $M$  es el número de filtros y  $M \ll N$ . El banco de filtros Mel es un conjunto de filtros triangulares, donde las frecuencias en escala Mel del banco de filtros se calcula con  $\varphi = 2595\log_{10}(f/700 + 1)$ , que es una aproximación común.

Convencionalmente, los filtros de forma triangular de banda crítica residen en el rango de Nyquist. Las transformaciones de los filtros se hacen simétricas con respecto a la frecuencia de Nyquist. Como se muestra en la Figura 2.5, el banco de filtros del eje Mel está construido con 40 filtros no uniformes. Para tener una transición suave entre bandas críticas adyacentes y preservar la correlación entre ellas, el banco de filtros gaussianos también se desarrolla con 40 filtros no uniformes como se muestra en la Figura 2.6.

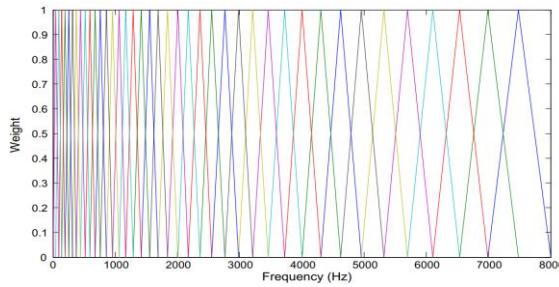


Figura 2.5: Estructura del banco de filtros triangulares [8].

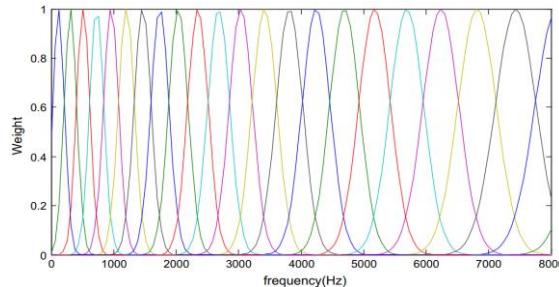


Figura 2.6: Estructura del banco de filtros Gaussianos [8].

Los MFCC se obtienen descorrelacionando el espectro  $X'(m)$  calculando la Transformada de coseno discreta usando (2.4),

$$c(l) = \sum_{m=1}^M X'(m) \cos\left[l \frac{\pi}{M} \left(m - \frac{1}{2}\right)\right] \quad (2.4)$$

para  $l = 1, 2, \dots, L$ , donde  $c(l)$  es el l-ésimo MFCC. Con este procedimiento, de cada frame se extrae un vector con coeficientes  $L$ .

Este trabajo, se centra en la Intelligent sound implementation (ISP) para informática MFCC [35], esta implementación considera un banco de filtros con logarítmico espaciado y amplitud constante, donde el número de filtros es personalizado para el parámetro.

## 2.6.2 Entropía espectral y entropía de Shannon

Cuando las señales de audio se degradan severamente, las características que las describen generalmente desaparecen, por lo tanto, el problema es encontrar las características que aún estarían presentes en la señal a pesar del nivel de degradación al que fue sometida. Los autores que se han centrado en este problema han explorado la entropía para caracterizar las señales de audio de la forma más sólida posible para diferentes tipos de degradaciones. En esta dirección, se empieza discutiendo sobre el concepto de entropía espectral y entropía de Shannon. En teoría de la información, la entropía de Shannon está relacionada con la incertidumbre de una fuente de información [36]. Por ejemplo, la entropía se usa para medir la predictibilidad de una señal aleatoria y el "pico" de una función de distribución de probabilidad. En investigación, es común usar (2.5) para medir, a través de la entropía, la cantidad de información que transporta la señal. Aquí,  $p_i$  es la probabilidad de que cualquier muestra de la señal tenga un valor  $i$  siendo  $n$  el número de valores posibles que pueden adoptar las muestras.

$$H = - \sum_{i=1}^n p_i \ln(p_i) \quad (2.5)$$

Se necesita alguna estimación de la función de distribución de probabilidad (PDF) para determinar la entropía de una señal, por lo tanto, se puede utilizar tanto métodos paramétricos como no paramétricos e histogramas. Si se eligen histogramas, se debe tener cuidado de que la cantidad de datos involucrados sea lo suficientemente alta como para evitar picos en el histograma. Cuando se habla de entropía espectral, es necesario revisar el trabajo de Shen [37], ya que ese concepto se introdujo por primera vez como una característica adicional para la detección de puntos finales (detección de actividad de voz). La idea de la entropía espectral compromete a considerar el espectro de una señal como un PDF para capturar los picos del espectro y su ubicación. Para convertir el espectro en un PDF, los componentes de frecuencia individuales del espectro se separan y se dividen por la suma de todos los componentes, a saber,  $p_k = X(k) / \sum_{i=1}^N X(i)$ , para  $k = 1, 2, \dots, N$ , donde  $X(k)$  es la energía del  $k$ -ésimo componente del espectro de frecuencia,  $p = (p_1, \dots, p_N)$  es la PDF del espectro y  $N$  es el número total de frecuencias componentes del espectro. Esto asegura que el área PDF sea uno y pueda usarse para calcular la entropía. El concepto de entropía espectral multibanda fue introducido por [38], y consiste en dividir el espectro en subbandas de igual tamaño para calcular la entropía en cada una de ellas usando (2.5), donde cada espectro de subbanda debe asumirse como un PDF. Además, [39] demostró que la entropía espectral multibanda funciona muy bien con ruido de banda ancha aditivo y con niveles bajos de SNR.

## 2.6.3 Firma de entropía espectral multibanda

Basado en la idea presentada por Misra et al. [38], [39], se utiliza el concepto de entropía espectral para obtener una firma robusta que se puede utilizar en diferentes problemas de reconocimiento de audio [40]–[44]. A diferencia de Misra et al., se calcula la entropía en cada subbanda utilizando la entropía de un proceso aleatorio. Sea  $x = [x_1, x_2, \dots, x_n]^T$  un vector de  $n$  variables aleatorias de valor real, entonces, se dice que  $x$  es un vector aleatorio gaussiano donde se dice que las variables

aleatorias  $x_i$  son conjuntamente gaussianas si la función de densidad de probabilidad conjunta de las  $n$  variables aleatorias  $x_i$  viene dada por  $p(x) = N(\mu_x, \Sigma_x)$ , donde  $\mu_x = [m_1, m_2, \dots, m_n]^T$  es un vector que contiene las medias de  $x_i$ , este es,  $m_i = E[x_i]$ .  $\Sigma_x$  es una matriz definida positiva simétrica con elementos  $\sigma_{ij}$  que son las covarianzas entre  $x_i$  y  $x_j$ , esto es,  $\sigma_{ij} = E[(x_i - m_i)(x_j - m_j)]$ .

Tomando algunas precauciones, la entropía de un vector aleatorio gaussiano se puede determinar usando la versión continua de la entropía de Shannon, que viene dada por (2.6).

$$H(x) = - \int_{-\infty}^{+\infty} p(x) \ln[p(x)] dx \quad (2.6)$$

Si se asume que el vector aleatorio sigue una distribución gaussiana con media cero y matriz de covarianza,  $N(0, \Sigma_x)$ , luego reemplazando  $p(x)$  en (2.6), se obtiene esta ecuación conocida para determinar la entropía de un vector en un proceso aleatorio [45], como se muestra en (2.7), donde  $|\Sigma_x|$  es el determinante de la matriz de covarianza.

$$H(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma_x|) \quad (2.7)$$

Para calcular MSES, la señal de audio debe dividirse en frames para extraer de cada uno un vector con  $L$  coeficientes de entropía. A continuación, se calcula la transformada de Fourier de tiempo corto en cada frame utilizando (2.2). Para dividir el espectro de banda completa en sub bandas, se tiene en cuenta la idea de cómo las personas identifican los sonidos. El oído humano percibe mejor las frecuencias más bajas que las más altas, pero no todas las frecuencias se pueden escuchar con la misma sensibilidad. Este proceso se puede modelar en todo el ancho de banda de la respuesta del oído utilizando la escala de Bark, que se divide en 25 bandas críticas [46], [47]. La Tabla 2.5 muestra las primeras 24 bandas críticas con sus respectivos anchos de banda.

Critical Band	Lower cut-off (Hz)	Central Frequency (Hz)	Higher cut-off (Hz)	Bandwidth (Hz)
1	0	50	100	100
2	100	150	200	100
3	200	250	300	100
4	300	350	400	100
5	400	450	510	110
6	510	570	630	120
7	630	700	770	140
8	770	840	920	150
9	920	1000	1080	160
10	1080	1170	1270	190
11	1270	1370	1480	210
12	1480	1600	1720	240
13	1720	1850	2000	280
14	2000	2150	2320	320
15	2320	2500	2700	380
16	2700	2900	3150	450
17	3150	3400	3700	550
18	3700	4000	4400	700
19	4400	4800	5300	900
20	5300	5800	6400	1100
21	6400	7000	7700	1300
22	7700	8500	9500	1800
23	9500	10500	12000	2500
24	12000	13500	15500	3500

Tabla 2.5: Bandas críticas para la escala Bark.

Se utiliza (2.8) para cambiar de Hertz a Barks, donde  $f$  es la frecuencia en Hertz.

$$\text{Barks} = 13\tan^{-1}\left(\frac{0.75f}{1000}\right) + 3.5\tan^{-1}\left[\left(\frac{f}{7500}\right)^2\right] \quad (2.8)$$

El proceso continúa calculando la entropía para cada una de las bandas críticas por (2.7). Se consideró para cada sub-banda que los coeficientes espectrales se distribuyen normalmente. Esta consideración se debe a que una buena estimación de la PDF no se puede determinar utilizando métodos no paramétricos, ya que las bandas más bajas del espectro tienen muy pocos coeficientes. Para calcular la entropía, se consideró un proceso aleatorio con dos variables aleatorias. Se asume que las partes reales e imaginarias de los coeficientes espectrales son variables aleatorias con una distribución normal y media cero, por lo tanto, para el caso bidimensional la entropía está determinada por  $H = \ln(2\pi) + (1/2)\ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)$ , donde  $\sigma_{xx}$  y  $\sigma_{yy}$  son las varianzas de las partes real e imaginaria, respectivamente, y  $\sigma_{xy}$  es la covarianza entre las partes real e imaginaria. El resultado de este proceso es una matriz  $L \times T$  (denominada como firma), donde  $L$  es el número de coeficientes de entropía y  $T$  denota el número de frames. Esta firma captura el nivel de contenido de información para cada banda crítica y posición de frame en el tiempo [48].

#### 2.6.4 Transformada discreta de Fourier

El espectro en frecuencia de una señal discreta se puede estimar mediante la transformada discreta de Fourier (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}} \quad (2.9)$$

Esta transformada proporciona un conjunto de  $N$  valores en el dominio de la frecuencia, en donde  $X(k)$  es la transformada discreta de Fourier en el dominio de la frecuencia, de la señal discreta  $x(n)$  en el dominio del tiempo. Sin embargo, en la práctica, la transformada discreta de Fourier se calcula de una manera mucho más eficiente mediante la transformada rápida de Fourier (FFT) de  $N$  puntos [49].

#### 2.6.5 RASTA-MFCC

El filtrado RASTA se aplica a la señal de voz en ventana para minimizar los efectos de ruido en la señal de voz, especialmente los efectos de ruido de convolución y aditivo [50]. El filtrado es seguido por la extracción de MFCC de la señal filtrada RASTA para producir características RASTA–MFCC. Los pasos seguidos para obtener la característica RASTA-MFCC se muestran en la Figura. 2.7.

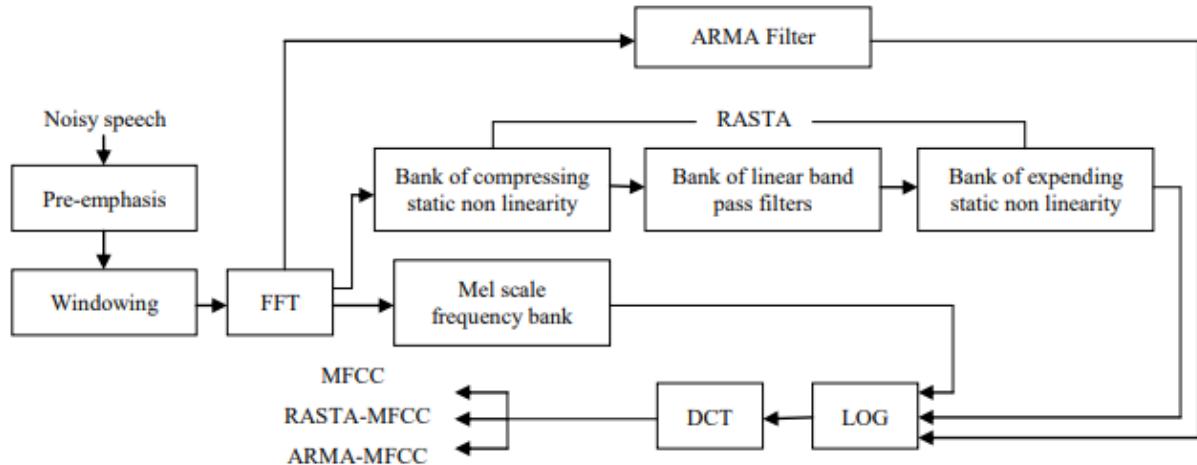


Figura 2.7: Diagrama a bloques para extraer MFCC, RASTA-MFCC y ARMA-MFCC [51].

El procesamiento RASTA mejora el rendimiento de un reconocedor en condiciones ruidosas. El procesamiento RASTA compensa el efecto del cambio espectral abrupto en la señal de voz mediante el filtrado. Los cambios espectrales rápidos en los frames consecutivos se alivian mediante el filtrado de paso bajo [52]. mediante un proceso de suavizado. En general, cuanto mayores son las estructuras auditivas, mayor es la sensibilidad a las frecuencias más bajas del habla / sonido. En la familia de los mamíferos, los seres humanos tienen relativamente menos sensibilidad a los sonidos de baja frecuencia. El procesamiento RASTA implica el cálculo del espectro de potencia de la banda crítica, filtrando la trayectoria temporal del componente espectral comprimido, transformación no lineal estática seguida de multiplicación con curvas de sonoridad iguales. Finalmente calcula el modelo de todos los polos del espectro. La frecuencia de corte más baja del filtro determina el cambio espectral más rápido, mientras que la frecuencia de corte más alta determina el cambio espectral preservado.

La técnica relativa espectral (RASTA) [50] suprime los componentes espectrales que cambian más lenta o rápidamente que el rango típico de cambio del habla. La Figura 2.8 ilustra el proceso de RASTA. Su función de transferencia del filtro IIR se muestra como

$$H(z) = 0.1z^4x \frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}} \quad (2.10)$$

La frecuencia de corte baja de este filtro determina el cambio espectral más rápido del espectro logarítmico, que se ignora en la salida, mientras que la frecuencia de corte alta determina el cambio espectral más rápido que se conserva en los parámetros de salida. Cuando opera en el dominio espectral logarítmico, RASTA disminuye de manera efectiva los componentes espectrales que son aditivos en el dominio espectral logarítmico, en particular, las características espirales fijas o que cambian lentamente en el entorno. Estos componentes espirales son convolutivos en el dominio del tiempo y, por lo tanto, aditivos en el dominio logarítmico espectral o cepstral.

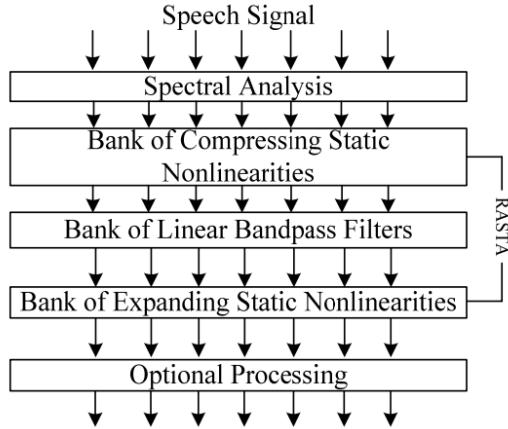


Figura 2.8: Diagrama a bloques de RASTA [53].

Sin embargo, los componentes de ruido aditivo no correlacionados que son aditivos en el dominio espectral de potencia se vuelven dependientes de la señal después de la operación logarítmica en el espectro y no pueden eliminarse eficazmente mediante el filtrado de paso de banda RASTA en el dominio logarítmico. Por lo tanto, el procesamiento RASTA original en el espectro logarítmico o cepstrum no es particularmente apropiado para el hablar con ruido aditivo significativo. La ecuación (2.11) se propone como un sustituto de la transformada logarítmica de RASTA como en [54].

$$y = \ln(1 + Jx) \quad (2.11)$$

donde  $J$  es una constante positiva dependiente de la señal. La transformada de deformación de amplitud es de tipo lineal para  $J \ll 1$  y de tipo logarítmico para  $J \gg 1$ . La transformada inversa exacta de la ecuación (2.11)

$$x = \frac{e^y - 1}{J} \quad (2.12)$$

donde  $e$  es la base del logaritmo natural. La inversa aproximada  $x = \frac{e^y}{J}$  se usa para asegurar que el valor de  $x$  sea positivo para todo  $y$  [50].

## 2.7 Tipos de clasificadores

Los vectores de características obtenidos de la técnica de extracción de características se comparan con el vector de características de prueba para averiguar el índice de similitud. Basado en las similitudes de los vectores de características obtenidos, el clasificador de patrones toma la decisión de aceptar o rechazar las muestras. La Tabla 2.6 muestra la comparación de varias técnicas de clasificación.

Clasificador	Descripción	Método de aprendizaje	Mérito	Demérito
DTW (Dynamic Time Warping)	Se utiliza principalmente para encontrar las similitudes entre secuencias basadas en dos tiempos utilizando medidas de distancia normalizadas en el tiempo. El que tiene mínima distancia está clasificado como el reconocido correctamente.	Método de aprendizaje sin supervisión.	<ul style="list-style-type: none"> <li>• Requiere menos espacio de almacenamiento.</li> <li>• Beneficioso para longitud variable.</li> </ul>	Hay problemas entre canales.
HMM (Hidden Markov Model)	HMM es un proceso estocástico. En el modelo Hidden Markov, el estado anterior no es directamente visible para el observador. Pero dependiendo de ese estado la salida es visible.	Método de aprendizaje sin supervisión.	Rendimiento eficiente en comparación con DTW.	<ul style="list-style-type: none"> <li>• Requiere más espacio de almacenamiento computacional.</li> <li>• Más complejo que DTW.</li> </ul>
GMM (Gaussian Mixture Model)	GMM es un método estadístico para estimar el parámetro espectral usando el algoritmo de maximización de expectativas.	Método de aprendizaje sin supervisión.	Requiere menos datos de entrenamiento y prueba.	Hay una compensación en el rendimiento entre el de DTW y HMM.
VQ (Vector Quantization)	VQ viene bajo la técnica de compresión con pérdida. VQ proporciona representación multidimensional de datos. El límite de decisión y los niveles de reconstrucción son dos términos importantes utilizados en VQ.	Método de aprendizaje sin supervisión.	<ul style="list-style-type: none"> <li>• Computacionalmente complejo.</li> <li>• Utilización eficaz del tiempo.</li> </ul>	La codificación en tiempo real es compleja.
SVM (Support Vector Machine)	Lo básico de SVM es crear un hiperplano. Este hiperplano diferencia las características. En SVM binario, las características son clasificadas en dos clases, una clase para oradores reconocidos y otra para no reconocidos.	Método de aprendizaje supervisado.	Operación simple.	SVM binario tiene limitación en el reconocimiento de oradores.
Neural Network (NN) Modeling Technique	Se utilizan para resolver tareas de identificación complejas.	Puede utilizar aprendizaje supervisado o no supervisado.	<ul style="list-style-type: none"> <li>• Puede controlar la baja calidad de señal de voz.</li> <li>• Puede usarse con datos ruidosos.</li> <li>• Proporciona una mayor precisión que HMM.</li> </ul>	La selección de la configuración óptima no es fácil.

Clasificador	Descripción	Método de aprendizaje	Mérito	Demérito
MLP	Las MLP son redes de retroalimentación en capas normalmente entrenadas con retro propagación estática para clasificar patrones estáticos.	Método de aprendizaje supervisado.	<ul style="list-style-type: none"> <li>• Esta red se puede construir a mano, crear mediante un algoritmo o ambos.</li> <li>• La red también se puede monitorear y modificar durante el tiempo de entrenamiento.</li> </ul>	MLP se entrena lentamente y requiere muchos datos de entrenamiento.
KNN (K Nearest Neighbors)	Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento $X$ pertenezca a la clase $C_j$ a partir de la información proporcionada por el conjunto de prototipos.	Método de aprendizaje supervisado.	<ul style="list-style-type: none"> <li>• La precisión es bastante alta pero no competitiva en comparación con modelos de aprendizaje mejor supervisados.</li> <li>• La precisión puede verse afectada por el ruido o las características irrelevantes.</li> </ul>	Computacionalmente costoso, ya que requiere mucha memoria.

Tabla 2.6: Comparación de varias técnicas de clasificación [7].

### 2.7.1 MLP

Citando a Wikipedia, "El aprendizaje automático es un subcampo de la informática que se desarrolló a partir del estudio de reconocimiento de patrones y la teoría del aprendizaje computacional en la inteligencia artificial. El aprendizaje automático explora el estudio y construcción de algoritmos que pueden aprender y hacer predicciones sobre los datos". Este tipo de aprendizaje, se basa en redes neuronales profundas (en inglés, Deep Neural Nets - DNN), que se asemejan a redes de neuronas biológicas. Los bloques de construcción de la DNN son los perceptrones, que son el equivalente a las neuronas. La información es procesada por una función de transferencia, resumiendo las señales de entrada, y una función de activación que decide si la salida está activa o no.

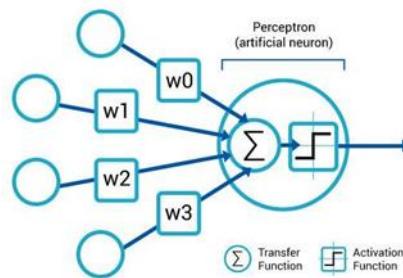
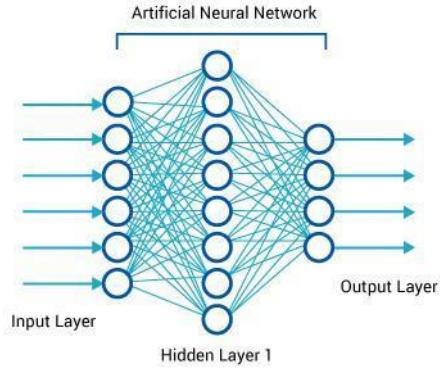


Figura 2.9: Activación de un perceptrón o neurona artificial [55].

En una red neuronal artificial, un algoritmo ajusta los valores de las variables internas que conectan las neuronas digitalmente. Por ejemplo, un "cero" significa que las neuronas no están conectadas, mientras que un "uno" representa la conexión entre dos neuronas.

Con el tiempo, la red "aprende" a reconocer correctamente los patrones mediante el cálculo de los valores de los parámetros internos. En un sentido, la red escribe sus propias reglas implícitas cuando se le proporcionan los datos de entrada y de salida de las etiquetas. Una vez completada la fase de aprendizaje, el conjunto de valores y la red neuronal digitales se pueden trasladar al sistema que ejecuta la aplicación.



*Figura 2.10: Red neuronal artificial de tipo perceptrón multicapa [55].*

En la Figura 2.10 se ilustra una red neuronal artificial de tipo Perceptrón Multicapa (MLP). Esta puede entenderse básicamente como una red de múltiples neuronas artificiales en múltiples capas. Aquí, la función de activación no es lineal, pero se utiliza una función de activación no lineal como el sigmoide logístico o la tangente hiperbólica, o una función de activación lineal por partes como la unidad rectificadora lineal (ReLU). Además, a menudo se usa una función softmax (una generalización del sigmoide logístico para problemas de clase múltiple) en la capa de salida, y una función de umbral para convertir las probabilidades predichas (por el softmax) en etiquetas de clase. Las redes neuronales artificiales pueden tener capas ocultas y diversas unidades de salida.

En este punto, se puede decir que la ventaja del MLP sobre el clásico Perceptrón y Adaline es que al conectar las neuronas artificiales a través de funciones de activación no lineales, se pueden crear límites de decisión complejos, no lineales, que nos permiten abordar problemas donde las diferentes clases no son separables linealmente [55].

### 2.7.2 KNN

El algoritmo de k vecinos más cercanos (k-NN), se divide en dos partes: la parte del entrenamiento y la parte de la clasificación [56], [57]. K-NN es una técnica que, dada una instancia a clasificar, se obtienen los k vecinos, y aplicándolos una función de distancia, determinará a qué clase pertenece de acuerdo a los vecinos más cercanos [58]. Es decir, esta técnica solo recuerda los ejemplos que se vieron en la etapa de entrenamiento. Los nuevos casos se clasifican según el comportamiento del dato más cercano [56], [59]. Este algoritmo tiene tres propiedades, la primera propiedad es que se trata de un algoritmo de aprendizaje perezoso. La segunda propiedad es que clasifica nuevos objetos comparando con objetos similares e ignora los que son distintos. La tercera propiedad es representar a los objetos como puntos de valores reales en un espacio euclíadiano de n dimensiones [60]. Para evitar el ruido que se puede presentar en el k-vecinos más cercanos, al clasificar nuevos objetos, se debe aumentar el número de vecinos(k), de este modo, al aumentar k, se asocia más rápido el nuevo dato a los elementos más representativos o con mayor presencia en el espacio [61]. Cabe señalar que el valor de k siempre debiera de ser un número impar [62]. Ver Figura 2.11.

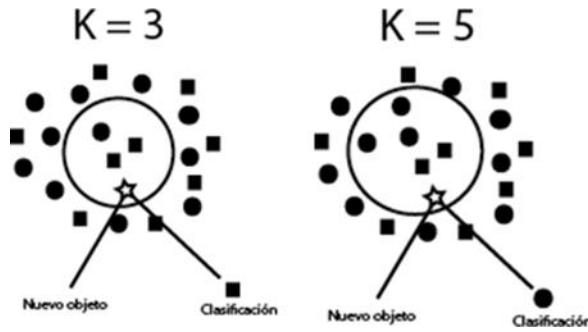


Figura 2.11: Ruido en clasificación [63].

En este punto, se puede decir que aumentar el número de vecinos o  $k$  al máximo no siempre es bueno, porque tiene que haber un balance entre su valor, ya que al aumentar su valor podríamos tener mayor presencia de una clase. Es decir, al algoritmo  $k$  vecinos le afecta el desbalance de clases, lo que causa errores de confiabilidad al clasificar los objetos [64], [65]. Esta clasificación también tiene sus desventajas ya que el costo de clasificar nuevos objetos suele ser muy alto [66], esto pasa por que el proceso se hace cuando se está clasificando y no cuando se está entrenando.

### 2.7.3 SVM

La tarea de entrenamiento de los SVM involucra la optimización de una función convexa, por ende, no hay mínimos locales que “compliquen” el proceso de aprendizaje. El enfoque tiene muchos otros beneficios, por ejemplo, el modelo construido tiene una dependencia explícita en los patrones de mayor aporte en los datos (los Support Vectors). Además, permite generalizar de mejor forma frente a nuevos objetos [67] dado que considera el principio de “minimización del riesgo estructural”.

La clasificación mediante Support Vector Machines permite obtener clasificadores lineales y no lineales.

#### Descripción del modelo

La técnica de Support Vector Machines fue propuesta por Vapnik [68], [69]. Ésta se basa en encontrar un hiperplano de separación que divide el espacio de entrada en dos regiones. Cada una de estas regiones corresponderá a una de las clases definidas, como se muestra en la Figura 2.12:

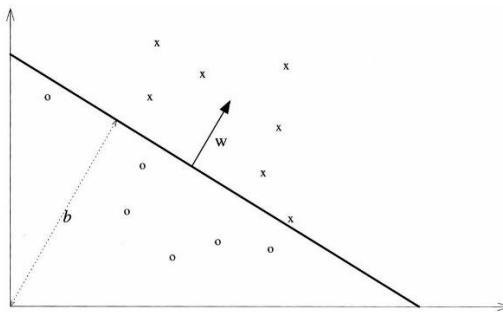


Figura 2.12: Un Hiperplano de clasificación  $\{w^{\rightarrow}, b\}$  para un conjunto de entrenamiento de dos dimensiones [70].

Por un lado, los hiperplanos que están más alejados de las fronteras de las clases de objetos corresponden mayores márgenes de separación. Por otro, hiperplanos que aciertan más en la asignación de objetos a las clases a las que efectivamente pertenecen, tienen un menor error de clasificación. Por lo tanto, un hiperplano de separación ideal debe maximizar el margen de separación

y minimizar el error de clasificación. Sin embargo, no siempre es posible cumplir los dos objetivos simultáneamente. Para salvar esta dificultad se plantea un problema de optimización cuya función objetivo combina ambos propósitos. Este problema de optimización resulta ser un problema de minimización cuadrático convexo [71]. En el caso que el número de objetos a clasificar es mayor que el número de atributos de cada objeto, lo que usualmente sucede, este problema tiene una única solución óptima. Lo descrito anteriormente corresponde al caso que existe un hiperplano de separación de las clases. En ese caso se dice que las clases son linealmente separables. Recientemente, el estudio de los SVM se ha extendido al caso de clases que no son linealmente separables mediante la introducción de las llamadas funciones Kernel [72], [73] o de variables de pérdida u holgura [74]. Esta metodología ha sido también extendida para problemas de regresión [75].

#### 2.7.4 Métricas

Para poder evaluar el rendimiento de los clasificadores utilizados, se usan métricas [76], las cuales son adaptadas en función de los casos correctamente e incorrectamente clasificados. En general, las métricas consideran los siguientes casos:

- True Positives (*tp*): elementos a los que el clasificador asignó la clase relevante y esta era correcta.
- False Positives (*fp*): elementos a los que el clasificador asignó la clase relevante y esta no era correcta.
- False Negatives (*fn*): elementos a los que el clasificador asignó la clase no-relevante y esta no era correcta.
- True Negatives (*tn*): elementos a los que el clasificador asignó la clase no-relevante y esta era correcta.

Esto quiere decir:

	Relevantes	No Relevantes
Seleccionados	<i>tp</i>	<i>fp</i>
No Seleccionados	<i>fn</i>	<i>tn</i>

Tabla 2.7: Evaluación de los clasificadores.

Es así cómo es posible realizar el cálculo del rendimiento de acuerdo a las siguientes métricas:

- Accuracy: Representa la porción de documentos que son clasificados correctamente sobre el total de casos (Ecuación 2.13).

$$Accuracy_{TOTAL} = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.13)$$

- Precisión: En ella se representa la porción de documentos que son clasificados correctamente para la clase A sobre el total de casos clasificados como clase A (Ecuación 2.14).

$$Precisión_A = \frac{tp}{tp + fp} \quad (2.14)$$

- Recall: Esta métrica representa la porción de documentos de clase A que son clasificados correctamente (Ecuación 2.15).

$$Recall_A = \frac{tp}{tp + fn} \quad (2.15)$$

- F-Measure: También llamada Medida-F, combina las medidas de precisión y recall a partir de la media armónica ponderada de estos dos valores (Ecuación 2.16).

$$F_1 = \frac{2PR}{P + R} \quad (2.16)$$

## CAPITULO 3 METODOLOGIA

En esta capítulo se presenta la metodología seguida para realizar los experimentos de esta tesis, todo el proceso se realizó en Python, por lo tanto, todos los pasos de la metodología vienen ejemplificados en Python. Como en ambas bases de datos se sigue el mismo proceso, solo se muestra exemplificación con código para una sola base de datos, solo en los casos en los que existen diferencias entre las bases de datos se muestra exemplificación con código para ambas bases de datos.

### 3.1 Análisis y división de datos

El primer paso de la metodología consiste en analizar y dividir los archivos de audio de las bases de datos, en el análisis de datos se determina el preprocesamiento necesario para poder extraer las características correctamente y la división de datos se realiza para tener conjuntos equilibrados de prueba y entrenamiento para los clasificadores.

Para analizar archivos de audio en Python se utilizan las siguientes bibliotecas:

- **IPython.display.Audio:** Permite reproducir el audio directamente en el IDE de Python.
- **Librosa:** Permite cargar el audio como un array numérico para su análisis y manipulación.
- **Librosa.display y matplotlib:** Permiten graficar la forma de onda del archivo de audio.

#### 3.1.1 Inspección auditiva

Inspeccionando los datos auditivamente se pueden encontrar archivos dañados o con bajo nivel de volumen, estos errores tienen que ser identificados y solucionados, ya que, dificultan la extracción de características.

A continuación, se muestra la codificación en Python para reproducir los archivos de audio y así poder inspeccionar auditivamente.

```
import IPython.display as ipd  
ipd.Audio('../EMODB/wav/03a04Ad.wav')
```

#### 3.1.2 Inspección visual

Inspeccionando los datos visualmente se pueden identificar zonas de silencio o de baja amplitud que dificultan la extracción de características.

A continuación, se muestra la codificación en Python para graficar la forma de onda de los archivos de audio y así poder inspeccionar visualmente.

```
# Cargar las importaciones  
import IPython.display as ipd  
import librosa  
import librosa.display  
import matplotlib.pyplot as plt  
  
filename = '../EMODB/wav/03a04Ad.wav'  
plt.figure(figsize=(12,4))  
data,sample_rate = librosa.load(filename)  
_ = librosa.display.waveplot(data,sr=sample_rate)  
ipd.Audio(filename)
```

A continuación, se muestra la forma de onda de 7 archivos, cada uno correspondiente a una clase de la base de datos EMODB.

Nótese que, aunque en las 7 formas de onda se está diciendo la misma frase, se observa que es difícil identificar las clases. Por las variaciones de energía de la señal y zonas de silencio que estas presentan a lo largo del tiempo.

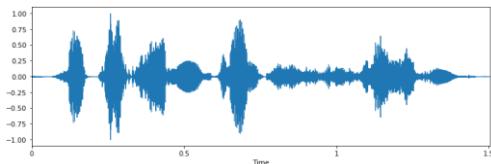


Figura 3.1: Forma de onda del archivo 03a04Ad.wav de la clase angustia de la base de datos EMODB.

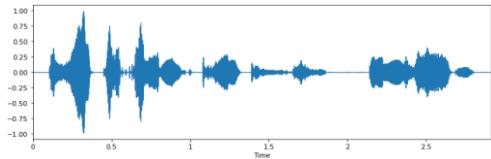


Figura 3.2: Forma de onda del archivo 14a04Ed.wav de la clase disgusto de la base de datos EMODB.

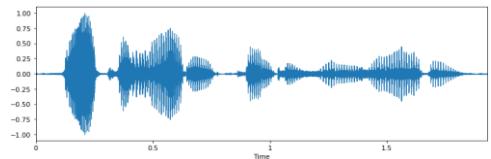


Figura 3.3: Forma de onda del archivo 03a04Lc.wav de la clase aburrimiento de la base de datos EMODB.

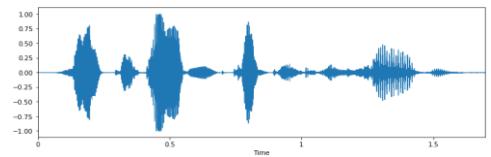


Figura 3.4: Forma de onda del archivo 03a04Fd.wav de la clase felicidad de la base de datos EMODB.

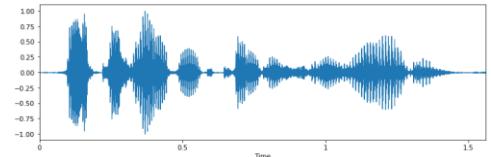


Figura 3.5: Forma de onda del archivo 03a04Nc.wav de la clase neutral de la base de datos EMODB.

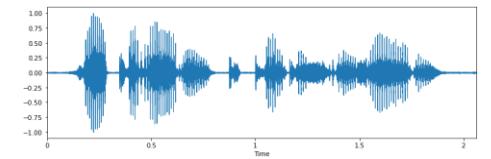


Figura 3.6: Forma de onda del archivo 03a04Ta.wav de la clase tristeza de la base de datos EMODB.

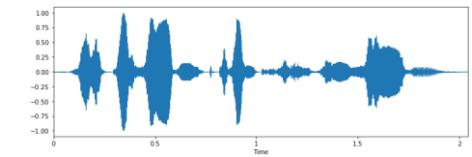


Figura 3.7: Forma de onda del archivo 03a04Wc.wav de la clase irá de la base de datos EMODB.

### 3.1.3 Propiedades de los archivos de audio

Es importante analizar las propiedades de los archivos de audio de las bases de datos, ya que, esto ayuda a identificar las propiedades de los archivos que deben ser normalizadas en la etapa de preprocesamiento.

Los archivos de audio de las bases de datos EMODB y EMOVO tienen formato wav, utilizan 16 bits de resolución y difieren en frecuencia de muestreo y numero de canales de audio; 16 kHz y un canal en el caso de EMODB y 48 kHz y 2 canales en el caso de EMOVO.

A continuación, se muestra la codificación en Python para iterar a través de cada uno de los archivos de muestra de audio y extraer, número de canales de audio, frecuencia de muestreo y resolución de bits.

```
# Cargar varias importaciones
import pandas as pd
import os
import librosa
import librosa.display
from helpers.wavfilehelper import WavFileHelper

wavfilehelper = WavFileHelper()
audiodata_EmoDB = []
metadata = pd.read_csv('../metadata/EMODB.csv')

for index, row in metadata.iterrows():

    file_name = os.path.join(os.path.abspath('..\EMODB\audio'),
    ,+str(row["fold"])+',',str(row["slice_file_name"]))
    data = wavfilehelper.read_file_properties(file_name)
    audiodata_EmoDB.append(data)

#Convertir en un Panda dataframe
audiodf_EmoDB = pd.DataFrame(audiodata_EmoDB,
columns=['num_channels','sample_rate','bit_depth'])
```

- **Canales de audio**

Todas las muestras de la base de datos EMODB tienen un solo canal, mientras que el 99.8% de las muestras de la base de datos EMOVO tienen 2 canales, por lo que las muestras de la base de datos EMOVO tendrán que ser normalizadas en la etapa de preprocesamiento para que tengan solo un canal.

A continuación, se muestra la codificación en Python para obtener el numero de canales de los archivos de audio de las bases de datos EMODB y EMOVO.

```
# Número de canales EMODB
print(audiodf_EmoDB.num_channels.value_counts(normalize=True))
1    1.0
# Número de canales EMOVO
print(audiodf_emovo.num_channels.value_counts(normalize=True))
2    0.998299
1    0.001701
```

- **Tasa de muestreo**

Todas las muestras de la base de datos EMODB tienen una taza de muestreo de 16 KHz, mientras que las muestras de la base de datos EMOVO tienen una taza de muestro de 48 KHz.

A continuación, se muestra la codificación en Python para obtener la taza de muestreo de los archivos de audio de las bases de datos EMODB y EMOVO.

```

# Tasa de muestreo EMODB
print(audiodef_EmoDB.sample_rate.value_counts(normalize=True))
16000    1.0

# Tasa de muestreo EMOVO
print(audiodef_EMOVO.sample_rate.value_counts(normalize=True))
48000    1.0

```

- **Resolución de bit**

La resolución de bit es uniforme para las muestras en ambas bases de datos

A continuación, se muestra la codificación en Python para obtener la resolución de bit de los archivos de audio de la base de datos EMODB.

```

#resolución de bit
print(audiodef_EmoDB.bit_depth.value_counts(normalize=True))
16    1.0

```

### 3.1.4 Distribuciones de clases

Es importante conocer la distribución de clases de cada conjunto de datos, para así, saber cómo balancear los conjuntos de prueba y entrenamiento, a continuación, en las figuras 3.8 y 3.9 se muestra la distribución de clases para los conjunto de datos EMODB y EMOVO respectivamente, como se puede observar el conjunto de datos EMOVO esta balanceado contando con 84 muestras en sus 7 clases, y el conjunto EMODB esta desbalanceado, siendo su clase con más muestras ira y su clase con menos muestras disgusto.

ira	127
aburrimiento	81
neutral	79
angustia	73
alegría	71
tristeza	62
disgusto	42

Figura 3.8: Distribución de clases de EMODB.

```

neutral     84
tristeza   84
disgusto   84
Sorpresa   84
angustia   84
ira        84
alegría    84

```

*Figura 3.9: Distribución de clases de EMOVO.*

### 3.1.5 Dividir el conjunto de datos

Para dividir los datos en conjuntos de prueba y entrenamiento, se creó un archivo de metadatos para cada base de datos, a continuación, se muestra cómo se carga en Python el archivo de metadatos EMODB.csv en un dataframe de Panda, se utiliza la columna if de los metadatos para dividir el conjunto de datos en conjuntos de entrenamiento y prueba. El tamaño del conjunto de prueba es del 30% y se estableció el criterio de balancear para que cada actor tuviera aproximadamente el 70% de sus interpretaciones de cada clase en el conjunto de entrenamiento.

```

import pandas as pd
metadata = pd.read_csv('../metadata/EMODB.csv')
metadata.head()

```

	slice_file_name	fold	class_name	if
0	03a04Ad.wav	A	A angustia	test
1	03a05Aa.wav	A	A angustia	train
2	03b02Aa.wav	A	A angustia	train
3	03b10Ab.wav	A	A angustia	train
4	08a01Ab.wav	A	A angustia	test

*Figura 3.10: Metadatos de EMODB*

## 3.2 Preprocesamiento

Ya habiendo analizado y dividido los datos lo siguiente es la etapa de preprocesamiento. En la Figura 3.11 se pueden observar los bloques que forman parte de la etapa de preprocesamiento, aquí es donde se prepara la señal para poder extraer características.

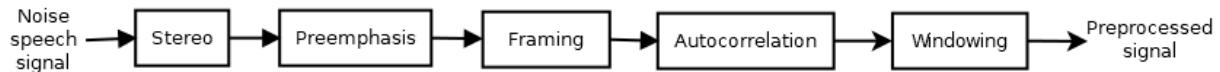


Figura 3.11: Diagrama a bloques del preprocesamiento de la señal del habla.

A continuación, se explica cada proceso de la Figura 3.11.

### 3.2.1 Stereo

Una señal de audio stereo es una señal que se compone de dos canales de audio diferentes, como se muestra en la Figura 3.12, esta señal tiene que ser normalizada a un solo canal para poder extraer características de ella.

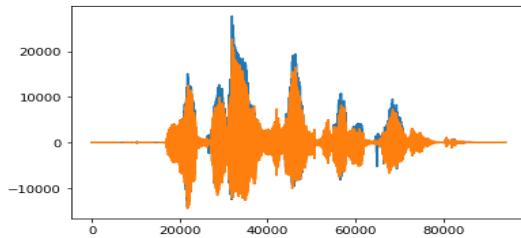


Figura 3.12: Grafica de la señal de audio stereo gio-f1-b1.wav

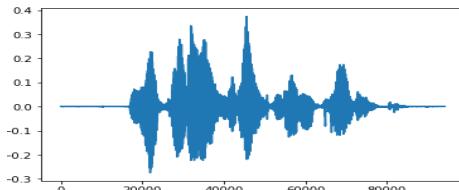


Figura 3.13: Grafica de la señal de audio monoaural gio-f1-b1.wav.

A continuación, se muestra la codificación en Python de la función `stereo`, donde se observa cómo se promedian los canales para convertir la señal de audio de stereo a monoaural, como se observa en la Figura 3.13.

```
def __Stereo(self):
    n_channel = len(np.shape(self.audio))
    if n_channel > 1:
        temp_1 = self.audio[:,0]
        temp_2 = self.audio[:,1]
        avg = (temp_1 + temp_2)/2
    else:
        avg = self.audio
```

### 3.2.2 Preemphasis

El bloque de preemphasis resalta las altas frecuencias que normalmente son de poca energía, en la Figura 3.14 se puede observar como actua el bloque de preemphasis en la señal de la figura 3.13.

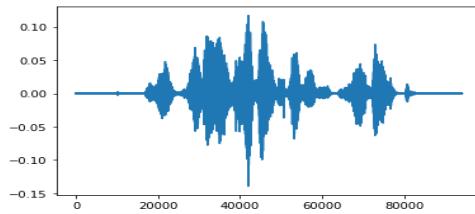


Figura 3.14: Grafica de la aplicación del filtro de preemphasis a la señal de audio gio-f1-b1.wav.

A continuación, se muestra la codificación en Python de la función preemphasis.

```
def __Preemphasis(self):
    Avg = np.zeros(len(avg))
    Avg[0] = avg[0]
    for n in range(1, len(avg)):
        Avg[n] = avg[n] - self.preemphasis*avg[n-1]
    return Avg
```

### 3.2.3 Framing

El habla es una señal casi estacionaria y por lo tanto, el procesamiento del habla debe realizarse en un análisis de tiempo corto el cual consiste en extraer de la señal de voz frames de longitud fija (número de muestras en el frame) que se superponen entre sí. Para lograr esto, la señal se divide en frames de 20-30 ms y superposición de 30-50% de cada frame con frame adyacentes.

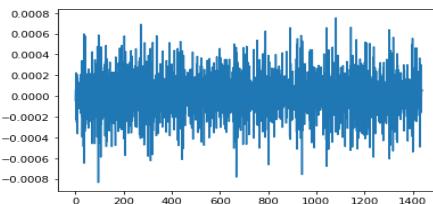


Figura 3.15: Framing, 50% traslape entre frames de 30ms.

A continuación, se muestra la codificación en Python del proceso de framing.

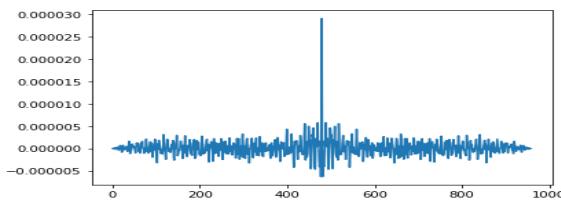
```
#Lee archivo de audio
self.fs, self.signal = rd.read(file_name)
#Tamaño de frame en segundos
self.frame=0.03
#Porcentaje de traslape
self.overlap = 50
self.n_frames = self.__Frames()
self.frame_size = np.fix(self.frame*self.fs)
self.overlap_size = np.fix(self.frame_size - self.frame_size*(self.overlap/100.0))
#Determina número de frames para analizar
def __Frames(self):
    i = 0
    j = 0
    while(j < self.audio_length):
        j = i*self.overlap_size + self.frame_size
        i = i + 1
    return i - 1
#Análisis de frames
for i in range(0,int(self.n_frames)):
    #Extrae frame de la señal
    frame = self.audio_avg[i*int(self.overlap_size):i*int(self.overlap_size)+int(self.frame_size)]
```

### 3.2.4 Autocorrelación

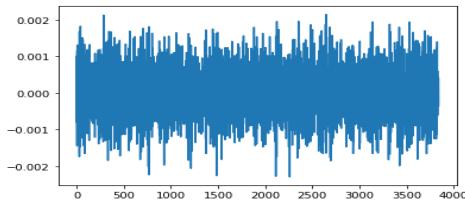
En el proceso para extraer las características se incluye la función de autocorrelación, en el siguiente código, se importa la librería numpy para usar su función de autocorrelación y aplicarla a todos los frames de la señal, si el frame a el cual se le aplica la función de autocorrelación supera el umbral de 0.1 quiere decir que el sonido es vocalizado, si no es así, el sonido es no vocalizado por lo tanto se omite en el procesamiento de la señal. La experimentación con varios valores para el umbral muestra que el valor de 0.1 es el mejor para discriminar sonidos no vocalizados.

```
import numpy as np
#función de autocorrelación
acorr = np.correlate(frame,frame,mode='full')
m = npamax(acorr[20:len(acorr)])
if m > 0.1:
```

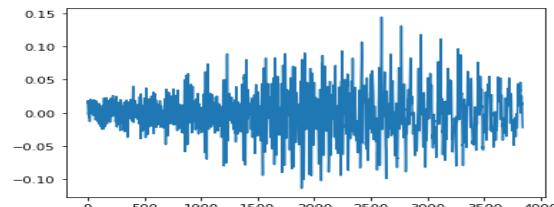
En la Figura 3.16(a): se muestra un frame cuando se le aplica la función de autocorrelación. En la figura(b): se muestra un frame que no supero el umbral de 0.1 y en la figura(c): se muestra un frame que si supero el umbral de 0.1.



(a): Despues de aplicar la función de autocorrelación.



(b): Frame de la señal de voz que no supero el umbral de 0.1.



(c): Frame de la señal de voz que si supero el umbral de 0.1.

Figura 3.16: Aplicación de la función de autocorrelación en frames de la señal de voz para descartar sonidos no vocalizados.

### 3.2.5 Windowing

los frames de la señal de voz son multiplicados con una función ventana, como se ve en la Figura 3.17, para minimizar el fenómeno de Gibbs que se ocasiona en el espectro de frecuencia por haber truncado la señal de voz.

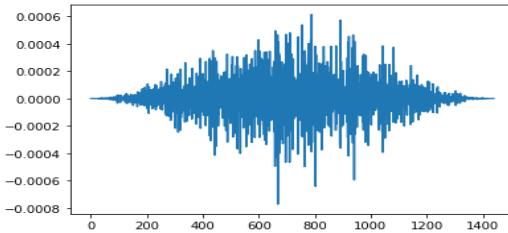


Figura 3.17: Grafica de frame multiplicado por la ventana de Hann.

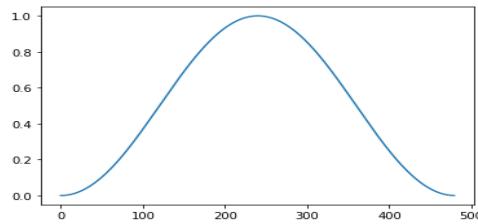


Figura 3.18: Grafica de la ventana de Hann.

### 3.3 Extracción de firmas espectrales

Una vez terminada la etapa de preprocesamiento, lo siguiente es extraer las características: MFCC, RASTA-MFCC, Entropy signature y MSES, el diagrama a bloques de la Figura 3.19 muestra el proceso seguido para obtener cada una de estas características de la señal de voz.

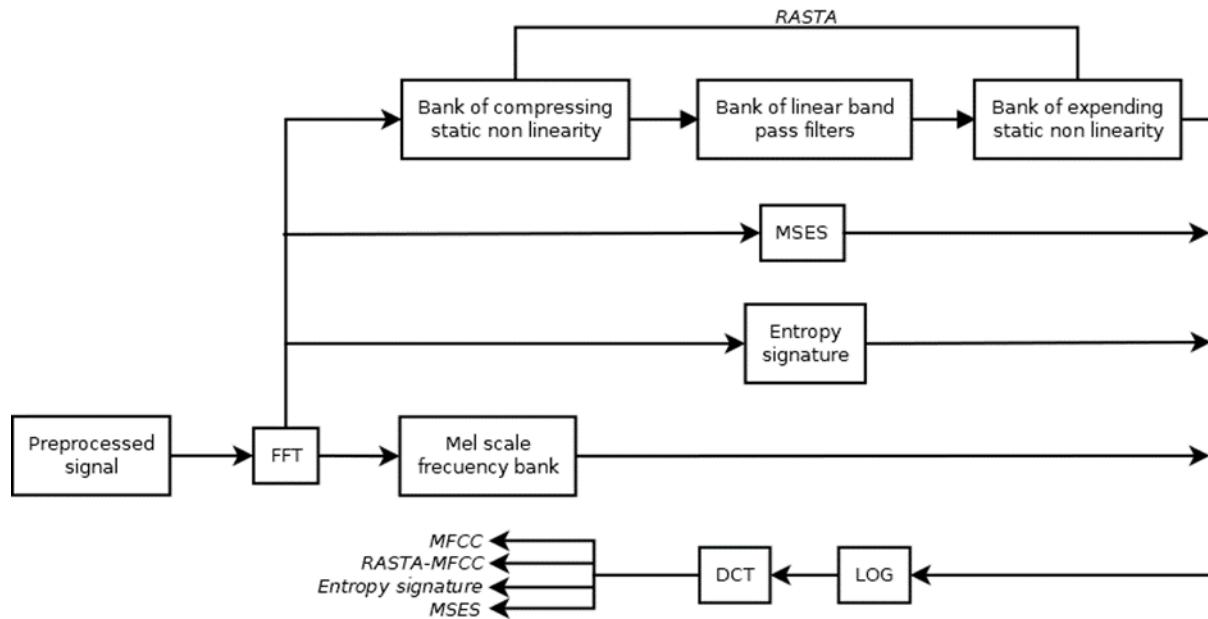


Figura 3.19: Diagrama a bloques con los pasos a seguir para obtener las características MFCC, RASTA-MFCC, Entropy signature y MSES.

A continuación, se detalla el proceso de extracción de cada característica.

### 3.3.1 MFCC

Para extraer la característica MFCC, lo primero es calcular la transformada de Fourier de tiempo corto para cada frame, de aquí se determina el espectrograma de la señal que se muestra en la Figura 3.20, luego se escala la frecuencia utilizando el banco de filtros de Mel, de aquí se obtiene el cepstrum de la señal que se muestra en la Figura 3.20 y por último la matriz de coeficientes MFCC se obtienen des correlacionando el espectro con la transformada de coseno discreta (DCT), el primer coeficiente de la matriz de coeficientes MFCC se muestra en la Figura 3.20, una vez obtenida la matriz de coeficientes, se obtiene su transpuesta y se promedia para tener un solo vector de coeficientes MFCC como se muestra en la Figura 3.20.

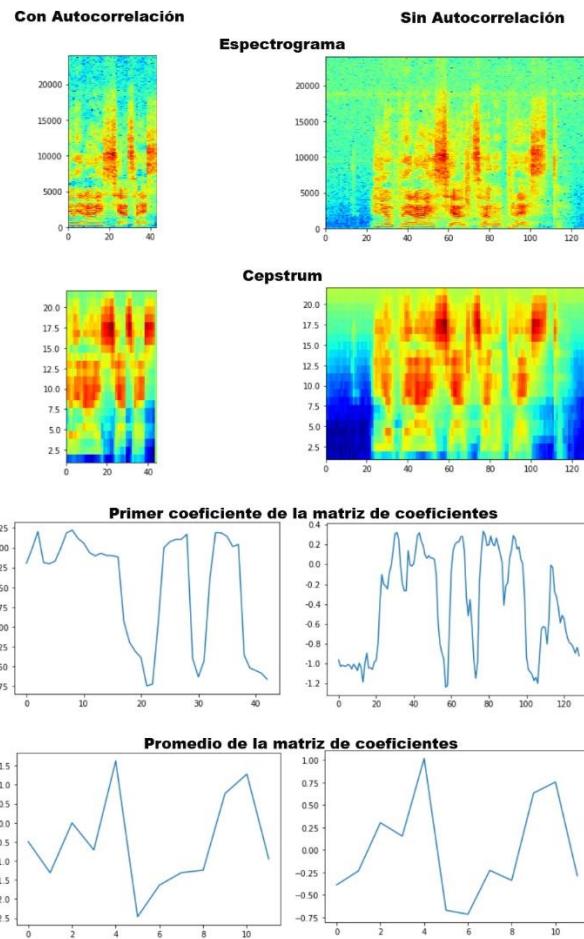
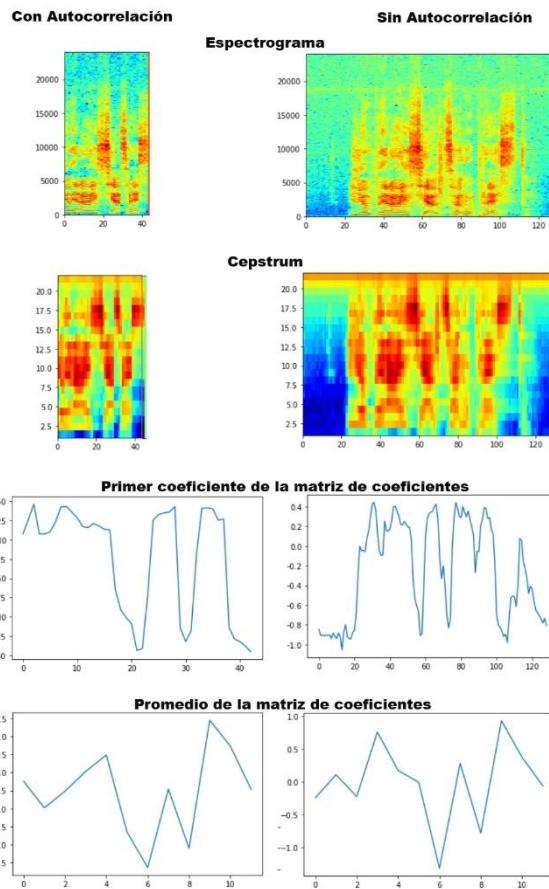


Figura 3.20: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica MFCC, con y sin función de autocorrelación.

### 3.3.2 RASTA-MFCC

Para extraer la característica RASTA-MFCC, el filtrado RASTA se aplica a la señal de voz en ventana para minimizar los efectos de ruido en la señal de voz, el filtrado es seguido por la extracción de MFCC de la señal filtrada RASTA para producir características RASTA–MFCC. El resultado de este proceso es una matriz  $L \times T$  (denominada como firma), donde  $L$  es el número de coeficientes y  $T$  denota el número de frames. A esta matriz se le saca la transpuesta, para luego obtener su promedio y así tener como resultado un solo vector de coeficientes.

En la Figura 3.21 se puede observar el espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y el promedio de la matriz de coeficientes, que se obtienen al extraer la característica RASTA-MFCC.

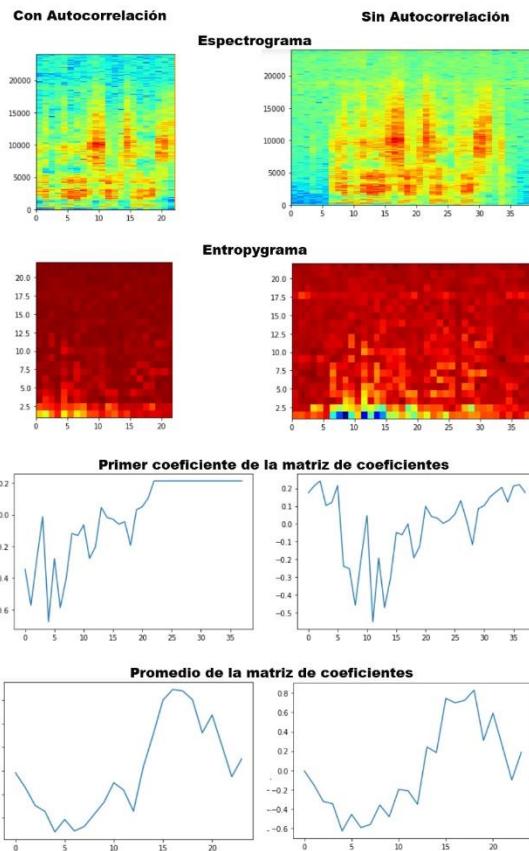


*Figura 3.21: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica RASTA-MFCC, con y sin función de autocorrelación.*

### 3.3.3 Entropy signature

Para extraer la característica Entropy signature, lo primero es calcular la transformada de Fourier de tiempo corto para cada frame, el proceso continúa calculando la entropía para cada una de las bandas críticas. El resultado de este proceso es una matriz  $L \times T$  (denominada como firma), donde  $L$  es el número de coeficientes de entropía y  $T$  denota el número de frames. A esta matriz se le saca la transpuesta, para luego obtener su promedio y así tener como resultado un solo vector de coeficientes.

En la Figura 3.22 se puede observar el espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y el promedio de la matriz de coeficientes, que se obtienen al extraer la característica Entropy Signature.

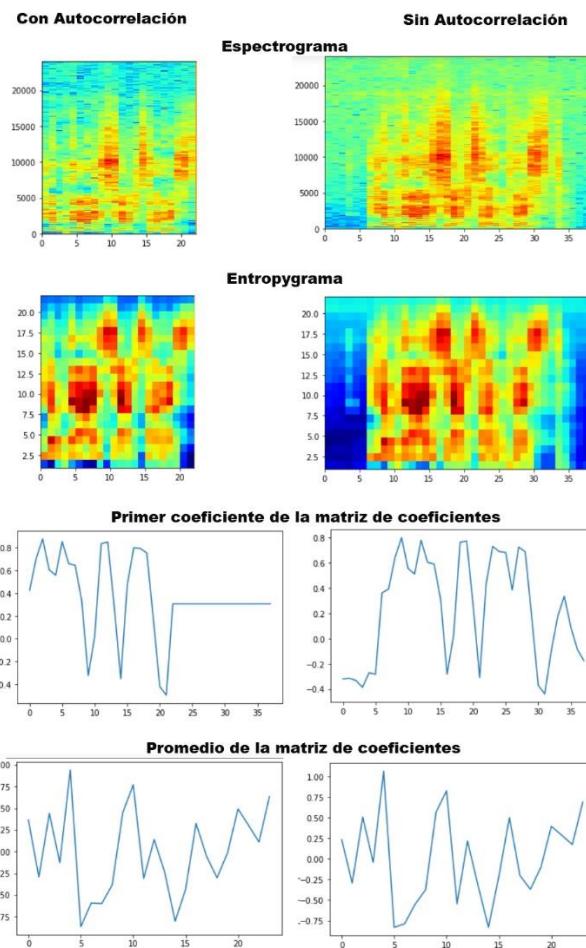


*Figura 3.22: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica Entropy Signature, con y sin función de autocorrelación.*

### 3.3.4 MSES

Para extraer la característica MSES, lo primero es calcular la transformada de Fourier de tiempo corto para cada frame, el proceso continúa calculando la entropía para cada una de las bandas críticas. Para calcular la entropía, se consideró un proceso aleatorio con dos variables aleatorias. Se asume que las partes reales e imaginarias de los coeficientes espectrales son variables aleatorias con una distribución normal y media cero. El resultado de este proceso es una matriz  $L \times T$  (denominada como firma), donde  $L$  es el número de coeficientes de entropía y  $T$  denota el número de frames. A esta matriz se le saca la transpuesta, para luego obtener su promedio y así tener como resultado un solo vector de coeficientes.

En la Figura 3.23 se puede observar el spectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y el promedio de la matriz de coeficientes, que se obtienen al extraer la característica MSES.



*Figura 3.23: Espectrograma, cepstrum, primer coeficiente de la matriz de coeficientes y promedio de la matriz de coeficientes, que se obtienen al extraer la característica MSES, con y sin función de autocorrelación.*

### 3.3.5 Extracción de características para cada archivo.

Habiendo mostrado como se extrae cada una de las características, ahora se muestra en el siguiente código de Python como se itera a través de cada archivo de audio para extraer una de las firmas espectrales para cada archivo de audio del conjunto de datos y se almacena en un Panda Dataframe junto con su etiqueta de clasificación.

A continuación, Se muestran una de las dos funciones creadas para este proceso, una para extraer el conjunto de prueba y otra para el de entrenamiento.

```
def extract_features_test_EMODB():

    # Se establece la ruta para EMODB dataset
    fulldatasetpath = '..\EMODB\audio'
    metadata = pd.read_csv('..\metadata\EMODB.csv')
    features = []

    # Itera a través de cada archivo de sonido y extrae las características
    for index, row in metadata.iterrows():

        file_name = os.path.join(os.path.abspath(fulldatasetpath),
                               str(row["fold"]) + '/' ,str(row["slice_file_name"]))

        class_label = row["class_name"]
        if row["if"] == "test" :
            data = mel freq ceps coef(file name)
            features.append([data.coefscalade, class_label])

    # Convertir en un Panda dataframe
    featuresdf = pd.DataFrame(features, columns=['feature',
                                                'class_label'])

    print('Finished feature extraction from ', len(featuresdf),
          'files')

    return featuresdf

featuresdf_EMODB_test=extract_features_test_EMODB()
featuresdf_EMODB_train=extract_features_train_EMODB()

Finished feature extraction from  161 files
Finished feature extraction from  374 files
```

### 3.3.6 Convertir los datos y las etiquetas

Ya habiendo extraído un vector característico para cada archivo de audio ahora es necesario codificar los datos de texto categóricos en datos numéricos comprensibles para el clasificador, para lograr esto se utiliza `sklearn.preprocessing.LabelEncoder` en el siguiente código de Python.

```
from sklearn.preprocessing import LabelEncoder
from keras.utils import to_categorical
import numpy as np

%store featuresdf_test_EMODB
%store featuresdf_train_EMODB

# Convierte las características y las correspondientes etiquetas de clasificación en matrices
# numéricas
x_test = np.array(featuresdf_test_EMODB.feature.tolist())
y_test = np.array(featuresdf_test_EMODB.class_label.tolist())

# Codifica las etiquetas de clasificación
le = LabelEncoder()
y_test = to_categorical(le.fit_transform(y_test))

x_train = np.array(featuresdf_train_EMODB.feature.tolist())
y_train = np.array(featuresdf_train_EMODB.class_label.tolist())

y_train = to_categorical(le.fit_transform(y_train))

%store x_train
%store x_test
%store y_train
%store y_test

y_train = np.array(featuresdf_train_EMODB.class_label.tolist())
y_test = np.array(featuresdf_test_EMODB.class_label.tolist())
y_train = np.append(y_train, y_test)
yy = to_categorical(le.fit_transform(y_train))

%store yy
%store le
```

## 3.4 Clasificadores

El último paso de la metodología es la clasificación en las 7 clases emocionales, de los vectores obtenidos para cada archivo de audio, en la etapa de extracción de características.

A continuación, se describe cada clasificador utilizado para los experimentos de esta tesis.

### 3.4.1 MLP

La configuración para el clasificador MLP se tomó del trabajo de [77], ya que consiguió buenos resultados al clasificar sonidos urbanos con la característica MFCC.

#### Arquitectura del modelo inicial

Se empieza con la construcción de una red neuronal de perceptrón multicapa (MLP) utilizando Keras y un backend de Tensorflow.

Se inicia con un sequential model para poder construir el modelo capa por capa.

Se empieza con una arquitectura de modelo simple, que consta de tres capas, una capa de entrada, una capa oculta y una capa de salida. Las tres capas son del tipo capa densa, que es un tipo de capa estándar que se utiliza en muchos casos para redes neuronales.

La primera capa recibirá la forma de entrada. Como cada muestra contiene 12 coeficientes (o columnas) se tiene una forma de (1x12) esto significa que se inicia con una forma de entrada de 12.

Las dos primeras capas tienen 500 nodos. La función de activación que se usa para las primeras dos capas es ReLU, o Rectified Linear Activation. Se ha demostrado que esta función de activación funciona bien en redes neuronales. También se aplica un valor Dropout del 50% en las dos primeras capas. Esto excluirá aleatoriamente los nodos de cada ciclo de actualización, lo que a su vez da como resultado una red que es capaz de una mejor generalización y es menos probable que se ajuste a los datos de entrenamiento. La capa de salida tiene 7 nodos que coinciden con el número de clasificaciones posibles. La activación para la capa de salida es softmax; Softmax hace que las salidas sumen 1 para que estas se puedan interpretar como probabilidades. Luego, el modelo hará su predicción basándose en qué opción tiene la mayor probabilidad.

```
import numpy as np
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D, MaxPooling2D
from keras.optimizers import Adam
from keras.utils import np_utils
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

num_labels = yy.shape[1]
filter_size = 2

# Construir el modelo
model = Sequential()

model.add(Dense(500, input_shape=(12,)))
model.add(Activation('relu'))
model.add(Dropout(0.5))

model.add(Dense(500))
model.add(Activation('relu'))
model.add(Dropout(0.5))

model.add(Dense(num_labels))
model.add(Activation('softmax'))
```

## Compilación del modelo

Para compilar el modelo, se usan los siguientes tres parámetros:

- Función de pérdida - se utiliza categorical\_crossentropy. Esta es la elección más común para la clasificación. Una puntuación más baja indica que el modelo está funcionando mejor.
- Métrica - se utiliza la métrica de accuracy que permite ver la puntuación de la precisión en los datos de validación cuando se entrena el modelo.
- Optimizador - se utiliza adam que es un optimizador generalmente bueno para muchos casos de uso.

```
# Compila el modelo
model.compile(loss='categorical_crossentropy', metrics=['accuracy'], optimizer='adam')
# Mostrar resumen de la arquitectura del modelo
model.summary()

# Calcular el accuracy previo al entrenamiento
score = model.evaluate(x_test, y_test, verbose=0)
accuracy = 100*score[1]

print("Pre-training accuracy EmoDB: %.4f%%" % accuracy)
```

A continuación, se muestran el resumen de la arquitectura del modelo junto con el accuracy previo al entrenamiento.

Layer(type)	Output Shape	Param #
dense_64(Dense)	(None, 500)	12500
activation_64(Activation)	(None, 500)	0
dropout_43(Dropout)	(None, 500)	0
dense_65(Dense)	(None, 500)	250500
activation_65(Activation)	(None, 500)	0
dropout_44(Dropout)	(None, 500)	0
dense_66(Dense)	(None, 7)	3507
activation_66(Activation)	(None, 7)	0
Total params:	266,507	
Trainable params:	266,507	
Non-trainable params:	0	

---

Pre-training accuracy EmoDB: 13.1429%

## Entrenamiento

Aquí se entrena el modelo 10 veces, ya que la inicialización de una red neuronal es aleatoria por eso es necesario promediar varias puntuaciones de la red neuronal. También se imprime la matriz de confusión y un reporte de los resultados obtenidos con diferentes métricas.

Se empieza con 100 épocas, que es el número de veces que el modelo circulará a través de los datos. El modelo mejorará en cada ciclo hasta que llegue a un punto determinado.

También se empieza con un tamaño de lote bajo, ya que tener un tamaño de lote grande puede reducir la capacidad de generalización del modelo.

```
from keras.callbacks import ModelCheckpoint
from datetime import datetime

trainscore = np.array([])
testscore = np.array([])

num_epochs = 100
num_batch_size = 32

checkpointer = ModelCheckpoint(filepath='saved_models/weights.best.basic_mlp.hdf5',
                               verbose=1, save_best_only=True)
start = datetime.now()

for n in range(0,10):

    model.fit(x_train, y_train, batch_size=num_batch_size, epochs=num_epochs,
               validation_data=(x_test, y_test), callbacks=[checkpointer], verbose=1)

    scoretrain = model.evaluate(x_train, y_train, verbose=0)
    trainscore = np.append(trainscore,scoretrain[1])
    scoretest = model.evaluate(x_test, y_test, verbose=0)
    testscore = np.append(testscore,scoretest[1])

    predicted_vector = model.predict_classes(x_test)
    y_pred = np.argmax(y_test, axis=1)
    conf_mat = confusion_matrix(y_pred,predicted_vector)
    report = classification_report(y_pred,predicted_vector)
    print(conf_mat)
    print(report)

duration = datetime.now() - start
print("Training completed in time: ", duration)

average = sum(trainscore)/len(trainscore)

print("Training Accuracy: ",trainscore)
print("Training Accuracy: ",average)

average = sum(testscore)/len(testscore)

print("Testing Accuracy: ",testscore)
print("Testing Accuracy: ",average)
```

A continuación, se imprime el entrenamiento de la red neuronal a partir de la época 98, se observa que en la época 100 el accuracy es de 0.6743.

```
Epoch 98/100
413/413 [=====] - 0s 286us/step - loss: 0.0089 - accuracy: 0.9976 - val_loss: 3.7195 -
val_accuracy: 0.6914
Epoch 00098: val_loss did not improve from 1.11009
```

```

Epoch 99/100
413/413 [=====] - 0s 290us/step - loss: 0.0064 - accuracy: 0.9952 - val_loss: 3.8402 -
val_accuracy: 0.6743
Epoch 00099: val_loss did not improve from 1.11009
Epoch 100/100
413/413 [=====] - 0s 290us/step - loss: 0.0089 - accuracy: 0.9952 - val_loss: 3.7976 -
val_accuracy: 0.6743
Epoch 00100: val_loss did not improve from 1.11009

```

A continuación, se muestra la matriz de confusión que se imprime al finalizar el entrenamiento de la red neuronal.

```

[[19 0 1 3 0 2 0]
 [ 1 19 1 4 0 0 0]
 [ 1 1 14 8 0 0 1]
 [ 1 2 2 20 0 0 0]
 [ 1 5 3 3 13 0 0]
 [ 3 1 0 4 1 16 0]
 [ 2 1 4 1 0 0 17]]

```

De esta matriz se obtiene un reporte con las métricas de precision, recall, f1-score y accuracy, los numeros del 0 al 6 hacen alusión a las 7 clases de emociones a clasificar, en este orden: Angustia, Disgusto, Alegría, Aburrimiento, Neutral, Tristeza, Ira.

	precision	recall	f1-score	support
0	0.68	0.76	0.72	25
1	0.66	0.76	0.70	25
2	0.56	0.56	0.56	25
3	0.47	0.80	0.59	25
4	0.93	0.52	0.67	25
5	0.89	0.64	0.74	25
6	0.94	0.68	0.79	25
accuracy		0.67		175
macro avg	0.73	0.67	0.68	175
weighted avg	0.73	0.67	0.68	175

A continuación, se imprimen los diez valores que fueron promediados para la obtención de puntaje recall final, tanto de entrenamiento como de prueba.

```

Training Recall: [0.99757868 1. 1. 1. 1. 1.
 1. 1. 1. 1. ]
Training Average: 0.9997578680515289

```

```

Testing Recall: [0.6857143 0.72000003 0.65714288 0.68000001 0.6857143 0.70285714
0.69714284 0.69142854 0.70857143 0.67428571]
Testing Average: 0.6902857184410095

```

### 3.4.2 SVM

Se construye una SVM utilizando sklearn esta librería pone a disposición cuatro diferentes kernels para abordar la regresión, que son:

- Linear: Kernel utilizado cuando se pretende aproximar una función con una función lineal.

$$K(x, x') = (x, x') \quad (3.1)$$

- Polynomial: Kernel que da la posibilidad de resolver problemas con un kernel polinomial con diferentes grados  $d$  y coeficientes  $R$ .

$$K_d(x, x') = ((x, x') + R)^d \quad (3.2)$$

- Radial: Kernel que da la posibilidad de resolver problemas con un kernel gaussiano con diferentes valores de  $\gamma$ .

$$K(x, x') = \exp(-\gamma ||x - x'||^2) \quad (3.3)$$

- Sigmoid: Kernel que da la posibilidad de resolver problemas con un kernel sigmoidal con diferentes valores de  $\gamma$  y  $R$ .

$$K(x, x') = \tanh(\gamma(x, x') + R) \quad (3.4)$$

### Compilación del modelo

Para compilar el modelo se usan los parámetros de gamma y degree, el valor de estos se obtuvo mediante experimentación buscando el mejor puntaje de recall.

```
from sklearn.svm import SVC
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

for kernel in('linear', 'rbf', 'poly', 'sigmoid'):
    svclassifier = SVC(kernel=kernel, gamma=1, degree=3)
    svclassifier.fit(X_train, y_train)
    print('Accuracy of SVM classifier on training set : {:.2f}'.format(svclassifier.score(X_train,y_train)))
    print('Accuracy of SVM classifier on test set : {:.2f}'.format(svclassifier.score(X_test,y_test)))
    pred = svclassifier.predict(X_test)
    print(confusion_matrix(y_test,pred))
    print(classification_report(y_test,pred))
```

A continuación, se imprime el puntaje recall, matriz de confusión y un reporte con las metricas precision, recall, f1-score y accuracy para los kernel Linear, RBF, Poly y Sigmoid, en el reporte los numeros del 0 al 6 hacen alusión a las 7 clases de emociones a clasificar, en este orden: Angustia, Disgusto, Alegría, Aburrimiento, Neutral, Tristeza, Ira.

#### LINEAR

Accuracy of SVM classifier on training set : 0.74

Accuracy of SVM classifier on test set : 0.48

```
[[12 2 0 9 1 1 0]
 [3 14 2 3 1 2 0]
 [2 5 7 6 3 1 1]
 [6 3 0 15 0 0 1]
 [3 4 3 5 8 1 1]
 [3 1 0 7 0 14 0]
 [2 2 3 2 2 0 14]]
```

	precision	recall	f1-score	support
0	0.39	0.48	0.43	25
1	0.45	0.56	0.50	25
2	0.47	0.28	0.35	25
3	0.32	0.60	0.42	25
4	0.53	0.32	0.40	25
5	0.74	0.56	0.64	25
6	0.82	0.56	0.67	25
accuracy		0.48		175
macro avg	0.53	0.48	0.49	175
weighted avg	0.53	0.48	0.49	175

### RBF

Accuracy of SVM classifier on training set : 0.99

Accuracy of SVM classifier on test set : 0.69

```
[[17 2 1 3 0 1 1]
 [ 0 21 0 2 0 1 1]
 [ 0 0 14 7 1 0 3]
 [ 1 2 2 19 0 0 1]
 [ 0 1 2 5 15 0 2]
 [ 4 1 1 4 0 15 0]
 [ 1 1 2 2 0 0 19]]
```

	precision	recall	f1-score	support
0	0.74	0.68	0.71	25
1	0.75	0.84	0.79	25
2	0.64	0.56	0.60	25
3	0.45	0.76	0.57	25
4	0.94	0.60	0.73	25
5	0.88	0.60	0.71	25
6	0.70	0.76	0.73	25
accuracy		0.69		175
macro avg	0.73	0.69	0.69	175
weighted avg	0.73	0.69	0.69	175

### POLY

Accuracy of SVM classifier on training set : 1.00

Accuracy of SVM classifier on test set : 0.66

```
[[19 2 2 1 0 1 0]
 [ 2 19 0 3 0 0 1]
 [ 3 4 10 5 2 0 1]
 [ 2 3 2 18 0 0 0]
 [ 1 2 1 3 18 0 0]
 [ 4 2 0 2 0 17 0]
 [ 3 2 4 1 1 0 14]]
```

	precision	recall	f1-score	support
0	0.56	0.76	0.64	25
1	0.56	0.76	0.64	25
2	0.53	0.40	0.45	25
3	0.55	0.72	0.62	25
4	0.86	0.72	0.78	25
5	0.94	0.68	0.79	25
6	0.88	0.56	0.68	25
accuracy		0.66		175
macro avg	0.70	0.66	0.66	175
weighted avg	0.70	0.66	0.66	175

**SIGMOID**

Accuracy of SVM classifier on training set : 0.12

Accuracy of SVM classifier on test set : 0.14

[ [ 0 2 2 1 1 1 8 1 ]

[ 3 5 1 0 1 1 3 2 ]

[ 0 4 1 0 1 5 3 2 ]

[ 3 2 3 0 9 6 2 ]

[ 2 4 4 0 8 3 4 ]

[ 2 3 0 2 1 0 7 1 ]

[ 2 1 1 2 1 4 2 3 ] ]

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.00	0.00	0.00	25
---	------	------	------	----

1	0.24	0.20	0.22	25
---	------	------	------	----

2	0.08	0.04	0.05	25
---	------	------	------	----

3	0.00	0.00	0.00	25
---	------	------	------	----

4	0.10	0.32	0.16	25
---	------	------	------	----

5	0.22	0.28	0.25	25
---	------	------	------	----

6	0.20	0.12	0.15	25
---	------	------	------	----

accuracy		0.14		175
----------	--	------	--	-----

macro avg	0.12	0.14	0.12	175
-----------	------	------	------	-----

weighted avg	0.12	0.14	0.12	175
--------------	------	------	------	-----

### 3.4.3 KNN

Se construye un clasificador KNN utilizando la librería sklearn, para obtener el puntaje de recall para los vecinos 1, 3 y 5.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

for n in range(1,6,2):
    n_neighbors = n
    knn = KNeighborsClassifier(n_neighbors, metric = 'minkowski', p = 2)
    knn.fit(X_train,y_train)
    print('Accuracy of KNN classifier on training set : {:.2f}'.format(knn.score(X_train,y_train)))
    print('Accuracy of KNN classifier on test set : {:.2f}'.format(knn.score(X_test,y_test)))

    pred = knn.predict(X_test)
    conf_mat = confusion_matrix(y_test,pred)
    report = classification_report(y_test,pred)
    print(conf_mat)
    print(report)
```

A continuación, se muestra la impresión del puntaje de recall, matriz de confusión y un reporte con las métricas precision, recall, f1-score y accuracy para los vecinos 1, 3 y 5, en el reporte los numeros del 0 al 6 hacen alusión a las 7 clases de emociones a clasificar, en este orden: Angustia, Disgusto, Alegría, Aburrimiento, Neutral, Tristeza, Ira.

#### Neighbor 1

Accuracy of KNN classifier on training set : 1.00

Accuracy of KNN classifier on test set : 0.63

```
[[18 2 2 1 0 2 0]
 [ 3 18 0 2 2 0 0]
 [ 1 1 9 8 2 0 4]
 [ 1 3 2 17 1 0 1]
 [ 0 2 3 2 17 0 1]
 [ 3 1 0 2 2 16 1]
 [ 2 3 4 1 0 0 15]]
```

	precision	recall	f1-score	support
0	0.64	0.72	0.68	25
1	0.60	0.72	0.65	25
2	0.45	0.36	0.40	25
3	0.52	0.68	0.59	25
4	0.71	0.68	0.69	25
5	0.89	0.64	0.74	25
6	0.68	0.60	0.64	25
accuracy		0.63		175
macro avg	0.64	0.63	0.63	175
weighted avg	0.64	0.63	0.63	175

#### Neighbor 3

Accuracy of KNN classifier on training set : 0.89

Accuracy of KNN classifier on test set : 0.65

```
[[18 2 1 2 0 2 0]
 [ 2 23 0 0 0 0 0]
 [ 3 1 13 4 1 0 3]
 [ 4 3 4 14 0 0 0]
 [ 2 4 3 2 14 0 0]
 [ 5 1 0 2 0 17 0]
 [ 2 3 5 1 0 0 14]]
```

```
precision recall f1-score support
```

0	0.50	0.72	0.59	25
1	0.62	0.92	0.74	25
2	0.50	0.52	0.51	25
3	0.56	0.56	0.56	25
4	0.93	0.56	0.70	25
5	0.89	0.68	0.77	25
6	0.82	0.56	0.67	25
accuracy		0.65		175
macro avg	0.69	0.65	0.65	175
weighted avg	0.69	0.65	0.65	175

### Neighbor 5

Accuracy of KNN classifier on training set : 0.85

Accuracy of KNN classifier on test set : 0.63

[[18 2 0 2 0 2 1]				
[ 2 21 0 1 0 1 0]				
[ 2 0 14 5 2 0 2]				
[ 5 3 2 14 0 0 1]				
[ 1 5 2 3 14 0 0]				
[ 8 0 0 1 0 16 0]				
[ 2 2 3 1 1 2 14]]				
precision	recall	f1-score	support	
0	0.47	0.72	0.57	25
1	0.64	0.84	0.72	25
2	0.67	0.56	0.61	25
3	0.52	0.56	0.54	25
4	0.82	0.56	0.67	25
5	0.76	0.64	0.70	25
6	0.78	0.56	0.65	25
accuracy		0.63		175
macro avg	0.67	0.63	0.64	175
weighted avg	0.67	0.63	0.64	175

## CAPITULO 4 EXPERIMENTOS

En esta sección se presentan los experimentos realizados con sus respectivos resultados al clasificar las emociones usando varias configuraciones en el proceso de extracción de características e incluyendo la función de autocorrelación para descartar sonidos no vocalizados y ruido.

En cada experimento se obtiene la tabla de resultados obtenidos al clasificar los vectores característicos de las firmas MFCC, RASTA-MFCC, Entropy Signature y MSES usando la métrica recall, con los clasificadores MLP, KNN y SVM. Se utiliza la metrica recall porque este es el principal criterio de comparación con otros trabajos, se pueden usar las otras métricas, pero en este caso solo se busca la comparación con otros trabajos en el estado del arte.

Los resultados reportados de cada clasificador son, para MLP el promedio de diez puntuaciones recall, para KNN el puntaje recall para los cinco primeros vecinos impares y de SVM el puntaje recall obtenido para los kernel linear, rbf, poly y sigmoid.

### 4.1 Experimentos con autocorrelación de la base de datos EMODB

#### Experimento 1

##### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, descartando el filtro de preéñfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

##### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, sin preéñfasis, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

##### Resumen de la Tabla 4.1

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.1 muestran que extrayendo 48 coeficientes se obtiene el mejor promedio de recall para la firma MFCC mientras que para RASTA-MFCC, Entropy Signature y MSES el mejor promedio de recall se obtiene con 24 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.73	0.75	0.47	0.69
24	0.77	<b>0.78</b>	<b>0.53</b>	<b>0.77</b>
36	0.73	0.73	0.49	0.75
48	<b>0.80</b>	0.77	0.43	0.73

Tabla 4.1: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.2

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.2 muestran que extrayendo 12 coeficientes se obtiene el mejor puntaje de recall para la firma RASTA-MFCC mientras que para MFCC, Entropy Signature y MSES el mejor promedio de recall se obtiene con 24 coeficientes. El número que aparece entre paréntesis en la Tabla 4.2 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.63(1) 0.65(3) 0.65(5)	0.62(1) <b>0.66(3)</b> 0.63(5)	0.37(1) 0.39(3) 0.43(5)	0.62(1) 0.61(3) 0.61(5)
24	0.71(1) <b>0.73(3)</b> 0.70(5)	0.63(1) 0.61(3) 0.65(5)	<b>0.46(1)</b> 0.43(3) 0.41(5)	0.69(1) <b>0.71(3)</b> 0.70(5)
36	0.62(1) 0.59(3) 0.65(5)	0.63(1) 0.61(3) 0.65(5)	0.43(1) 0.42(3) 0.45(5)	0.65(1) 0.63(3) 0.65(5)
48	0.59(1) 0.57(3) 0.57(5)	0.63(1) 0.61(3) 0.65(5)	0.34(1) 0.30(3) 0.35(5)	0.60(1) 0.58(3) 0.57(5)

Tabla 4.2: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.3

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.3 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las 4 firmas espectrales MFCC, RASTA-MFCC, Entropy Signature y MSES. El texto que aparece entre paréntesis en la Tabla 4.3 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.65(linear) 0.72(rbf) 0.65(poly) 0.24(sigmoid)	0.65(linear) <b>0.74(rbf)</b> 0.70(poly) 0.30(sigmoid)	0.42(linear) 0.45(rbf) 0.46(poly) 0.23(sigmoid)	0.59(linear) 0.70(rbf) 0.70(poly) 0.21(sigmoid)
24	0.68(linear) 0.72(rbf) <b>0.76(poly)</b> 0.25(sigmoid)	0.72(linear) <b>0.74(rbf)</b> 0.72(poly) 0.27(sigmoid)	0.46(linear) 0.45(rbf) <b>0.55(poly)</b> 0.23(sigmoid)	0.72(linear) <b>0.76(rbf)</b> 0.71(poly) 0.26(sigmoid)
36	0.68(linear) 0.56(rbf) 0.65(poly) 0.22(sigmoid)	0.66(linear) 0.58(rbf) 0.63(poly) 0.23(sigmoid)	0.43(linear) 0.48(rbf) 0.47(poly) 0.22(sigmoid)	0.66(linear) 0.68(rbf) 0.71(poly) 0.28(sigmoid)
48	0.69(linear) 0.58(rbf) 0.68(poly) 0.20(sigmoid)	0.68(linear) 0.58(rbf) 0.71(poly) 0.24(sigmoid)	0.34(linear) 0.35(rbf) 0.52(poly) 0.24(sigmoid)	0.65(linear) 0.63(rbf) 0.63(poly) 0.24(sigmoid)

Tabla 4.3: Resultados de recall usando el clasificador SVM.

## Experimento 2

### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, usando el filtro de preéñfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, con preéñfasis de 0.97, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

### Resumen de la Tabla 4.4

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.4 muestran que extrayendo 48 coeficientes se obtiene el mejor promedio de recall para las firma MSES y MFCC mientras que para RASTA-MFCC, Entropy Signature y nuevamente MFCC el mejor promedio de recall se obtiene con 24 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.69	0.73	0.46	0.66
24	<b>0.77</b>	<b>0.76</b>	<b>0.55</b>	0.75
36	0.73	0.71	0.49	0.76
48	<b>0.77</b>	0.73	0.41	<b>0.77</b>

Tabla 4.4: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.5

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.5 muestran que extrayendo 12 coeficientes se obtiene el mejor puntaje de recall para las firmas RASTA-MFCC y MFCC mientras que para Entropy Signature, MSES y nuevamente RASTA-MFCC el mejor promedio de recall se obtiene con 24 coeficientes. El número que aparece entre paréntesis en la Tabla 4.5 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.65(1) 0.66(3) <b>0.69(5)</b>	0.66(1) <b>0.68(3)</b> 0.65(5)	0.41(1) 0.37(3) 0.38(5)	0.55(1) 0.61(3) 0.59(5)
24	0.67(1) 0.68(3) 0.66(5)	<b>0.68(1)</b> 0.64(3) 0.63(5)	0.43(1) 0.38(3) <b>0.50(5)</b>	<b>0.69(1)</b> <b>0.69(3)</b> 0.71(5)
36	0.61(1) 0.60(3) 0.60(5)	0.60(1) 0.60(3) 0.60(5)	0.43(1) 0.42(3) 0.48(5)	0.56(1) 0.59(3) 0.65(5)
48	0.59(1) 0.48(3) 0.50(5)	0.54(1) 0.55(3) 0.55(5)	0.29(1) 0.32(3) 0.32(5)	0.62(1) 0.59(3) 0.60(5)

Tabla 4.5: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.6

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.6 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales RASTA-MFCC, Entropy Signature y MSES mientras que para MFCC con 48 coeficientes. El texto que aparece entre paréntesis en la Tabla 4.6 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.63(linear) 0.67(rbf) 0.65(poly) 0.27(sigmoid)	0.62(linear) <b>0.68(rbf)</b> 0.65(poly) 0.28(sigmoid)	0.36(linear) 0.40(rbf) 0.39(poly) 0.24(sigmoid)	0.62(linear) 0.64(rbf) 0.63(poly) 0.43(sigmoid)
24	0.68(linear) 0.53(rbf) 0.70(poly) 0.22(sigmoid)	0.67(linear) 0.55(rbf) <b>0.68(poly)</b> 0.17(sigmoid)	0.42(linear) 0.48(rbf) <b>0.52(poly)</b> 0.24(sigmoid)	0.66(linear) <b>0.76(rbf)</b> 0.69(poly) 0.26(sigmoid)
36	0.67(linear) 0.32(rbf) 0.59(poly) 0.19(sigmoid)	0.60(linear) 0.32(rbf) 0.61(poly) 0.16(sigmoid)	0.40(linear) 0.48(rbf) 0.47(poly) 0.24(sigmoid)	0.63(linear) 0.66(rbf) 0.66(poly) 0.18(sigmoid)
48	0.68(linear) 0.56(rbf) <b>0.71(poly)</b> 0.24(sigmoid)	0.62(linear) 0.60(rbf) 0.67(poly) 0.24(sigmoid)	0.36(linear) 0.38(rbf) <b>0.52(poly)</b> 0.24(sigmoid)	0.61(linear) 0.69(rbf) 0.60(poly) 0.24(sigmoid)

Tabla 4.6: Resultados de recall usando el clasificador SVM.

## Experimento 3

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 3 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y descartando el filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, sin preénfasis, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.7

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.7 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor promedio de recall para las firmas MFCC y RASTA-MFCC mientras que para Entropy Signature y MSES el mejor promedio de recall se obtiene usando un tamaño de frame de 50 y 20 ms respectivamente.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	<b>0.81</b>	<b>0.82</b>	0.13	0.72
20	0.79	0.80	0.49	<b>0.79</b>
30	0.78	0.77	0.52	0.78
40	0.78	0.77	0.49	0.78
50	0.77	0.77	<b>0.54</b>	0.77
60	0.78	0.77	0.53	0.76
70	0.76	0.78	0.50	0.74
80	0.77	0.77	0.48	0.75
90	0.77	0.76	0.47	0.77
100	0.75	0.76	0.48	0.75

Tabla 4.7: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.8

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.8 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor puntaje de recall para las firmas RASTA-MFCC y MFCC mientras que para Entropy Signature y MSES el mejor puntaje de recall se obtiene con 50 y 20 ms respectivamente. El número que aparece entre paréntesis en la Tabla 4.8 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.74(1) 0.76(3) <b>0.81(5)</b>	0.74(1) 0.78(3) <b>0.81(5)</b>	0.12(1) 0.12(3) 0.12(5)	0.58(1) 0.57(3) 0.57(5)
20	0.71(1) 0.73(3) 0.68(5)	0.70(1) 0.73(3) 0.70(5)	0.35(1) 0.39(3) 0.40(5)	0.74(1) <b>0.76(3)</b> 0.71(5)
30	0.71(1) 0.73(3) 0.70(5)	0.66(1) 0.68(3) 0.70(5)	0.41(1) 0.43(3) 0.46(5)	0.70(1) 0.71(3) 0.69(5)
40	0.70(1) 0.71(3) 0.69(5)	0.63(1) 0.70(3) 0.68(5)	0.47(1) 0.45(3) 0.46(5)	0.71(1) 0.70(3) 0.67(5)
50	0.69(1) 0.70(3) 0.69(5)	0.63(1) 0.69(3) 0.65(5)	0.47(1) 0.43(3) <b>0.50(5)</b>	0.71(1) 0.70(3) 0.67(5)
60	0.68(1) 0.70(3) 0.71(5)	0.68(1) 0.70(3) 0.69(5)	0.47(1) 0.45(3) 0.45(5)	0.71(1) 0.73(3) 0.71(5)
70	0.68(1) 0.71(3) 0.70(5)	0.65(1) 0.70(3) 0.71(5)	0.44(1) 0.40(3) 0.46(5)	0.70(1) 0.71(3) 0.71(5)
80	0.73(1) 0.71(3) 0.70(5)	0.67(1) 0.74(3) 0.71(5)	0.47(1) 0.43(3) 0.41(5)	0.71(1) 0.75(3) 0.71(5)
90	0.71(1) 0.72(3) 0.72(5)	0.67(1) 0.74(3) 0.68(5)	0.43(1) 0.43(3) 0.40(5)	0.71(1) 0.70(3) 0.69(5)
100	0.70(1) 0.69(3) 0.70(5)	0.70(1) 0.73(3) 0.70(5)	0.47(1) 0.39(3) 0.39(5)	0.72(1) 0.70(3) 0.69(5)

Tabla 4.8: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.9

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.9 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor puntaje de recall para la firma MFCC, para RASTA-MFCC 50 ms, Entropy Signature 30ms y MSES 20 ms El texto que aparece entre paréntesis en la Tabla 4.9 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.70(linear) 0.72(rbf) <b>0.78(poly)</b> 0.25(sigmoid)	0.71(linear) 0.71(rbf) 0.76(poly) 0.24(sigmoid)	0.13(linear) 0.13(rbf) 0.13(poly) 0.13(sigmoid)	0.49(linear) 0.57(rbf) 0.60(poly) 0.35(sigmoid)
20	0.70(linear) 0.72(rbf) 0.76(poly) 0.21(sigmoid)	0.69(linear) 0.72(rbf) 0.74(poly) 0.20(sigmoid)	0.42(linear) 0.51(rbf) 0.49(poly) 0.24(sigmoid)	0.71(linear) <b>0.78(rbf)</b> 0.71(poly) 0.39(sigmoid)
30	0.70(linear) 0.72(rbf) 0.76(poly) 0.21(sigmoid)	0.72(linear) 0.74(rbf) 0.72(poly) 0.27(sigmoid)	0.46(linear) 0.45(rbf) <b>0.55(poly)</b> 0.23(sigmoid)	0.72(linear) 0.76(rbf) 0.71(poly) 0.26(sigmoid)
40	0.67(linear) 0.70(rbf) 0.76(poly) 0.27(sigmoid)	0.68(linear) 0.73(rbf) 0.73(poly) 0.27(sigmoid)	0.47(linear) 0.48(rbf) 0.48(poly) 0.24(sigmoid)	0.68(linear) 0.75(rbf) 0.73(poly) 0.27(sigmoid)
50	0.66(linear) 0.71(rbf) 0.76(poly) 0.25(sigmoid)	0.66(linear) 0.71(rbf) <b>0.77(poly)</b> 0.22(sigmoid)	0.52(linear) 0.49(rbf) 0.53(poly) 0.19(sigmoid)	0.67(linear) 0.75(rbf) 0.73(poly) 0.24(sigmoid)
60	0.67(linear) 0.73(rbf) 0.73(poly) 0.26(sigmoid)	0.68(linear) 0.71(rbf) 0.74(poly) 0.22(sigmoid)	0.43(linear) 0.45(rbf) 0.49(poly) 0.21(sigmoid)	0.68(linear) 0.75(rbf) 0.71(poly) 0.22(sigmoid)
70	0.69(linear) 0.71(rbf) 0.74(poly) 0.25(sigmoid)	0.70(linear) 0.71(rbf) 0.73(poly) 0.23(sigmoid)	0.45(linear) 0.45(rbf) 0.51(poly) 0.24(sigmoid)	0.70(linear) 0.74(rbf) 0.70(poly) 0.23(sigmoid)
80	0.70(linear) 0.72(rbf) 0.73(poly) 0.26(sigmoid)	0.66(linear) 0.71(rbf) 0.75(poly) 0.20(sigmoid)	0.43(linear) 0.48(rbf) 0.45(poly) 0.20(sigmoid)	0.71(linear) 0.75(rbf) 0.68(poly) 0.26(sigmoid)
90	0.70(linear) 0.72(rbf) 0.73(poly) 0.26(sigmoid)	0.66(linear) 0.70(rbf) 0.75(poly) 0.20(sigmoid)	0.44(linear) 0.48(rbf) 0.49(poly) 0.22(sigmoid)	0.66(linear) 0.73(rbf) 0.70(poly) 0.22(sigmoid)
100	0.67(linear) 0.70(rbf) 0.72(poly) 0.25(sigmoid)	0.67(linear) 0.68(rbf) 0.74(poly) 0.19(sigmoid)	0.48(linear) 0.48(rbf) 0.50(poly) 0.21(sigmoid)	0.66(linear) 0.74(rbf) 0.73(poly) 0.22(sigmoid)

Tabla 4.9: Resultados de recall usando el clasificador SVM.

## Experimento 4

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 4 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y usando filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.10

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.10 muestran que usando un tamaño de frame de 40 ms se obtiene el mejor promedio de recall para la firma MFCC, 80 ms para RASTA-MFCC, 90 ms para Entropy Signature y 60 ms para MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.71	0.71	0.13	0.67
20	0.77	0.73	0.49	0.75
30	0.75	0.75	0.54	0.75
40	<b>0.78</b>	0.75	0.54	0.76
50	0.74	0.76	0.50	0.74
60	0.76	0.75	0.52	<b>0.77</b>
70	0.76	0.75	0.47	0.73
80	0.75	<b>0.77</b>	0.43	0.75
90	0.75	0.75	<b>0.55</b>	0.75
100	0.74	0.75	0.51	0.73

Tabla 4.10: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.11

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.11 muestran que usando un tamaño de frame de 20 ms se obtiene el mejor puntaje de recall para las firmas RASTA-MFCC y MFCC mientras que para Entropy Signature y MSES el mejor puntaje de recall se obtiene con 30 y 40 ms respectivamente. El número que aparece entre paréntesis en la Tabla 4.11 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.70(1)	0.70(1)	0.13(1)	0.55(1)
	0.66(3)	0.66(3)	0.13(3)	0.52(3)
	0.68(5)	0.67(5)	0.13(5)	0.53(5)
20	0.69(1)	<b>0.71(1)</b>	0.40(1)	0.69(1)
	0.70(3)	0.70(3)	0.39(3)	0.71(3)
	<b>0.79(5)</b>	0.68(5)	0.36(5)	0.70(5)
30	0.67(1)	0.68(1)	<b>0.50(1)</b>	0.71(1)
	0.68(3)	0.64(3)	0.43(3)	0.69(3)
	0.66(5)	0.63(5)	0.38(5)	0.69(5)
40	0.68(1)	0.65(1)	0.47(1)	<b>0.74(1)</b>
	0.68(3)	0.66(3)	0.39(3)	0.70(3)
	0.68(5)	0.68(5)	0.43(5)	0.70(5)
50	0.68(1)	0.67(1)	0.42(1)	0.71(1)
	0.71(3)	0.67(3)	0.37(3)	0.70(3)
	0.68(5)	0.66(5)	0.42(5)	0.68(5)
60	0.71(1)	0.66(1)	0.45(1)	0.73(1)
	0.68(3)	0.67(3)	0.42(3)	0.66(3)
	0.69(5)	0.66(5)	0.44(5)	0.69(5)
70	0.70(1)	0.66(1)	0.45(1)	<b>0.74(1)</b>
	0.65(3)	0.65(3)	0.41(3)	0.67(3)
	0.71(5)	0.67(5)	0.48(5)	0.69(5)
80	0.66(1)	0.63(1)	0.45(1)	0.70(1)
	0.65(3)	0.66(3)	0.42(3)	0.73(3)
	0.70(5)	0.68(5)	0.44(5)	0.67(5)
90	0.70(1)	0.66(1)	0.41(1)	0.72(1)
	0.68(3)	0.66(3)	0.42(3)	0.68(3)
	0.68(5)	0.69(5)	0.42(5)	0.68(5)
100	0.68(1)	0.65(1)	0.47(1)	0.70(1)
	0.68(3)	0.65(3)	0.43(3)	0.68(3)
	0.71(5)	0.69(5)	0.43(5)	0.69(5)

Tabla 4.11: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.12

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.12 muestran que usando un tamaño de frame de 100 ms se obtiene el mejor puntaje de recall para la firma MFCC, para RASTA-MFCC 80 ms, Entropy Signature 20 y 60 ms y MSES 30 ms. El texto que aparece entre paréntesis en la Tabla 4.12 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.64(linear) 0.70(rbf) 0.67(poly) 0.20(sigmoid)	0.64(linear) 0.28(rbf) 0.65(poly) 0.22(sigmoid)	0.13(linear) 0.13(rbf) 0.13(poly) 0.13(sigmoid)	0.35(linear) 0.40(rbf) 0.32(poly) 0.37(sigmoid)
20	0.66(linear) 0.40(rbf) 0.66(poly) 0.22(sigmoid)	0.65(linear) 0.42(rbf) 0.68(poly) 0.20(sigmoid)	0.40(linear) 0.44(rbf) <b>0.55(poly)</b> 0.24(sigmoid)	0.70(linear) 0.75(rbf) 0.70(poly) 0.35(sigmoid)
30	0.68(linear) 0.53(rbf) 0.70(poly) 0.22(sigmoid)	0.67(linear) 0.55(rbf) 0.68(poly) 0.17(sigmoid)	0.42(linear) 0.48(rbf) 0.52(poly) 0.24(sigmoid)	0.66(linear) <b>0.76(rbf)</b> 0.69(poly) 0.26(sigmoid)
40	0.66(linear) 0.60(rbf) 0.66(poly) 0.20(sigmoid)	0.64(linear) 0.61(rbf) 0.68(poly) 0.14(sigmoid)	0.44(linear) 0.43(rbf) 0.53(poly) 0.24(sigmoid)	0.68(linear) 0.74(rbf) 0.71(poly) 0.23(sigmoid)
50	0.69(linear) 0.65(rbf) 0.68(poly) 0.20(sigmoid)	0.68(linear) 0.65(rbf) 0.70(poly) 0.18(sigmoid)	0.48(linear) 0.53(rbf) 0.52(poly) 0.20(sigmoid)	0.65(linear) 0.75(rbf) 0.67(poly) 0.24(sigmoid)
60	0.64(linear) 0.63(rbf) 0.67(poly) 0.20(sigmoid)	0.66(linear) 0.66(rbf) 0.70(poly) 0.18(sigmoid)	0.43(linear) 0.46(rbf) <b>0.55(poly)</b> 0.22(sigmoid)	0.64(linear) 0.71(rbf) 0.69(poly) 0.24(sigmoid)
70	0.68(linear) 0.65(rbf) 0.68(poly) 0.17(sigmoid)	0.66(linear) 0.66(rbf) 0.70(poly) 0.18(sigmoid)	0.47(linear) 0.51(rbf) 0.54(poly) 0.23(sigmoid)	0.62(linear) 0.71(rbf) 0.73(poly) 0.21(sigmoid)
80	0.66(linear) 0.66(rbf) 0.70(poly) 0.19(sigmoid)	0.63(linear) 0.66(rbf) <b>0.72(poly)</b> 0.20(sigmoid)	0.47(linear) 0.50(rbf) 0.47(poly) 0.25(sigmoid)	0.63(linear) 0.73(rbf) 0.73(poly) 0.22(sigmoid)
90	0.67(linear) 0.65(rbf) <b>0.71(poly)</b> 0.19(sigmoid)	0.63(linear) 0.66(rbf) 0.71(poly) 0.24(sigmoid)	0.46(linear) 0.48(rbf) 0.47(poly) 0.24(sigmoid)	0.60(linear) 0.73(rbf) 0.71(poly) 0.21(sigmoid)
100	0.64(linear) 0.64(rbf) <b>0.71(poly)</b> 0.17(sigmoid)	0.70(linear) 0.63(rbf) 0.71(poly) 0.22(sigmoid)	0.48(linear) 0.50(rbf) 0.46(poly) 0.22(sigmoid)	0.61(linear) 0.75(rbf) 0.72(poly) 0.20(sigmoid)

Tabla 4.12: Resultados de recall usando el clasificador SVM.

## 4.2 Experimentos con autocorrelación de la base de datos EMOVO

### Experimento 1

#### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, descartando el filtro de preéñfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

#### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, sin preéñfasis, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

#### Resumen de la Tabla 4.13

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.13 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas MFCC, Entropy Signature y MSES mientras que para RASTA-MFCC el mejor promedio de recall se obtiene tanto con 24 como con 48 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.58	0.57	0.37	0.58
24	<b>0.69</b>	<b>0.71</b>	<b>0.46</b>	<b>0.68</b>
36	0.67	0.67	0.41	0.67
48	0.67	<b>0.71</b>	0.37	0.65

Tabla 4.13: Resultados de recall usando el clasificador MLP.

#### Resumen de la Tabla 4.14

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.14 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas MFCC, RASTA-MFCC, Entropy Signature y MSES. El número que aparece entre paréntesis en la Tabla 4.14 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.62(1)	0.55(5)	0.30(1)	0.61(1)
	0.58(3)	0.57(3)	0.26(3)	0.59(3)
	0.58(5)	0.57(1)	0.33(5)	0.58(5)
24	<b>0.68(1)</b>	<b>0.63(5)</b>	<b>0.38(1)</b>	<b>0.63(1)</b>
	0.65(3)	0.60(3)	0.36(3)	0.58(3)
	0.62(5)	<b>0.63(1)</b>	0.32(5)	0.63(5)
36	0.62(1)	0.59(5)	0.35(1)	0.62(1)
	0.63(3)	0.58(3)	0.31(3)	0.62(3)
	0.57(5)	0.62(1)	0.34(5)	0.60(5)
48	0.54(1)	0.53(5)	0.28(1)	0.53(1)
	0.51(3)	0.53(3)	0.26(3)	0.56(3)
	0.50(5)	0.57(1)	0.31(5)	0.50(5)

Tabla 4.14: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.15

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.15 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales MFCC y MSES mientras que para RASTA y Entropy Signature 36 y 48 coeficientes respectivamente. El texto que aparece entre paréntesis en la Tabla 4.15 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.44(linear) 0.56(rbf) 0.51(poly) 0.12(sigmoid)	0.38(linear) 0.57(rbf) 0.51(poly) 0.18(sigmoid)	0.21(linear) 0.31(rbf) 0.31(poly) 0.22(sigmoid)	0.38(linear) 0.58(rbf) 0.59(poly) 0.13(sigmoid)
24	0.47(linear) 0.54(rbf) <b>0.63(poly)</b> 0.08(sigmoid)	0.50(linear) 0.53(rbf) 0.60(poly) 0.13(sigmoid)	0.30(linear) 0.37(rbf) 0.35(poly) 0.13(sigmoid)	0.50(linear) <b>0.69(rbf)</b> 0.64(poly) 0.12(sigmoid)
36	0.51(linear) 0.39(rbf) 0.54(poly) 0.15(sigmoid)	0.48(linear) 0.45(rbf) <b>0.61(poly)</b> 0.11(sigmoid)	0.31(linear) 0.37(rbf) 0.38(poly) 0.10(sigmoid)	0.48(linear) 0.62(rbf) 0.58(poly) 0.11(sigmoid)
48	0.53(linear) 0.48(rbf) 0.56(poly) 0.17(sigmoid)	0.50(linear) 0.51(rbf) 0.58(poly) 0.19(sigmoid)	0.21(linear) 0.29(rbf) <b>0.41(poly)</b> 0.10(sigmoid)	0.42(linear) 0.56(rbf) 0.62(poly) 0.11(sigmoid)

Tabla 4.15: Resultados de recall usando el clasificador SVM.

## Experimento 2

### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, usando el filtro de preénfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

### Resumen de la Tabla 4.16

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.16 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas RASTA-MFCC, Entropy Signature y MSES mientras que para MFCC el mejor promedio de recall se obtiene con 48 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.60	0.55	0.40	0.59
24	0.67	<b>0.68</b>	<b>0.44</b>	<b>0.67</b>
36	0.64	0.64	0.40	0.64
48	<b>0.73</b>	0.67	0.34	0.66

Tabla 4.16: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.17

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.17 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas MFCC, RASTA-MFCC, Entropy Signature y MSES. El número que aparece entre paréntesis en la Tabla 4.17 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.63(1)	0.58(1)	0.31(1)	0.56(1)
	0.58(3)	0.53(3)	0.30(3)	0.54(3)
	0.55(5)	0.51(5)	0.35(5)	0.59(5)
24	<b>0.70(1)</b>	<b>0.65(1)</b>	<b>0.36(1)</b>	0.59(1)
	0.69(3)	0.63(3)	0.34(3)	0.59(3)
	0.64(5)	0.63(5)	0.34(5)	<b>0.64(5)</b>
36	0.62(1)	0.59(1)	0.35(1)	0.58(1)
	0.61(3)	0.59(3)	0.31(3)	0.62(3)
	0.58(5)	0.58(5)	0.35(5)	0.61(5)
48	0.59(1)	0.54(1)	0.33(1)	0.53(1)
	0.51(3)	0.50(3)	0.33(3)	0.54(3)
	0.51(5)	0.55(5)	0.30(5)	0.53(5)

Tabla 4.17: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.18

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.18 muestran que extrayendo 48 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales MFCC, RASTA-MFCC y Entropy Signature mientras que para MSES 24 coeficientes. El texto que aparece entre paréntesis en la Tabla 4.18 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.38(linear)	0.36(linear)	0.27(linear)	0.33(linear)
	0.54(rbf)	0.53(rbf)	0.31(rbf)	0.59(rbf)
	0.53(poly)	0.47(poly)	0.29(poly)	0.53(poly)
	0.13(sigmoid)	0.10(sigmoid)	0.21(sigmoid)	0.20(sigmoid)
24	0.49(linear)	0.48(linear)	0.34(linear)	0.45(linear)
	0.45(rbf)	0.45(rbf)	0.37(rbf)	<b>0.69(rbf)</b>
	0.57(poly)	0.52(poly)	0.35(poly)	0.54(poly)
	0.13(sigmoid)	0.14(sigmoid)	0.17(sigmoid)	0.17(sigmoid)
36	0.47(linear)	0.45(linear)	0.30(linear)	0.44(linear)
	0.33(rbf)	0.35(rbf)	0.37(rbf)	0.62(rbf)
	0.54(poly)	0.50(poly)	0.34(poly)	0.54(poly)
	0.15(sigmoid)	0.14(sigmoid)	0.13(sigmoid)	0.13(sigmoid)
48	0.44(linear)	0.51(linear)	0.19(linear)	0.45(linear)
	0.49(rbf)	0.45(rbf)	0.23(rbf)	0.58(rbf)
	<b>0.60(poly)</b>	<b>0.58(poly)</b>	<b>0.38(poly)</b>	0.51(poly)
	0.14(sigmoid)	0.14(sigmoid)	0.12(sigmoid)	0.15(sigmoid)

Tabla 4.18: Resultados de recall usando el clasificador SVM.

## Experimento 3

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 3 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y descartando el filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, sin preénfasis, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.19

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.19 muestran que usando un tamaño de frame de 80 ms se obtiene el mejor promedio de recall para la firma MFCC, 30 y 40 ms para RASTA-MFCC, 20 ms para Entropy Signature y 40, 80 y 90 ms para MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.66	0.66	0.42	0.68
20	0.70	0.69	<b>0.46</b>	0.69
30	0.70	<b>0.73</b>	0.45	0.69
40	0.70	<b>0.73</b>	0.44	<b>0.71</b>
50	0.71	0.71	0.45	0.70
60	0.70	0.70	0.39	0.70
70	0.71	0.69	0.38	0.70
80	<b>0.72</b>	0.70	0.41	<b>0.71</b>
90	0.71	0.69	0.41	<b>0.71</b>
100	0.71	0.69	0.39	0.70

Tabla 4.19: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.20

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.20 muestran que usando un tamaño de frame de 10, 20, 40 y 50 ms se obtiene el mejor puntaje de recall para la firma MFCC mientras que para RASTA-MFCC 40 y 50 ms, Entropy Signature 20 y 30ms y MSES 80 y 100 ms El número que aparece entre paréntesis en la Tabla 4.20 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.65(1)	0.61(1)	0.35(1)	0.61(1)
	0.65(3)	0.56(3)	0.35(3)	0.61(3)
	<b>0.69(5)</b>	0.59(5)	0.33(5)	0.64(5)
20	<b>0.69(1)</b>	0.62(1)	<b>0.38(1)</b>	0.63(1)
	0.63(5)	0.61(3)	0.35(3)	0.60(3)
	0.62(3)	0.62(5)	0.34(5)	0.63(5)
30	0.68(1)	0.63(1)	0.32(1)	0.63(1)
	0.65(3)	0.60(3)	0.36(3)	0.58(3)
	0.58(5)	0.62(5)	<b>0.38(5)</b>	0.63(5)
40	0.68(1)	<b>0.66(1)</b>	0.33(1)	0.59(1)
	<b>0.69(3)</b>	0.62(3)	0.35(3)	0.62(3)
	0.63(5)	0.63(5)	0.37(5)	0.63(5)
50	<b>0.69(1)</b>	<b>0.66(1)</b>	0.35(1)	0.62(1)
	0.64(3)	0.58(3)	0.36(3)	0.63(3)
	0.64(5)	0.63(5)	0.37(5)	0.63(5)
60	0.66(1)	0.65(1)	0.30(1)	0.62(1)
	0.62(3)	0.61(3)	0.32(3)	0.62(3)
	0.62(5)	0.58(5)	0.30(5)	0.63(5)
70	0.66(1)	0.62(1)	0.29(1)	0.61(1)
	0.63(3)	0.60(3)	0.30(3)	0.62(3)
	0.62(5)	0.60(5)	0.30(5)	0.63(5)
80	0.67(1)	0.65(1)	0.30(1)	0.62(1)
	0.62(3)	0.63(3)	0.25(3)	0.61(3)
	0.63(5)	0.62(5)	0.31(5)	<b>0.65(5)</b>
90	0.64(1)	0.63(1)	0.30(1)	0.62(1)
	0.63(3)	0.60(3)	0.29(3)	0.62(3)
	0.62(5)	0.61(5)	0.30(5)	0.63(5)
100	0.67(1)	0.65(1)	0.31(1)	0.63(1)
	0.62(3)	0.62(3)	0.29(3)	<b>0.65(3)</b>
	0.65(5)	0.61(5)	0.30(5)	0.63(5)

Tabla 4.20: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.21

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.21 muestran que usando un tamaño de frame de 80 ms se obtiene el mejor puntaje de recall para la firma MFCC, para RASTA-MFCC y Entropy Signature 20 ms y para MSES 30, 40 y 100 ms. El texto que aparece entre paréntesis en la Tabla 4.21 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.45(linear) 0.43(rbf) 0.57(poly) 0.13(sigmoid)	0.46(linear) 0.45(rbf) 0.56(poly) 0.13(sigmoid)	0.29(linear) 0.37(rbf) 0.37(poly) 0.17(sigmoid)	0.43(linear) 0.65(rbf) 0.58(poly) 0.18(sigmoid)
20	0.51(linear) 0.54(rbf) 0.61(poly) 0.11(sigmoid)	0.51(linear) 0.50(rbf) <b>0.62(poly)</b> 0.12(sigmoid)	0.30(linear) 0.38(rbf) <b>0.41(poly)</b> 0.14(sigmoid)	0.49(linear) 0.67(rbf) 0.61(poly) 0.12(sigmoid)
30	0.47(linear) 0.54(rbf) 0.63(poly) 0.08(sigmoid)	0.50(linear) 0.53(rbf) 0.60(poly) 0.13(sigmoid)	0.30(linear) 0.37(rbf) 0.35(poly) 0.13(sigmoid)	0.50(linear) <b>0.69(rbf)</b> 0.64(poly) 0.12(sigmoid)
40	0.47(linear) 0.54(rbf) 0.63(poly) 0.07(sigmoid)	0.50(linear) 0.54(rbf) 0.61(poly) 0.17(sigmoid)	0.33(linear) 0.37(rbf) 0.35(poly) 0.14(sigmoid)	0.48(linear) 0.67(rbf) <b>0.69(poly)</b> 0.11(sigmoid)
50	0.48(linear) 0.56(rbf) 0.63(poly) 0.11(sigmoid)	0.46(linear) 0.58(rbf) 0.61(poly) 0.17(sigmoid)	0.31(linear) 0.37(rbf) 0.37(poly) 0.13(sigmoid)	0.49(linear) 0.67(rbf) 0.65(poly) 0.10(sigmoid)
60	0.50(linear) 0.55(rbf) 0.61(poly) 0.09(sigmoid)	0.47(linear) 0.59(rbf) 0.61(poly) 0.18(sigmoid)	0.30(linear) 0.35(rbf) 0.31(poly) 0.11(sigmoid)	0.47(linear) 0.65(rbf) 0.67(poly) 0.10(sigmoid)
70	0.49(linear) 0.58(rbf) 0.62(poly) 0.09(sigmoid)	0.45(linear) 0.59(rbf) 0.59(poly) 0.15(sigmoid)	0.29(linear) 0.32(rbf) 0.31(poly) 0.15(sigmoid)	0.49(linear) 0.66(rbf) 0.66(poly) 0.12(sigmoid)
80	0.50(linear) 0.56(rbf) <b>0.65(poly)</b> 0.07(sigmoid)	0.46(linear) 0.59(rbf) 0.57(poly) 0.13(sigmoid)	0.30(linear) 0.37(rbf) 0.33(poly) 0.15(sigmoid)	0.47(linear) 0.67(rbf) 0.67(poly) 0.14(sigmoid)
90	0.48(linear) 0.57(rbf) 0.62(poly) 0.09(sigmoid)	0.46(linear) 0.59(rbf) 0.57(poly) 0.13(sigmoid)	0.33(linear) 0.37(rbf) 0.33(poly) 0.15(sigmoid)	0.48(linear) 0.67(rbf) 0.66(poly) 0.13(sigmoid)
100	0.50(linear) 0.58(rbf) 0.61(poly) 0.10(sigmoid)	0.48(linear) 0.58(rbf) 0.58(poly) 0.13(sigmoid)	0.31(linear) 0.35(rbf) 0.32(poly) 0.17(sigmoid)	0.48(linear) <b>0.69(rbf)</b> 0.66(rbf) 0.14(sigmoid)

Tabla 4.21: Resultados de recall usando el clasificador SVM.

## Experimento 4

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 4 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y usando filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.22

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.22 muestran que usando un tamaño de frame de 80 ms se obtiene el mejor promedio de recall para las firmas MFCC y MSES, 60 ms para RASTA-MFCC y 20 ms para Entropy Signature.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.66	0.65	0.41	0.61
20	0.67	0.66	<b>0.47</b>	0.65
30	0.67	0.68	0.44	0.67
40	0.69	0.68	0.45	0.66
50	0.70	0.70	0.44	0.68
60	0.69	<b>0.71</b>	0.38	0.68
70	0.69	0.70	0.39	0.70
80	<b>0.72</b>	0.70	0.39	<b>0.72</b>
90	0.69	0.68	0.39	0.70
100	0.71	0.70	0.38	0.71

Tabla 4.22: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.23

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.23 muestran que usando un tamaño de frame de 30 ms se obtiene el mejor puntaje de recall para la firma MFCC, con 30, 70 y 100 ms para MSES, 80 y 90 ms para RASTA-MFCC y 20 ms para Entropy Signature. El número que aparece entre paréntesis en la Tabla 4.23 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.61(1) 0.60(3) 0.55(5)	0.57(1) 0.57(3) 0.53(5)	0.34(1) 0.36(3) 0.35(5)	0.59(1) 0.62(3) 0.58(5)
20	0.69(1) 0.67(3) 0.66(5)	0.64(1) 0.62(3) 0.59(5)	0.35(1) <b>0.37(3)</b> 0.35(5)	0.63(1) 0.59(3) 0.62(5)
30	<b>0.70(1)</b> 0.69(3) 0.64(5)	0.65(1) 0.63(3) 0.63(5)	0.34(1) 0.34(3) 0.36(5)	<b>0.64(1)</b> 0.59(3) 0.59(5)
40	0.68(1) 0.66(3) 0.62(5)	0.65(1) 0.62(3) 0.61(5)	0.34(1) 0.30(3) 0.32(5)	0.61(1) 0.61(3) 0.61(5)
50	0.65(1) 0.67(3) 0.62(5)	0.63(1) 0.62(3) 0.59(5)	0.31(1) 0.31(3) 0.35(5)	0.59(1) 0.58(3) 0.59(5)
60	0.62(1) 0.66(3) 0.61(5)	0.64(1) 0.65(3) 0.59(5)	0.31(1) 0.30(3) 0.31(5)	0.60(1) 0.57(3) 0.61(5)
70	0.63(1) 0.65(3) 0.62(5)	0.65(1) 0.65(3) 0.61(5)	0.30(1) 0.24(3) 0.27(5)	0.63(1) 0.58(3) <b>0.64(5)</b>
80	0.63(1) 0.63(3) 0.58(5)	<b>0.67(1)</b> 0.64(3) 0.62(5)	0.33(1) 0.29(3) 0.35(5)	0.59(1) 0.58(3) 0.62(5)
90	0.65(1) 0.63(3) 0.58(5)	<b>0.67(1)</b> 0.62(3) 0.60(5)	0.27(1) 0.29(3) 0.27(5)	0.61(1) 0.59(3) 0.63(5)
100	0.65(1) 0.62(3) 0.58(5)	0.65(1) 0.63(3) 0.59(5)	0.30(1) 0.26(3) 0.32(5)	0.59(1) 0.59(3) <b>0.64(5)</b>

Tabla 4.23: Resultados de recall usando el clasificador KNN.

#### Resumen de la Tabla 4.24

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.24 muestran que usando un tamaño de frame de 60 ms se obtiene el mejor puntaje de recall para las firmas MFCC y RASTA-MFCC mientras que para Entropy Signature y MSES 40 ms. El texto que aparece entre paréntesis en la Tabla 4.24 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.45(linear) 0.37(rbf) 0.51(poly) 0.15(sigmoid)	0.40(linear) 0.34(rbf) 0.54(poly) 0.12(sigmoid)	0.28(linear) 0.33(rbf) 0.35(poly) 0.22(sigmoid)	0.42(linear) 0.57(rbf) 0.42(poly) 0.26(sigmoid)
20	0.46(linear) 0.40(rbf) 0.53(poly) 0.13(sigmoid)	0.45(linear) 0.42(rbf) 0.47(poly) 0.15(sigmoid)	0.35(linear) 0.40(rbf) 0.37(poly) 0.20(sigmoid)	0.45(linear) <b>0.70(rbf)</b> 0.58(poly) 0.17(sigmoid)
30	0.49(linear) 0.45(rbf) 0.57(poly) 0.13(sigmoid)	0.48(linear) 0.45(rbf) 0.52(poly) 0.14(sigmoid)	0.34(linear) 0.37(rbf) 0.35(poly) 0.17(sigmoid)	0.45(linear) 0.69(rbf) 0.54(poly) 0.17(sigmoid)
40	0.49(linear) 0.49(rbf) 0.59(poly) 0.11(sigmoid)	0.50(linear) 0.49(rbf) 0.58(poly) 0.14(sigmoid)	0.38(linear) <b>0.41(rbf)</b> 0.34(poly) 0.15(sigmoid)	0.43(linear) <b>0.70(rbf)</b> 0.56(poly) 0.16(sigmoid)
50	0.48(linear) 0.48(rbf) 0.61(poly) 0.10(sigmoid)	0.48(linear) 0.53(rbf) 0.57(poly) 0.14(sigmoid)	0.37(linear) 0.36(rbf) 0.35(poly) 0.19(sigmoid)	0.49(linear) 0.68(rbf) 0.57(poly) 0.15(sigmoid)
60	0.51(linear) 0.47(rbf) <b>0.65(poly)</b> 0.13(sigmoid)	0.44(linear) 0.59(rbf) <b>0.61(poly)</b> 0.13(sigmoid)	0.28(linear) 0.31(rbf) 0.33(poly) 0.12(sigmoid)	0.47(linear) 0.68(rbf) 0.59(poly) 0.14(sigmoid)
70	0.53(linear) 0.46(rbf) 0.64(poly) 0.10(sigmoid)	0.44(linear) 0.49(rbf) 0.58(poly) 0.14(sigmoid)	0.27(linear) 0.31(rbf) 0.35(poly) 0.12(sigmoid)	0.49(linear) 0.66(rbf) 0.58(poly) 0.14(sigmoid)
80	0.51(linear) 0.47(rbf) 0.61(poly) 0.12(sigmoid)	0.47(linear) 0.50(rbf) 0.55(poly) 0.12(sigmoid)	0.29(linear) 0.34(rbf) 0.36(poly) 0.17(sigmoid)	0.49(linear) 0.66(rbf) 0.59(poly) 0.12(sigmoid)
90	0.54(linear) 0.47(rbf) 0.61(poly) 0.10(sigmoid)	0.49(linear) 0.50(rbf) 0.60(poly) 0.13(sigmoid)	0.27(linear) 0.31(rbf) 0.37(poly) 0.14(sigmoid)	0.47(linear) 0.66(rbf) 0.63(poly) 0.14(sigmoid)
100	0.51(linear) 0.47(rbf) 0.61(poly) 0.15(sigmoid)	0.46(linear) 0.49(rbf) 0.57(poly) 0.12(sigmoid)	0.29(linear) 0.31(rbf) 0.31(poly) 0.15(sigmoid)	0.47(linear) 0.65(rbf) 0.61(poly) 0.15(sigmoid)

Tabla 4.24: Resultados de recall usando el clasificador SVM.

## 4.3 Mejores resultados de los experimentos con autocorrelación

A continuación, se presenta una tabla de cada una de las firmas espectrales que muestra los mejores puntajes recall obtenidos en los experimentos para clasificar las dos bases de datos.

Se puede observar que para MFCC el clasificador que más destaca es MLP teniendo los mejores resultados en las dos bases de datos, a excepción del experimento 4 donde se ve superior el clasificador KNN por una pequeña diferencia de un 0.01.

El mejor resultado para EMODB es de 0.81, se obtiene con MLP, en el experimento 3, con 10 ms, y para EMOVO es de 0.73, se obtiene con MLP, en el experimento 2, con 48 coeficientes.

MFCC				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	48 coeficientes	<b>0.80</b>	24 coeficientes	<b>0.69</b>
KNN	24 coeficientes	0.73(3)	24 coeficientes	68(1)
SVM	24 coeficientes	0.76(poly)	24 coeficientes	0.63(poly)
Experimento 2				
MLP	24 y 48 coeficientes	<b>0.77</b>	48 coeficientes	<b>0.73</b>
KNN	12 coeficientes	0.69(5)	24 coeficientes	0.70(1)
SVM	48 coeficientes	0.71(poly)	48 coeficientes	0.60(poly)
Experimento 3				
MLP	Frame de 10 ms	<b>0.81</b>	Frame de 80 ms	<b>0.72</b>
KNN	Frame de 10 ms	<b>0.81(5)</b>	Frame de 10, 20, 40 y 50 ms	0.69(1, 3, 5)
SVM	Frame de 10 ms	0.78(poly)	Frame de 80 ms	0.65(poly)
Experimento 4				
MLP	Frame de 40 ms	0.78	Frame de 80 ms	<b>0.72</b>
KNN	Frame de 20 ms	<b>0.79(5)</b>	Frame de 30 ms	0.70(1)
SVM	Frame de 100 ms	0.71(poly)	Frame de 60 ms	0.65(poly)

Tabla 4.25: Mejores resultados para la firma MFCC.

Se puede observar que para RASTA-MFCC el clasificador que más destaca es MLP teniendo los mejores resultados en las dos bases de datos.

El mejor resultado para EMODB es de 0.82, se obtiene con MLP, en el experimento 3, con 10 ms, y para EMOVO es de 0.73, se obtiene con MLP, en el experimento 3, con 50 ms.

RASTA-MFCC				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	24 coeficientes	<b>0.78</b>	24 y 48 coeficientes	<b>0.71</b>
KNN	12 coeficientes	0.66(3)	24 coeficientes	0.63(1, 5)
SVM	12 y 24 coeficientes	0.74(rbf)	36 coeficientes	0.61(poly)
Experimento 2				
MLP	24 coeficientes	<b>0.76</b>	24 coeficientes	<b>0.68</b>
KNN	12 y 24 coeficientes	0.68(1, 3)	24 coeficientes	0.65(1)
SVM	12 y 24 coeficientes	0.68(poly, rbf)	48 coeficientes	0.58(poly)
Experimento 3				
MLP	Frame de 10 ms	<b>0.82</b>	Frame de 50 ms	<b>0.73</b>
KNN	Frame de 10 ms	0.81(5)	Frame de 40 y 50 ms	0.66(1)
SVM	Frame de 50 ms	0.77(poly)	Frame de 20 ms	0.62(poly)
Experimento 4				
MLP	Frame de 80 ms	<b>0.77</b>	Frame de 60 ms	<b>0.71</b>
KNN	Frame de 20 ms	0.71(1)	Frame de 80 y 90 ms	0.67(1)
SVM	Frame de 80 ms	0.72(poly)	Frame de 60 ms	0.61(poly)

Tabla 4.26: Mejores resultados para la firma RASTA-MFCC.

Se puede observar que para Entropy Signature en la base de datos EMODB el clasificador que más destaca en los experimentos sin filtro de preénfasis (experimento 1 y 3) es SVM y en los experimentos con filtro de preénfasis (experimento 2 y 4) es MLP, mientras que para EMOVO el clasificador que más destaca para todos los experimentos es MLP.

El mejor resultado para EMODB es de 0.55, se obtiene en los 4 experimentos, con 24 coeficientes y 20, 30, 60 y 90 ms y para EMOVO es de 0.47, se obtiene con MLP en el experimento 4, con 20 ms.

Entropy Signature				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	24 coeficientes	0.53	24 coeficientes	<b>0.46</b>
KNN	24 coeficientes	0.46(1)	24 coeficientes	0.38(1)
SVM	24 coeficientes	<b>0.55(poly)</b>	48 coeficientes	0.41(poly)
Experimento 2				
MLP	24 coeficientes	<b>0.55</b>	24 coeficientes	<b>0.44</b>
KNN	24 coeficientes	0.50(5)	24 coeficientes	0.36(1)
SVM	12 y 48 coeficientes	0.52(poly)	48 coeficientes	0.38(poly)
Experimento 3				
MLP	Frame de 50 ms	0.54	Frame de 30 ms	<b>0.46</b>
KNN	Frame de 50 ms	0.50(5)	Frame de 20 y 30 ms	0.38(1, 5)
SVM	Frame de 30 ms	<b>0.55(poly)</b>	Frame de 20 ms	0.41(poly)
Experimento 4				
MLP	Frame de 90 ms	<b>0.55</b>	Frame de 20 ms	<b>0.47</b>
KNN	Frame de 30 ms	0.50(1)	Frame de 20 ms	0.37(3)
SVM	Frame de 20 y 60 ms	<b>0.55(poly)</b>	Frame de 40 ms	0.41(rbf)

Tabla 4.27: Mejores resultados para la firma Entropy Signature.

Se puede observar que para MSES el clasificador que más destaca para la base de datos EMODB es MLP, mientras que para EMOVO en los experimentos 1 y 2 es SVM y en los experimentos 3 y 4 MLP.

El mejor resultado para EMODB es de 0.79, se obtiene con MLP, en el experimento 3, con 20 ms, y para EMOVO es de 0.72, se obtiene con MLP, en el experimento 4, con 80 ms.

MSES				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	24 coeficientes	<b>0.77</b>	24 coeficientes	0.68
KNN	24 coeficientes	0.71(3)	24 coeficientes	0.63(1)
SVM	24 coeficientes	0.76(rbf)	24 coeficientes	<b>0.69(rbf)</b>
Experimento 2				
MLP	48 coeficientes	<b>0.77</b>	24 coeficientes	0.67
KNN	24 coeficientes	0.69(1,3)	24 coeficientes	0.64(5)
SVM	24 coeficientes	0.76(rbf)	24 coeficientes	<b>0.69(rbf)</b>
Experimento 3				
MLP	Frame de 20 ms	<b>0.79</b>	Frame de 40 ms	<b>0.71</b>
KNN	Frame de 20 ms	0.76(3)	Frame de 80 y 100 ms	0.65(3, 5)
SVM	Frame de 20 ms	0.78(rbf)	Frame de 30, 40 y 100 ms	0.69(rbf, poly)
Experimento 4				
MLP	Frame de 60 ms	<b>0.77</b>	Frame de 80 ms	<b>0.72</b>
KNN	Frame de 40 y 70 ms	0.74(1)	Frame de 30, 70 y 100 ms	0.64(1, 5)
SVM	Frame de 30 ms	0.76(rbf)	Frame de 40 ms	0.70(rbf)

Tabla 4.28: Mejores resultados para la firma MFCC.

## 4.4 Experimentos sin autocorrelación de la base de datos EMODB

### Experimento 1

#### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, descartando el filtro de preéñfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

#### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, sin preéñfasis, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

#### Resumen de la Tabla 4.29

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.29 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas Entropy Signature y MSES mientras que para RASTA-MFCC y MFCC el mejor promedio de recall se obtiene con 48 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.70	0.70	0.42	0.71
24	0.77	0.74	<b>0.48</b>	<b>0.77</b>
36	0.76	0.72	0.45	0.75
48	<b>0.78</b>	<b>0.77</b>	0.41	0.76

Tabla 4.29: Resultados de recall usando el clasificador MLP.

#### Resumen de la Tabla 4.30

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.30 muestran que extrayendo 48 coeficientes se obtiene el mejor puntaje de recall para la firma Entropy Signature mientras que para MFCC, RASTA-MFCC y MSES el mejor promedio de recall se obtiene con 24 coeficientes. El número que aparece entre paréntesis en la Tabla 4.30 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.63(1) 0.66(3) 0.66(5)	0.61(1) 0.67(3) 0.63(5)	0.35(1) 0.33(3) 0.43(5)	0.64(1) 0.68(3) 0.68(5)
24	0.70(1) 0.70(3) <b>0.73(5)</b>	0.65(1) <b>0.71(3)</b> <b>0.71(5)</b>	0.36(1) 0.37(3) 0.39(5)	0.71(1) 0.73(3) <b>0.77(5)</b>
36	0.65(1) 0.63(3) 0.65(5)	0.63(1) 0.66(3) 0.68(5)	0.36(1) 0.35(3) 0.42(5)	0.67(1) 0.61(3) 0.64(5)
48	0.50(1) 0.49(3) 0.52(5)	0.58(1) 0.55(3) 0.53(5)	<b>0.47(1)</b> 0.42(3) 0.44(5)	0.55(1) 0.55(3) 0.58(5)

Tabla 4.30: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.31

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.31 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las 4 firmas espectrales MFCC, RASTA-MFCC, Entropy Signature y MSES. El texto que aparece entre paréntesis en la Tabla 4.31 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.60(linear) 0.68(rbf) 0.71(poly) 0.23(sigmoid)	0.55(linear) 0.69(rbf) 0.68(poly) 0.34(sigmoid)	0.36(linear) 0.39(rbf) 0.42(poly) 0.22(sigmoid)	0.58(linear) 0.68(rbf) 0.72(poly) 0.22(sigmoid)
24	0.71(linear) <b>0.75(rbf)</b> 0.73(poly) 0.25(sigmoid)	0.68(linear) <b>0.76(rbf)</b> 0.71(poly) 0.29(sigmoid)	0.40(linear) 0.43(rbf) <b>0.52(poly)</b> 0.22(sigmoid)	0.69(linear) <b>0.78(rbf)</b> 0.73(poly) 0.25(sigmoid)
36	0.66(linear) 0.65(rbf) 0.71(poly) 0.21(sigmoid)	0.66(linear) 0.66(rbf) 0.66(poly) 0.20(sigmoid)	0.39(linear) 0.45(rbf) 0.47(poly) 0.24(sigmoid)	0.65(linear) 0.70(rbf) 0.68(poly) 0.26(sigmoid)
48	0.68(linear) 0.58(rbf) 0.61(poly) 0.25(sigmoid)	0.66(linear) 0.63(rbf) 0.63(poly) 0.24(sigmoid)	0.31(linear) 0.30(rbf) 0.49(poly) 0.24(sigmoid)	0.66(linear) 0.65(rbf) 0.66(poly) 0.24(sigmoid)

Tabla 4.31: Resultados de recall usando el clasificador SVM.

## Experimento 2

### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, usando el filtro de preénfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

### Resumen de la Tabla 4.32

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.32 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas MFCC, RASTA-MFCC y MSES mientras que para Entropy Signature el mejor promedio de recall se obtiene con 12 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.68	0.70	<b>0.50</b>	0.71
24	<b>0.77</b>	<b>0.73</b>	0.48	<b>0.76</b>
36	0.74	0.71	0.47	0.74
48	0.72	0.70	0.42	0.73

Tabla 4.32: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.33

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.33 muestran que extrayendo 48 coeficientes se obtiene el mejor puntaje de recall para la firma Entropy Signature mientras que para MFCC, RASTA-MFCC y MSES el mejor promedio de recall se obtiene con 24 coeficientes. El número que aparece entre paréntesis en la Tabla 4.33 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.62(1)	0.60(1)	0.42(1)	0.63(1)
	0.68(3)	0.68(3)	0.41(3)	0.68(3)
	0.67(5)	0.61(5)	0.45(5)	0.70(5)
24	0.69(1)	0.69(1)	0.40(1)	0.70(1)
	0.71(3)	0.69(3)	0.41(3)	0.73(3)
	<b>0.73(5)</b>	<b>0.72(5)</b>	0.48(5)	<b>0.75(5)</b>
36	0.65(1)	0.65(1)	0.39(1)	0.68(1)
	0.65(3)	0.65(3)	0.40(3)	0.64(3)
	0.63(5)	0.64(5)	0.39(5)	0.66(5)
48	0.54(1)	0.59(1)	0.45(1)	0.60(1)
	0.50(3)	0.55(3)	0.48(3)	0.58(3)
	0.50(5)	0.57(5)	<b>0.50(5)</b>	0.61(5)

Tabla 4.33: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.34

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.34 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales MFCC, RASTA-MFCC y MSES mientras que para Entropy Signature 48 coeficientes. El texto que aparece entre paréntesis en la Tabla 4.34 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.61(linear)	0.58(linear)	0.40(linear)	0.57(linear)
	0.70(rbf)	0.70(rbf)	0.42(rbf)	0.68(rbf)
	0.72(poly)	0.71(poly)	0.47(poly)	0.71(poly)
	0.23(sigmoid)	0.31(sigmoid)	0.20(sigmoid)	0.27(sigmoid)
24	0.71(linear)	0.70(linear)	0.45(linear)	0.69(linear)
	<b>0.74(rbf)</b>	<b>0.73(rbf)</b>	0.49(rbf)	<b>0.77(rbf)</b>
	0.70(poly)	0.70(poly)	0.50(poly)	0.73(poly)
	0.27(sigmoid)	0.24(sigmoid)	0.24(sigmoid)	0.26(sigmoid)
36	0.65(linear)	0.64(linear)	0.42(linear)	0.65(linear)
	0.67(rbf)	0.66(rbf)	0.47(rbf)	0.68(rbf)
	0.68(poly)	0.70(poly)	0.50(poly)	0.68(poly)
	0.21(sigmoid)	0.24(sigmoid)	0.24(sigmoid)	0.24(sigmoid)
48	0.63(linear)	0.59(linear)	0.31(linear)	0.60(linear)
	0.60(rbf)	0.64(rbf)	0.31(rbf)	0.63(rbf)
	0.63(poly)	0.61(poly)	<b>0.52(poly)</b>	0.66(poly)
	0.24(sigmoid)	0.24(sigmoid)	0.24(sigmoid)	0.24(sigmoid)

Tabla 4.34: Resultados de recall usando el clasificador SVM.

## Experimento 3

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 3 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y descartando el filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, sin preénfasis, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.35

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.35 muestran que usando un tamaño de frame de 90 ms se obtiene el mejor promedio de recall para la firma MFCC, 70 ms para RASTA-MFCC, 50 ms para Entropy Signature y 30 ms para MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.77	0.77	0.13	0.65
20	0.77	0.77	0.51	0.77
30	0.76	0.75	0.50	<b>0.78</b>
40	0.77	0.76	0.46	0.76
50	0.76	0.74	<b>0.53</b>	0.75
60	0.76	0.75	0.51	0.75
70	0.77	<b>0.79</b>	0.49	0.75
80	0.76	0.78	0.47	0.75
90	<b>0.78</b>	0.77	0.45	0.74
100	0.76	0.76	0.47	0.76

Tabla 4.35: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.36

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.36 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor puntaje de recall para las firmas MFCC y RASTA-MFCC con 30 ms para MSES, con 50 y 80 ms para Entropy Signature. El número que aparece entre paréntesis en la Tabla 4.36 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	<b>0.78(1)</b>	0.73(1)	0.12(1)	0.52(1)
	0.76(3)	<b>0.76(3)</b>	0.12(3)	0.55(3)
	0.74(5)	0.75(5)	0.12(5)	0.57(5)
20	0.70(1)	0.65(1)	0.39(1)	0.72(1)
	0.73(3)	0.71(3)	0.44(3)	0.75(3)
	0.73(5)	0.73(5)	0.38(5)	0.75(5)
30	0.70(1)	0.65(1)	0.36(1)	0.71(1)
	0.70(3)	0.71(3)	0.37(3)	0.73(3)
	0.73(5)	0.71(5)	0.39(5)	<b>0.77(5)</b>
40	0.67(1)	0.66(1)	0.37(1)	0.70(1)
	0.68(3)	0.70(3)	0.40(3)	0.73(3)
	0.71(5)	0.70(5)	0.45(5)	0.75(5)
50	0.66(1)	0.64(1)	<b>0.48(1)</b>	0.66(1)
	0.70(3)	0.71(3)	<b>0.48(3)</b>	0.72(3)
	0.70(5)	0.70(5)	0.45(5)	0.75(5)
60	0.68(1)	0.65(1)	0.38(1)	0.68(1)
	0.70(3)	0.69(3)	0.43(3)	0.72(3)
	0.70(5)	0.70(5)	0.46(5)	0.71(5)
70	0.67(1)	0.65(1)	0.44(1)	0.70(1)
	0.68(3)	0.71(3)	0.42(3)	0.70(3)
	0.71(5)	0.73(5)	0.43(5)	0.72(5)
80	0.66(1)	0.65(1)	0.44(1)	0.70(1)
	0.71(3)	0.69(3)	0.42(3)	0.73(3)
	0.70(5)	0.70(5)	<b>0.48(5)</b>	0.68(5)
90	0.71(1)	0.68(1)	0.43(1)	0.70(1)
	0.68(3)	0.67(3)	0.42(3)	0.71(3)
	0.71(5)	0.67(5)	0.42(5)	0.70(5)
100	0.68(1)	0.68(1)	0.39(1)	0.68(1)
	0.68(3)	0.74(3)	0.39(3)	0.71(3)
	0.70(5)	0.66(5)	0.43(5)	0.70(5)

Tabla 4.36: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.37

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.37 muestran que usando un tamaño de frame de 20 ms se obtiene el mejor puntaje de recall para las firmas MFCC y MSES mientras que para RASTA-MFCC con 10 ms y Entropy Signature 30 y 50 ms. El texto que aparece entre paréntesis en la Tabla 4.37 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.70(linear) 0.76(rbf) 0.71(poly) 0.25(sigmoid)	0.69(linear) <b>0.80(rbf)</b> 0.72(poly) 0.23(sigmoid)	0.13(linear) 0.13(rbf) 0.13(poly) 0.13(sigmoid)	0.41(linear) 0.48(rbf) 0.47(poly) 0.38(sigmoid)
20	0.70(linear) <b>0.78(rbf)</b> 0.75(poly) 0.27(sigmoid)	0.67(linear) 0.78(rbf) 0.73(poly) 0.27(sigmoid)	0.45(linear) 0.46(rbf) 0.50(poly) 0.24(sigmoid)	0.72(linear) <b>0.79(rbf)</b> 0.71(poly) 0.32(sigmoid)
30	0.71(linear) 0.75(rbf) 0.73(poly) 0.25(sigmoid)	0.68(linear) 0.76(rbf) 0.71(poly) 0.29(sigmoid)	0.40(linear) 0.43(rbf) <b>0.52(poly)</b> 0.22(sigmoid)	0.69(linear) 0.78(rbf) 0.73(poly) 0.25(sigmoid)
40	0.69(linear) 0.73(rbf) 0.70(poly) 0.26(sigmoid)	0.68(linear) 0.75(rbf) 0.72(poly) 0.26(sigmoid)	0.42(linear) 0.44(rbf) 0.43(poly) 0.24(sigmoid)	0.68(linear) 0.76(rbf) 0.74(poly) 0.27(sigmoid)
50	0.67(linear) 0.73(rbf) 0.68(poly) 0.26(sigmoid)	0.66(linear) 0.73(rbf) 0.72(poly) 0.22(sigmoid)	0.48(linear) <b>0.52(rbf)</b> <b>0.52(poly)</b> 0.21(sigmoid)	0.69(linear) 0.76(rbf) 0.71(poly) 0.24(sigmoid)
60	0.70(linear) 0.73(rbf) 0.67(poly) 0.24(sigmoid)	0.65(linear) 0.73(rbf) 0.68(poly) 0.25(sigmoid)	0.48(linear) 0.47(rbf) 0.48(poly) 0.23(sigmoid)	0.67(linear) 0.75(rbf) 0.68(poly) 0.22(sigmoid)
70	0.65(linear) 0.73(rbf) 0.68(poly) 0.24(sigmoid)	0.66(linear) 0.75(rbf) 0.71(poly) 0.21(sigmoid)	0.42(linear) 0.46(rbf) 0.51(poly) 0.24(sigmoid)	0.68(linear) 0.75(rbf) 0.67(poly) 0.22(sigmoid)
80	0.68(linear) 0.72(rbf) 0.67(poly) 0.24(sigmoid)	0.68(linear) 0.72(rbf) 0.72(poly) 0.22(sigmoid)	0.41(linear) 0.46(rbf) 0.45(poly) 0.22(sigmoid)	0.66(linear) 0.75(rbf) 0.68(poly) 0.22(sigmoid)
90	0.66(linear) 0.71(rbf) 0.67(poly) 0.25(sigmoid)	0.65(linear) 0.70(rbf) 0.69(poly) 0.21(sigmoid)	0.43(linear) 0.45(rbf) 0.48(poly) 0.24(sigmoid)	0.65(linear) 0.71(rbf) 0.69(poly) 0.22(sigmoid)
100	0.67(linear) 0.74(rbf) 0.68(poly) 0.25(sigmoid)	0.70(linear) 0.71(rbf) 0.70(poly) 0.20(sigmoid)	0.44(linear) 0.48(rbf) 0.47(poly) 0.24(sigmoid)	0.68(linear) 0.73(rbf) 0.69(poly) 0.24(sigmoid)

Tabla 4.37: Resultados de recall usando el clasificador SVM.

## Experimento 4

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 4 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y usando filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.38

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.38 muestran que usando un tamaño de frame de 80 ms se obtiene el mejor promedio de recall para la firma MFCC, 20 ms para RASTA-MFCC, 50 y 60 ms para Entropy Signature y 20 ms para MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.75	0.77	0.13	0.67
20	0.76	<b>0.78</b>	0.49	<b>0.78</b>
30	0.76	0.74	0.49	0.77
40	0.76	0.73	0.48	0.75
50	0.76	0.74	<b>0.51</b>	0.75
60	0.75	0.76	<b>0.51</b>	0.74
70	0.75	0.77	0.49	0.75
80	<b>0.77</b>	0.76	0.47	0.75
90	0.75	0.77	0.49	0.73
100	0.76	0.76	0.50	0.73

Tabla 4.38: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.39

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.39 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor puntaje de recall para las firmas MFCC y RASTA-MFCC, 20 y 30 ms para MSES y 60 y 70 ms para Entropy Signature. El número que aparece entre paréntesis en la Tabla 4.39 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	<b>0.75(1)</b>	0.74(1)	0.13(1)	0.46(1)
	0.74(3)	0.74(3)	0.13(3)	0.47(3)
	0.73(5)	<b>0.76(5)</b>	0.13(5)	0.52(5)
20	0.68(1)	0.65(1)	0.41(1)	0.69(1)
	0.72(3)	0.68(3)	0.42(3)	0.72(3)
	0.74(5)	0.74(5)	0.41(5)	<b>0.75(5)</b>
30	0.69(1)	0.69(1)	0.40(1)	0.70(1)
	0.71(3)	0.69(3)	0.41(3)	0.73(3)
	0.73(5)	0.72(5)	0.48(5)	<b>0.75(5)</b>
40	0.68(1)	0.66(1)	0.41(1)	0.68(1)
	0.70(3)	0.70(3)	0.48(3)	0.70(3)
	0.71(5)	0.68(5)	0.47(5)	0.74(5)
50	0.68(1)	0.68(1)	0.48(1)	0.70(1)
	0.71(3)	0.71(3)	0.42(3)	0.71(3)
	0.71(5)	0.68(5)	0.44(5)	0.74(5)
60	0.68(1)	0.65(1)	0.44(1)	0.70(1)
	0.68(3)	0.68(3)	0.44(3)	0.71(3)
	0.68(5)	0.66(5)	<b>0.50(5)</b>	0.71(5)
70	0.69(1)	0.67(1)	<b>0.50(1)</b>	0.70(1)
	0.68(3)	0.67(3)	0.42(3)	0.68(3)
	0.70(5)	0.66(5)	0.40(5)	0.71(5)
80	0.66(1)	0.66(1)	0.42(1)	0.70(1)
	0.68(3)	0.70(3)	0.36(3)	0.71(3)
	0.67(5)	0.65(5)	0.40(5)	0.69(5)
90	0.69(1)	0.68(1)	0.32(1)	0.71(1)
	0.68(3)	0.65(3)	0.37(3)	0.68(3)
	0.69(5)	0.68(5)	0.45(5)	0.68(5)
100	0.70(1)	0.66(1)	0.44(1)	0.69(1)
	0.68(3)	0.72(3)	0.40(3)	0.71(3)
	0.69(5)	0.65(5)	0.46(5)	0.68(5)

Tabla 4.39: Resultados de recall usando el clasificador KNN.

#### Resumen de la Tabla 4.40

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.40 muestran que usando un tamaño de frame de 10 ms se obtiene el mejor puntaje de recall para las firmas MFCC y RASTA-MFCC mientras que para Entropy Signature 50ms y MSES 20 ms. El texto que aparece entre paréntesis en la Tabla 4.40 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.66(linear) <b>0.78(rbf)</b> 0.71(poly) 0.30(sigmoid)	0.66(linear) <b>0.78(rbf)</b> 0.74(poly) 0.29(sigmoid)	0.13(linear) 0.13(rbf) 0.13(poly) 0.13(sigmoid)	0.39(linear) 0.44(rbf) 0.40(poly) 0.35(sigmoid)
20	0.70(linear) 0.77(rbf) 0.72(poly) 0.25(sigmoid)	0.68(linear) 0.73(rbf) 0.71(poly) 0.27(sigmoid)	0.43(linear) 0.45(rbf) 0.48(poly) 0.24(sigmoid)	0.67(linear) <b>0.80(rbf)</b> 0.71(poly) 0.31(sigmoid)
30	0.71(linear) 0.74(rbf) 0.70(poly) 0.27(sigmoid)	0.70(linear) 0.73(rbf) 0.70(poly) 0.24(sigmoid)	0.45(linear) 0.49(rbf) 0.50(poly) 0.24(sigmoid)	0.69(linear) 0.77(rbf) 0.73(poly) 0.26(sigmoid)
40	0.71(linear) 0.71(rbf) 0.71(poly) 0.22(sigmoid)	0.68(linear) 0.72(rbf) 0.69(poly) 0.20(sigmoid)	0.47(linear) 0.50(rbf) 0.54(poly) 0.24(sigmoid)	0.68(linear) 0.74(rbf) 0.72(poly) 0.22(sigmoid)
50	0.70(linear) 0.71(rbf) 0.70(poly) 0.24(sigmoid)	0.67(linear) 0.71(rbf) 0.69(poly) 0.25(sigmoid)	0.45(linear) 0.49(rbf) <b>0.55(poly)</b> 0.22(sigmoid)	0.69(linear) 0.73(rbf) 0.71(poly) 0.21(sigmoid)
60	0.70(linear) 0.71(rbf) 0.68(poly) 0.20(sigmoid)	0.66(linear) 0.71(rbf) 0.70(poly) 0.21(sigmoid)	0.47(linear) 0.47(rbf) 0.48(poly) 0.24(sigmoid)	0.68(linear) 0.71(rbf) 0.67(poly) 0.19(sigmoid)
70	0.69(linear) 0.72(rbf) 0.71(poly) 0.23(sigmoid)	0.70(linear) 0.71(rbf) 0.70(poly) 0.20(sigmoid)	0.44(linear) 0.48(rbf) 0.51(poly) 0.24(sigmoid)	0.68(linear) 0.71(rbf) 0.71(poly) 0.21(sigmoid)
80	0.68(linear) 0.70(rbf) 0.69(poly) 0.21(sigmoid)	0.68(linear) 0.70(rbf) 0.68(poly) 0.23(sigmoid)	0.44(linear) 0.48(rbf) 0.47(poly) 0.25(sigmoid)	0.68(linear) 0.74(rbf) 0.69(poly) 0.21(sigmoid)
90	0.65(linear) 0.70(rbf) 0.68(poly) 0.22(sigmoid)	0.68(linear) 0.70(rbf) 0.70(poly) 0.22(sigmoid)	0.46(linear) 0.48(rbf) 0.48(poly) 0.24(sigmoid)	0.66(linear) 0.70(rbf) 0.70(poly) 0.22(sigmoid)
100	0.67(linear) 0.72(rbf) 0.68(poly) 0.20(sigmoid)	0.68(linear) 0.71(rbf) 0.71(poly) 0.22(sigmoid)	0.46(linear) 0.48(rbf) 0.49(poly) 0.24(sigmoid)	0.67(linear) 0.71(rbf) 0.71(poly) 0.19(sigmoid)

Tabla 4.40: Resultados de recall usando el clasificador SVM.

## 4.5 Experimentos sin autocorrelación de la base de datos EMOVO

### Experimento 1

#### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, descartando el filtro de preéñfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

#### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, sin preéñfasis, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

#### Resumen de la Tabla 4.41

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.41 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas RASTA-MFCC y Entropy Signature mientras que para MFCC y MSES el mejor promedio de recall se obtiene con 36 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.61	0.56	0.41	0.60
24	0.72	<b>0.72</b>	<b>0.50</b>	0.71
36	<b>0.73</b>	0.70	0.49	<b>0.73</b>
48	0.63	0.68	0.25	0.70

Tabla 4.41: Resultados de recall usando el clasificador MLP.

#### Resumen de la Tabla 4.42

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.42 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas MFCC, RASTA-MFCC, Entropy Signature y MSES. El número que aparece entre paréntesis en la Tabla 4.42 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.62(1) 0.56(3) 0.61(5)	0.58(1) 0.58(3) 0.63(5)	0.38(1) 0.34(3) 0.31(5)	0.59(1) 0.59(3) 0.63(5)
24	0.67(1) <b>0.69(3)</b> 0.65(5)	<b>0.67(1)</b> <b>0.67(3)</b> 0.65(5)	<b>0.42(1)</b> 0.37(3) 0.43(5)	<b>0.67(1)</b> 0.66(3) 0.63(5)
36	0.64(1) 0.63(3) 0.57(5)	0.63(1) 0.61(3) 0.61(5)	0.37(1) 0.37(3) 0.38(5)	0.65(1) 0.61(3) 0.56(5)
48	0.50(1) 0.46(3) 0.47(5)	0.60(1) 0.55(3) 0.57(5)	0.39(1) 0.36(3) 0.35(5)	0.58(1) 0.49(3) 0.52(5)

Tabla 4.42: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.43

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.43 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales MFCC, RASTA-MFCC y MSES mientras que para Entropy Signature 36 coeficientes. El texto que aparece entre paréntesis en la Tabla 4.43 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.38(linear) 0.64(rbf) 0.59(poly) 0.13(sigmoid)	0.38(linear) 0.63(rbf) 0.57(poly) 0.15(sigmoid)	0.31(linear) 0.28(rbf) 0.33(poly) 0.15(sigmoid)	0.37(linear) 0.63(rbf) 0.53(poly) 0.17(sigmoid)
24	0.55(linear) <b>0.70(rbf)</b> 0.67(poly) 0.15(sigmoid)	0.54(linear) 0.67(rbf) <b>0.70(poly)</b> 0.15(sigmoid)	0.35(linear) 0.41(rbf) 0.45(poly) 0.12(sigmoid)	0.52(linear) <b>0.69(rbf)</b> 0.65(poly) 0.15(sigmoid)
36	0.50(linear) 0.63(rbf) 0.59(poly) 0.10(sigmoid)	0.51(linear) 0.61(rbf) 0.67(poly) 0.12(sigmoid)	0.33(linear) 0.38(rbf) <b>0.49(poly)</b> 0.15(sigmoid)	0.50(linear) 0.63(rbf) 0.63(poly) 0.09(sigmoid)
48	0.41(linear) 0.54(rbf) 0.54(poly) 0.14(sigmoid)	0.42(linear) 0.54(rbf) 0.63(poly) 0.09(sigmoid)	0.17(linear) 0.15(rbf) 0.46(poly) 0.14(sigmoid)	0.39(linear) 0.52(rbf) 0.57(poly) 0.14(sigmoid)

Tabla 4.43: Resultados de recall usando el clasificador MLP.

## Experimento 2

### Objetivo

Encontrar el número de coeficientes en los vectores característicos que hacen obtener el mayor promedio de recall, usando el filtro de preénfasis, ya que este número será usado en la configuración de los experimentos 3 y 4.

### Configuración

Cantidad de coeficientes 12, 24, 36 y 48, Usando la función ventana de Hann, con preénfasis de 0.97, ancho de banda completo, tamaño de frame de 30 ms y traslape de 50%.

### Resumen de la Tabla 4.44

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.44 muestran que extrayendo 24 coeficientes se obtiene el mejor promedio de recall para las firmas RASTA-MFCC, Entropy Signature y MSES, mientras que para MFCC el mejor promedio de recall se obtiene con 48 coeficientes.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.59	0.56	0.39	0.57
24	0.72	<b>0.74</b>	<b>0.52</b>	<b>0.73</b>
36	0.73	0.71	0.51	0.71
48	<b>0.75</b>	0.68	0.27	0.67

Tabla 4.44: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.45

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.45 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas MFCC, RASTA-MFCC, Entropy Signature y MSES para esta última también se obtiene el mismo promedio de recall con 36 coeficientes. El número que aparece entre paréntesis en la Tabla 4.45 es el número de vecinos utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.61(1) 0.59(3) 0.58(5)	0.59(1) 0.54(3) 0.54(5)	0.35(1) 0.35(3) 0.33(5)	0.61(1) 0.58(3) 0.62(5)
24	0.69(1) <b>0.70(3)</b> 0.66(5)	0.66(1) <b>0.67(3)</b> <b>0.67(5)</b>	<b>0.50(1)</b> 0.38(3) 0.35(5)	<b>0.65(1)</b> 0.63(3) 0.63(5)
36	0.65(1) 0.62(3) 0.61(5)	0.62(1) 0.57(3) 0.58(5)	0.42(1) 0.41(3) 0.34(5)	<b>0.65(1)</b> 0.58(3) 0.61(5)
48	0.60(1) 0.58(3) 0.55(5)	0.59(1) 0.59(3) 0.64(5)	0.41(1) 0.36(3) 0.39(5)	0.55(1) 0.57(3) 0.59(5)

Tabla 4.45: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.46

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.46 muestran que extrayendo 24 coeficientes se obtiene el mejor puntaje de recall para las firmas espectrales MFCC, RASTA-MFCC y MSES mientras que para Entropy Signature 48 coeficientes. El texto que aparece entre paréntesis en la Tabla 4.46 es el kernel utilizado para conseguir el puntaje de recall.

Número de Coeficientes	MFCC	RASTA-MFCC	Entropy Signature	MSES
12	0.36(linear) 0.66(rbf) 0.57(poly) 0.13(sigmoid)	0.39(linear) 0.61(rbf) 0.53(poly) 0.13(sigmoid)	0.32(linear) 0.35(rbf) 0.38(poly) 0.14(sigmoid)	0.37(linear) 0.61(rbf) 0.57(poly) 0.11(sigmoid)
24	0.51(linear) <b>0.71(rbf)</b> 0.69(poly) 0.14(sigmoid)	0.53(linear) 0.67(rbf) <b>0.69(poly)</b> 0.16(sigmoid)	0.37(linear) 0.43(rbf) 0.42(poly) 0.18(sigmoid)	0.50(linear) <b>0.70(rbf)</b> 0.66(poly) 0.18(sigmoid)
36	0.46(linear) 0.62(rbf) 0.59(poly) 0.11(sigmoid)	0.46(linear) 0.56(rbf) 0.66(poly) 0.13(sigmoid)	0.37(linear) 0.42(rbf) 0.42(poly) 0.16(sigmoid)	0.46(linear) 0.63(rbf) 0.64(poly) 0.09(sigmoid)
48	0.43(linear) 0.65(rbf) 0.64(poly) 0.09(sigmoid)	0.41(linear) 0.59(rbf) 0.58(poly) 0.14(sigmoid)	0.16(linear) 0.17(rbf) <b>0.46(poly)</b> 0.14(sigmoid)	0.41(linear) 0.56(rbf) 0.59(poly) 0.14(sigmoid)

Tabla 4.46: Resultados de recall usando el clasificador SVM.

## Experimento 3

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 3 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y descartando el filtro de preénfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, sin preénfasis, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.47

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.47 muestran que usando un tamaño de frame de 20 ms se obtiene el mejor promedio de recall para la firma MFCC, 40 ms para RASTA-MFCC y 50 ms para Entropy Signature y MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.70	0.70	0.50	0.70
20	<b>0.73</b>	0.72	0.48	0.72
30	0.71	0.71	0.49	0.73
40	0.72	<b>0.74</b>	0.52	0.73
50	0.72	0.72	<b>0.52</b>	<b>0.74</b>
60	0.71	0.72	0.44	0.72
70	0.71	0.73	0.44	0.73
80	0.71	0.73	0.45	0.73
90	0.72	0.72	0.44	0.73
100	0.70	0.73	0.44	0.71

Tabla 4.47: Resultados de recall usando el clasificador MLP.

#### Resumen de la Tabla 4.48

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.48 muestran que usando un tamaño de frame de 10, 30, 40, 50 y 70 ms se obtiene el mejor puntaje de recall para la firma MFCC, con 20 y 30 ms para MSES, 70 y 80 ms para RASTA-MFCC y 20 ms para Entropy Signature. El número que aparece entre paréntesis en la Tabla 4.48 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	<b>0.69(1)</b> 0.68(3) 0.65(5)	0.66(1) 0.64(3) 0.60(5)	0.38(1) 0.38(3) 0.39(5)	0.67(1) 0.63(3) 0.59(5)
20	0.68(1) 0.67(3) 0.68(5)	0.69(1) 0.63(3) 0.63(5)	<b>0.45(1)</b> 0.41(3) 0.42(5)	<b>0.70(1)</b> 0.66(3) 0.63(5)
30	0.67(1) <b>0.69(3)</b> 0.65(5)	0.67(1) 0.67(3) 0.65(5)	0.44(1) 0.38(3) 0.41(5)	<b>0.70(1)</b> 0.69(3) 0.64(5)
40	0.67(1) <b>0.69(3)</b> 0.65(5)	0.69(1) 0.68(3) 0.65(5)	0.43(1) 0.39(3) 0.41(5)	0.69(1) 0.66(3) 0.65(5)
50	<b>0.69(1)</b> <b>0.69(3)</b> 0.66(5)	0.66(1) 0.65(3) 0.65(5)	0.43(1) 0.35(3) 0.39(5)	0.69(1) 0.65(3) 0.65(5)
60	0.67(1) 0.67(3) 0.66(5)	0.70(1) 0.66(3) 0.66(5)	0.32(1) 0.31(3) 0.33(5)	0.67(1) 0.65(3) 0.64(5)
70	<b>0.69(1)</b> 0.67(3) 0.66(5)	<b>0.71(1)</b> 0.65(3) 0.65(5)	0.36(1) 0.33(3) 0.34(5)	0.69(1) 0.63(3) 0.64(5)
80	0.67(1) 0.66(3) 0.66(5)	<b>0.71(1)</b> 0.66(3) 0.66(5)	0.34(1) 0.34(3) 0.34(5)	0.67(1) 0.63(3) 0.65(5)
90	0.67(1) 0.66(3) 0.66(5)	0.69(1) 0.64(3) 0.65(5)	0.40(1) 0.34(3) 0.33(5)	0.68(1) 0.65(3) 0.65(5)
100	0.67(1) 0.67(3) 0.64(5)	0.70(1) 0.66(3) 0.67(5)	0.31(1) 0.34(3) 0.34(5)	0.69(1) 0.65(3) 0.65(5)

Tabla 4.48: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.49

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.49 muestran que usando un tamaño de frame de 50 ms se obtiene el mejor puntaje de recall para la firma MFCC, mientras que RASTA-MFCC 30 ms, Entropy Signature 20 ms y MSES 50 y 70 ms. El texto que aparece entre paréntesis en la Tabla 4.49 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.51(linear) 0.69(rbf) 0.64(poly) 0.16(sigmoid)	0.54(linear) 0.65(rbf) 0.65(poly) 0.15(sigmoid)	0.30(linear) 0.38(rbf) 0.44(poly) 0.13(sigmoid)	0.44(linear) 0.62(rbf) 0.59(poly) 0.21(sigmoid)
20	0.54(linear) 0.69(rbf) 0.67(poly) 0.14(sigmoid)	0.53(linear) 0.67(rbf) 0.69(poly) 0.17(sigmoid)	0.35(linear) 0.40(rbf) <b>0.48(poly)</b> 0.14(sigmoid)	0.53(linear) 0.69(rbf) 0.68(poly) 0.14(sigmoid)
30	0.55(linear) 0.70(rbf) 0.67(poly) 0.15(sigmoid)	0.54(linear) 0.67(rbf) <b>0.70(poly)</b> 0.15(sigmoid)	0.36(linear) 0.38(rbf) 0.39(poly) 0.13(sigmoid)	0.53(linear) <b>0.71(rbf)</b> 0.68(poly) 0.15(sigmoid)
40	0.54(linear) 0.70(rbf) 0.70(poly) 0.15(sigmoid)	0.51(linear) 0.65(rbf) 0.65(poly) 0.19(sigmoid)	0.36(linear) 0.42(rbf) 0.43(poly) 0.15(sigmoid)	0.50(linear) 0.70(rbf) 0.70(poly) 0.14(sigmoid)
50	0.55(linear) 0.68(rbf) <b>0.71(poly)</b> 0.14(sigmoid)	0.50(linear) 0.64(rbf) 0.62(poly) 0.18(sigmoid)	0.38(linear) 0.39(rbf) 0.44(poly) 0.14(sigmoid)	0.51(linear) <b>0.71(rbf)</b> <b>0.71(poly)</b> 0.15(sigmoid)
60	0.52(linear) 0.66(rbf) 0.69(poly) 0.11(sigmoid)	0.48(linear) 0.66(rbf) 0.63(poly) 0.18(sigmoid)	0.32(linear) 0.37(rbf) 0.32(poly) 0.13(sigmoid)	0.51(linear) 0.69(rbf) 0.66(poly) 0.11(sigmoid)
70	0.51(linear) 0.67(rbf) 0.70(poly) 0.11(sigmoid)	0.50(linear) 0.66(rbf) 0.65(poly) 0.18(sigmoid)	0.31(linear) 0.37(rbf) 0.37(poly) 0.14(sigmoid)	0.50(linear) 0.70(rbf) <b>0.71(poly)</b> 0.13(sigmoid)
80	0.53(linear) 0.66(rbf) 0.69(poly) 0.10(sigmoid)	0.49(linear) 0.69(rbf) 0.68(poly) 0.15(sigmoid)	0.34(linear) 0.39(rbf) 0.37(poly) 0.13(sigmoid)	0.49(linear) 0.68(rbf) 0.66(poly) 0.14(sigmoid)
90	0.49(linear) 0.67(rbf) 0.69(poly) 0.10(sigmoid)	0.49(linear) 0.66(rbf) 0.66(poly) 0.16(sigmoid)	0.33(linear) 0.37(rbf) 0.35(poly) 0.13(sigmoid)	0.48(linear) 0.68(rbf) 0.69(poly) 0.13(sigmoid)
100	0.49(linear) 0.66(rbf) 0.69(poly) 0.10(sigmoid)	0.49(linear) 0.69(rbf) 0.67(poly) 0.15(sigmoid)	0.36(linear) 0.39(rbf) 0.31(poly) 0.13(sigmoid)	0.49(linear) 0.69(rbf) 0.69(poly) 0.12(sigmoid)

Tabla 4.49: Resultados de recall usando el clasificador SVM.

## Experimento 4

### Objetivo

De los resultados del experimento 1 y 2, se observó que el mayor puntaje de recall se consiguió utilizando 24 coeficientes por lo que a continuación en este experimento se trabaja con esta configuración. El experimento 4 consiste en hacer la clasificación de las emociones modificando el tamaño de frame, desde 10 ms hasta 100 ms con incrementos de 10 ms usando un traslape del 50% y usando filtro de preéñfasis en el proceso de extracción de características.

### Configuración

Usando la función ventana de Hann, con preéñfasis de 0.97, ancho de banda completo y traslape de 50%.

### Resumen de la Tabla 4.50

Para el clasificador MLP, usando la métrica de recall, los resultados de la Tabla 4.50 muestran que usando un tamaño de frame de 80 ms se obtiene el mejor promedio de recall para la firma MFCC, 20 ms para RASTA-MFCC y Entropy Signature y 50 ms para MSES.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.69	0.72	0.47	0.71
20	0.71	<b>0.77</b>	<b>0.51</b>	0.73
30	0.72	0.74	0.50	0.73
40	0.71	0.73	0.48	0.71
50	0.73	0.71	0.49	<b>0.74</b>
60	0.71	0.72	0.45	0.72
70	0.71	0.70	0.45	0.73
80	<b>0.74</b>	0.72	0.44	0.73
90	0.71	0.70	0.42	0.72
100	0.70	0.73	0.42	0.73

Tabla 4.50: Resultados de recall usando el clasificador MLP.

### Resumen de la Tabla 4.51

Para el clasificador KNN, usando la métrica de recall, los resultados de la Tabla 4.51 muestran que usando un tamaño de frame de 50 ms se obtiene el mejor puntaje de recall para la firma MFCC, con 10 ms para MSES, 60 ms para RASTA-MFCC y 30 ms para Entropy Signature. El número que aparece entre paréntesis en la Tabla 4.51 es el número de vecinos utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.66(1)	0.65(1)	0.44(1)	<b>0.70(1)</b>
	0.67(3)	0.61(3)	0.42(3)	0.64(3)
	0.65(5)	0.61(5)	0.39(5)	0.66(5)
20	0.67(1)	0.67(1)	0.43(1)	0.69(1)
	0.69(3)	0.67(3)	0.37(3)	0.67(3)
	0.65(5)	0.69(5)	0.39(5)	0.63(5)
30	0.69(1)	0.66(1)	<b>0.47(1)</b>	0.67(1)
	0.70(3)	0.67(3)	0.39(3)	0.65(3)
	0.66(5)	0.67(5)	0.38(5)	0.66(5)
40	0.66(1)	0.67(1)	0.46(1)	0.69(1)
	0.69(3)	0.67(3)	0.41(3)	0.64(3)
	0.65(5)	0.68(5)	0.39(5)	0.65(5)
50	0.68(1)	0.66(1)	0.41(1)	0.68(1)
	<b>0.71(3)</b>	0.65(3)	0.35(3)	0.65(3)
	0.66(5)	0.64(5)	0.38(5)	0.67(5)
60	0.68(1)	<b>0.71(1)</b>	0.35(1)	0.68(1)
	0.67(3)	0.65(3)	0.36(3)	0.67(3)
	0.66(5)	0.66(5)	0.36(5)	0.65(5)
70	0.67(1)	0.69(1)	0.35(1)	0.68(1)
	0.66(3)	0.65(3)	0.32(3)	0.66(3)
	0.64(5)	0.65(5)	0.35(5)	0.65(5)
80	0.67(1)	0.70(1)	0.35(1)	0.69(1)
	0.65(3)	0.66(3)	0.33(3)	0.65(3)
	0.65(5)	0.65(5)	0.37(5)	0.64(5)
90	0.67(1)	0.67(1)	0.37(1)	0.67(1)
	0.64(3)	0.65(3)	0.35(3)	0.66(3)
	0.66(5)	0.61(5)	0.35(5)	0.65(5)
100	0.67(1)	0.69(1)	0.35(1)	0.68(1)
	0.66(3)	0.63(3)	0.36(3)	0.65(3)
	0.65(5)	0.63(5)	0.34(5)	0.64(5)

Tabla 4.51: Resultados de recall usando el clasificador KNN.

### Resumen de la Tabla 4.52

Para el clasificador SVM, usando la métrica de recall, los resultados de la Tabla 4.52 muestran que usando un tamaño de frame de 20 y 30 ms se obtiene el mejor puntaje de recall para la firma MFCC, con RASTA-MFCC con un frame en los rangos de 20 a 60 ms y de 80 a 100 ms mientras que para Entropy Signature y MSES 30 ms. El texto que aparece entre paréntesis en la Tabla 4.52 es el kernel utilizado para conseguir el puntaje de recall.

Tamaño de Frame	MFCC	RASTA-MFCC	Entropy Signature	MSES
10	0.53(linear) 0.67(rbf) 0.70(poly) 0.14(sigmoid)	0.53(linear) 0.67(rbf) 0.68(poly) 0.16(sigmoid)	0.32(linear) 0.39(rbf) 0.39(poly) 0.14(sigmoid)	0.43(linear) 0.65(rbf) 0.62(poly) 0.28(sigmoid)
20	0.52(linear) <b>0.71(rbf)</b> 0.67(poly) 0.17(sigmoid)	0.50(linear) 0.67(rbf) <b>0.69(poly)</b> 0.14(sigmoid)	0.37(linear) 0.40(rbf) 0.43(poly) 0.16(sigmoid)	0.51(linear) 0.69(rbf) 0.68(poly) 0.19(sigmoid)
30	0.51(linear) <b>0.71(rbf)</b> 0.69(poly) 0.14(sigmoid)	0.53(linear) 0.67(rbf) <b>0.69(poly)</b> 0.16(sigmoid)	0.39(linear) <b>0.44(rbf)</b> 0.42(poly) 0.14(sigmoid)	0.51(linear) <b>0.71(rbf)</b> 0.69(poly) 0.19(sigmoid)
40	0.54(linear) 0.66(rbf) 0.69(poly) 0.15(sigmoid)	0.50(linear) <b>0.69(rbf)</b> <b>0.69(poly)</b> 0.19(sigmoid)	0.38(linear) 0.43(rbf) 0.40(poly) 0.14(sigmoid)	0.50(linear) 0.69(rbf) 0.70(poly) 0.18(sigmoid)
50	0.55(linear) 0.66(rbf) 0.68(poly) 0.15(sigmoid)	0.51(linear) 0.67(rbf) <b>0.69(poly)</b> 0.17(sigmoid)	0.38(linear) 0.44(rbf) 0.41(poly) 0.13(sigmoid)	0.50(linear) 0.68(rbf) 0.69(poly) 0.17(sigmoid)
60	0.53(linear) 0.65(rbf) 0.69(poly) 0.14(sigmoid)	0.53(linear) <b>0.69(rbf)</b> <b>0.69(poly)</b> 0.14(sigmoid)	0.33(linear) 0.42(rbf) 0.33(poly) 0.15(sigmoid)	0.50(linear) 0.68(rbf) 0.69(poly) 0.17(sigmoid)
70	0.51(linear) 0.63(rbf) 0.70(poly) 0.12(sigmoid)	0.53(linear) 0.65(rbf) 0.67(poly) 0.17(sigmoid)	0.34(linear) 0.38(rbf) 0.34(poly) 0.14(sigmoid)	0.49(linear) 0.66(rbf) 0.71(poly) 0.19(sigmoid)
80	0.53(linear) 0.64(rbf) 0.69(poly) 0.10(sigmoid)	0.51(linear) <b>0.69(rbf)</b> 0.68(poly) 0.18(sigmoid)	0.31(linear) 0.38(rbf) 0.36(poly) 0.13(sigmoid)	0.46(linear) 0.66(rbf) 0.69(poly) 0.15(sigmoid)
90	0.51(linear) 0.65(rbf) 0.68(poly) 0.12(sigmoid)	0.53(linear) 0.66(rbf) <b>0.69(poly)</b> 0.18(sigmoid)	0.31(linear) 0.35(rbf) 0.33(poly) 0.14(sigmoid)	0.49(linear) 0.67(rbf) 0.67(poly) 0.16(sigmoid)
100	0.51(linear) 0.66(rbf) 0.68(poly) 0.10(sigmoid)	0.51(linear) 0.67(rbf) <b>0.69(poly)</b> 0.19(sigmoid)	0.31(linear) 0.37(rbf) 0.32(poly) 0.10(sigmoid)	0.46(linear) 0.67(rbf) 0.67(poly) 0.14(sigmoid)

Tabla 4.52: Resultados de recall usando el clasificador SVM.

## 4.6 Mejores resultados de los experimentos sin autocorrelación

A continuación, se presenta una tabla de cada una de las firmas espectrales que muestra los mejores puntajes recall obtenidos en los experimentos para clasificar las dos bases de datos.

Se puede observar que para MFCC el clasificador que más destaca para la base de datos EMODB en los experimentos 1 y 2 es MLP, en el experimento 3 los tres clasificadores dan un puntaje de 0.78 y en el experimento 4 es SVM. Para EMOVO en todos los experimentos es MLP.

El mejor resultado para EMODB es de 0.78, se obtiene en los experimentos 1, 3 y 4, con 48 coeficientes, y 10, 20 y 90 ms, para EMOVO es de 0.75, se obtiene con MLP, en el experimento 2, con 48 coeficientes.

MFCC				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	48 coeficientes	<b>0.78</b>	36 coeficientes	<b>0.73</b>
KNN	24 coeficientes	0.73(5)	24 coeficientes	0.69(3)
SVM	24 coeficientes	0.75(rbf)	24 coeficientes	0.70(rbf)
Experimento 2				
MLP	24 coeficientes	<b>0.77</b>	48 coeficientes	<b>0.75</b>
KNN	24 coeficientes	0.73(5)	24 coeficientes	0.70(3)
SVM	24 coeficientes	0.74(rbf)	24 coeficientes	0.71(rbf)
Experimento 3				
MLP	Frame de 90 ms	0.78	Frame de 20 ms	<b>0.73</b>
KNN	Frame de 10 ms	0.78(1)	Frame de 10, 30, 40, 50 y 70 ms	0.69(1, 3)
SVM	Frame de 20 ms	0.78(rbf)	Frame de 50 ms	0.71(poly)
Experimento 4				
MLP	Frame de 80 ms	0.77	Frame de 80 ms	<b>0.74</b>
KNN	Frame de 10 ms	0.75(1)	Frame de 50 ms	0.71(3)
SVM	Frame de 10 ms	<b>0.78(rbf)</b>	Frame de 20 y 30 ms	0.71(rbf)

Tabla 4.53: Mejores resultados para la firma MFCC.

Se puede observar que para RASTA-MFCC el clasificador que más destaca para la base de datos EMODB en los experimentos 1 y 2 es MLP, en el experimento 3 es SVM y en el 4 son SVM y MLP ambos con un resultado de 0.78, mientras que para EMOVO en todos los experimentos es MLP.

El mejor resultado para EMODB es de 0.80, se obtiene con SVM, en el experimento 3, con 10 ms, y para EMOVO es de 0.76, se obtiene con MLP, en el experimento 4, con 20 ms.

RASTA-MFCC				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	48 coeficientes	<b>0.77</b>	24 coeficientes	<b>0.72</b>
KNN	24 coeficientes	0.71(3, 5)	24 coeficientes	0.67(1, 3)
SVM	24 coeficientes	0.76(rbf)	24 coeficientes	0.70(poly)
Experimento 2				
MLP	24 coeficientes	<b>0.73</b>	24 coeficientes	<b>0.74</b>
KNN	24 coeficientes	0.72(5)	24 coeficientes	0.67(3, 5)
SVM	24 coeficientes	<b>0.73(rbf)</b>	24 coeficientes	0.69(poly)
Experimento 3				
MLP	Frame de 70 ms	0.79	Frame de 40 ms	<b>0.74</b>
KNN	Frame de 10 ms	0.76(3)	Frame de 70 y 80 ms	0.71(1)
SVM	Frame de 10 ms	<b>0.80(rbf)</b>	Frame de 30 ms	0.70(poly)
Experimento 4				
MLP	Frame de 20 ms	<b>0.78</b>	Frame de 20 ms	<b>0.77</b>
KNN	Frame de 10 ms	0.76(5)	Frame de 60 ms	0.71(1)
SVM	Frame de 10 ms	<b>0.78(rbf)</b>	Frame de 20-60 y 80-100 ms	0.69(rbf, poly)

Tabla 4.54: Mejores resultados para la firma RASTA-MFCC.

Se puede observar que para Entropy Signature el clasificador que más destaca para la base de datos EMODB en los experimentos 1, 2 y 4 es SVM, y en el 3 MLP, mientras que para EMOVO en todos los experimentos es MLP.

El mejor resultado para EMODB es de 0.53, se obtiene con MLP, en el experimento 3, con 50 ms, y para EMOVO es de 0.52, se obtiene con MLP, en los experimento 2 y 3, con 24 coeficientes y 50 ms.

Entropy Signature				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	24 coeficientes	0.48	24 coeficientes	<b>0.50</b>
KNN	48 coeficientes	0.47(1)	24 coeficientes	0.42(1)
SVM	24 coeficientes	<b>0.52(poly)</b>	36 coeficientes	0.49(poly)
Experimento 2				
MLP	12 coeficientes	0.50	24 coeficientes	<b>0.52</b>
KNN	48 coeficientes	0.50(5)	24 coeficientes	0.50(1)
SVM	48 coeficientes	<b>0.52(poly)</b>	48 coeficientes	0.46(poly)
Experimento 3				
MLP	Frame de 50 ms	<b>0.53</b>	Frame de 50 ms	<b>0.52</b>
KNN	Frame de 50 y 80 ms	0.48(1, 3, 5)	Frame de 20 ms	0.45(1)
SVM	Frame de 30 y 50 ms	0.52(rbf, poly)	Frame de 20 ms	0.48(poly)
Experimento 4				
MLP	Frame de 50 y 60 ms	0.51	Frame de 20 ms	<b>0.51</b>
KNN	Frame de 60 y 70 ms	0.50(1, 5)	Frame de 30 ms	0.47(1)
SVM	Frame de 50 ms	<b>0.55(poly)</b>	Frame de 30 ms	0.44(rbf)

Tabla 4.55: Mejores resultados para la firma Entropy Signature.

Se puede observar que para MSES el clasificador que más destaca para la base de datos EMODB es SVM, mientras que para EMOVO es MLP.

El mejor resultado para EMODB es de 0.80, se obtiene con SVM, en el experimento 4, con 20 ms, y para EMOVO es de 0.74, se obtiene con MLP, en los experimentos 3 y 4, con 50 ms.

MSES				
	EMODB		EMOVO	
Clasificador	Parámetros	Recall	Parámetros	Recall
Experimento 1				
MLP	24 coeficientes	0.77	36 coeficientes	<b>0.73</b>
KNN	24 coeficientes	0.77(5)	24 coeficientes	0.67(1)
SVM	24 coeficientes	<b>0.78(rbf)</b>	24 coeficientes	0.69(rbf)
Experimento 2				
MLP	24 coeficientes	0.76	24 coeficientes	<b>0.73</b>
KNN	24 coeficientes	0.75(5)	24 coeficientes	0.65(1)
SVM	24 coeficientes	<b>0.77(rbf)</b>	24 coeficientes	0.70(rbf)
Experimento 3				
MLP	Frame de 30 ms	0.78	Frame de 50 ms	<b>0.74</b>
KNN	Frame de 30 ms	0.77(5)	Frame de 20 y 30 ms	0.70(1)
SVM	Frame de 20 ms	<b>0.79(rbf)</b>	Frame de 50 y 70 ms	0.71(rbf, poly)
Experimento 4				
MLP	Frame de 20 ms	0.78	Frame de 50 ms	<b>0.74</b>
KNN	Frame de 20 y 30 ms	0.75(5)	Frame de 10 ms	0.70(1)
SVM	Frame de 20 ms	<b>0.80(rbf)</b>	Frame de 30 ms	0.71(rbf)

Tabla 4.56: Mejores resultados para la firma MSES.

## 4.7 Mejor resultado EMODB

El mejor puntaje recall para EMODB de 0.82, se obtuvo usando autocorrelación, en el experimento 3, sin filtro de preénfasis, con un tamaño de frame de 10 ms, con la característica RASTA-MFCC y el clasificador MLP.

A continuación, se muestran los diez puntajes recall que fueron promediados para la obtención del mejor resultado para la base de datos EMODB, se resalta en negritas el mejor de los diez.

Diez puntajes recall: [0.83229816 0.82608694 0.81987578 **0.83850932** 0.80124223 0.78881985  
0.81366462 **0.83850932** 0.80124223 0.81987578]

Promedio de los Diez puntajes recall: 0.8180124223232269

A continuación, se muestra la matriz de confusión para el mejor de los diez puntajes recall, se puede ver que la clase mejor clasificada es Ira y las que tienen más error son Neutral y Alegría.

	Angustia	Disgusto	Alegría	Aburrimiento	Neutral	Tristeza	Ira
Angustia	19	0	1	0	0	2	0
Disgusto	0	10	0	0	0	0	3
Alegría	1	0	15	1	0	0	4
Aburrimiento	0	1	0	19	3	1	0
Neutral	0	2	1	4	17	0	0
Tristeza	1	0	0	0	0	18	0
Ira	0	0	1	0	0	0	37

Tabla 4.57: Matriz de confusión del mejor resultado de la base de datos EMODB.

## 4.8 Mejor resultado EMOVO

El mejor puntaje recall para EMOVO de 0.77, se obtuvo sin usar autocorrelación, en el experimento 4, con filtro de preénfasis, con un tamaño de frame de 20 ms, con la característica RASTA-MFCC y el clasificador MLP

A continuación, se muestran los diez puntajes recall que fueron promediados para la obtención del mejor resultado para la base de datos EMOVO, se resalta en negritas el mejor de los diez.

Diez puntajes recall: [0.76571429 0.74857146 0.75428569 0.77714288 0.74857146 0.78285712  
0.74857146 0.75999999 0.78285712 **0.79428571**]

Promedio de los Diez puntajes recall: 0.7662857174873352

A continuación, se muestra la matriz de confusión para el mejor de los diez puntajes recall, se puede ver que la clase mejor clasificada es disgusto y la que tiene más error es sorpresa.

	Angustia	Disgusto	Alegría	Sorpresa	Neutral	Tristeza	Ira
Angustia	21	0	0	3	0	1	0
Disgusto	0	22	0	1	1	1	0
Alegría	0	0	18	4	3	0	0
Sorpresa	2	4	0	16	0	2	1
Neutral	1	2	0	1	21	0	0
Tristeza	2	1	0	0	1	21	0
Ira	1	0	2	1	1	0	20

Tabla 4.58: Matriz de confusión del mejor resultado de la base de datos EMOVO.

## 4.9 Comparación con los mejores resultados en el estado del arte

A continuación, en la Tabla 4.59, se muestran los mejores resultados en el estado del arte para la base de datos EMODB, se observa que el mejor resultado para este conjunto de datos lo tiene el trabajo [16] con un puntaje recall de 84.62%, un 2.82% superior al puntaje obtenido en esta tesis.

Paper	Dataset	Features	Classifier	Results
Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo [1].	EMODB	Características acústicas.	SVM.	Se obtuvo un puntaje recall de 84.13%
Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace [14].	EMODB correlacionada con SAVEE	Se extrajeron 384 características usando openSMILE.	Logistico simple con Kernel Canonical Correlation Analysis (KCCA)	Se consiguió un puntaje recall de 71.9%
Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique [13].	EMODB	eGemaps feature set.	Deep Belief Networks (DBNs)	El puntaje recall obtenido es de 72.38%
A novel feature selection method for speech emotion recognition [16].	EMODB	Se extrajeron 1582 características usando OpenSMILE.	SVM	El puntaje recall obtenido es de 84.62%
Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods [9].	EMODB	Emobase con Selección de Infinite Latent Feature Selection (ILFS).	SVM	El puntaje recall obtenido es de 76.9%
Wavelet packet analysis for speaker-independent emotion recognition [17].	EMODB	Wavelet Packet Coefficient (WPC) con Sequential Floating Forward Search (SFFS)	RSVM	El puntaje recall obtenido es de 79.2%
Esta tesis.	EMODB	RASTA-MFCC	MLP	El puntaje recall obtenido es de 81.8%

Tabla 4.59: Mejores resultados en el estado del arte para la base de datos EMODB.

A continuación, en la Tabla 4.60, se muestran los mejores resultados en el estado del arte para la base de datos EMOVO, se observa que el mejor resultado para este conjunto de datos es de 76.62% el cual se logra en esta tesis.

Paper	Dataset	Features	Classifier	Results
Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace [14].	EMOVO correlacionadas con SAVEE	Se extrajeron 384 características usando openSMILE.	Logistico simple con Kernel Canonical Correlation Analysis (KCCA)	Se consiguió un puntaje recall de 59.5%
Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique [13].	EMOVO	eGemaps feature set.	Deep Belief Networks(DBNs)	El puntaje recall obtenido es de 76.22%
A novel feature selection method for speech emotion recognition [16].	EMOVO	Se extrajeron 1582 características usando OpenSMILE.	SVM	El puntaje recall obtenido es de 60.4%
Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods [9].	EMOVO	eGeMAPS con puntuación generalizada de Fisher.	SVM	El puntaje recall obtenido es de 41.0%
Esta tesis.	EMOVO	RASTA-MFCC	MLP	El mejor puntaje recall es de 76.62%

Tabla 4.60: Mejores resultados en el estado del arte para la base de datos EMOVO.

## CAPITULO 5 CONCLUSIÓN

### 5.1 Conclusión

Este trabajo es bueno comparado con otros en el estado del arte, ya que, como se muestra en las tablas 4.59 y 4.60 se obtienen buenos resultados, en especial para EMOVO, donde se obtuvo el mejor resultado en el estado del arte, además se utilizaron características novedosas como RASTA-MFCC, MSES y Entropy Signatura las cuales no han sido exploradas en otros trabajos y se aportó un enfoque discreto de clasificación de 7 clases de emociones usando autocorrelación.

La característica que dio los mejores resultados en ambas bases de datos fue RASTA-MFCC, sin embargo, MFCC y MSES dieron buenos resultados siendo solo superadas por 0.01 y un 0.02 respectivamente. Entropy Signature dio muy mal rendimiento para el reconocimiento de emociones en el habla, siendo su puntaje más alto de 0.55. Cada característica tiene su utilidad combinarlas como en el caso de RASTA-MFCC puede resultar útil para futuros trabajos.

El clasificador que dio los mejores resultados en ambas bases de datos fue MLP, sin embargo, SVM con los kernel rbf y poly se desempeñó mejor que MLP en los experimentos sin autocorrelación para EMODB y en algunos casos al momento de clasificar la característica Entropy Signature. El clasificador KNN aunque se desempeñó bien no tuvo ningún resultado destacable. Se vio que en los trabajos [1], [16] SVM dio mejores resultados que los obtenidos en este trabajo para EMODB, esto motiva a probar otras configuraciones de SVM en futuros trabajos, así como también, probar otras configuraciones de MLP y KNN.

El mejor puntaje recall para EMODB de 0.82, se obtuvo usando autocorrelación, en el experimento 3, sin filtro de preénfasis, con un tamaño de frame de 10 ms, con la característica RASTA-MFCC y el clasificador MLP. Y el mejor puntaje recall para EMOVO de 0.77, se obtuvo sin usar autocorrelación, en el experimento 4, con filtro de preénfasis, con un tamaño de frame de 20 ms, con la característica RASTA-MFCC y el clasificador MLP. Esto valida la hipótesis de esta tesis solo para el caso de EMODB, esta se puede deber a que la autocorrelación resulta útil en la clasificación dependiendo el idioma, puede ser que en alemán funcionó bien porque los sonidos vocalizados son más expresivos en alemán y en italiano pueden llegar a ser más expresivos los sonidos no vocalizados y los silencios por eso al eliminarlos con la autocorrelación se le quita expresividad a la emoción lo que lleva a un bajo nivel de precisión en el clasificador. También se puede deber a la diferencia de calidad entre las bases de datos. La calidad del conjunto de datos de EMODB fue evaluada por 20 codificadores humanos con una tasa de reconocimiento promedio del 86% y las grabaciones de audio con un acuerdo entre codificadores por debajo del 80% fueron eliminadas (no se tomó tal medida para EMOVO). Para EMOVO, la precisión informada por el equipo de prueba es del 80% [9], se debe tener en cuenta que en lugar de evaluar el conjunto de datos EMOVO completo solo se seleccionaron dos frases y cada codificador tuvo que elegir solo entre dos emociones propuestas en lugar de siete. El hecho de que el enfoque de aprendizaje automático para la clasificación de EMOVO, en esta tesis, sea de un problema de siete clases explica los resultados mucho más bajos obtenidos en comparación con el rendimiento humano. Hay que tener en cuenta que el proceso para reconocimiento de voz tiene que estar centrado en la aplicación que se le quiere dar, ya que, incluir emociones como ira y disgusto que meten ruido a la clasificación puede no ser necesario en una aplicación que busca identificar emociones para captar el interés de un niño en algún videojuego, en dado caso incluir solo 3 emociones garantizaría una precisión por arriba del 90% la cual sería más que suficiente para una aplicación. También hay que tener en cuenta que el número de características a extraer de la señal de voz influye en la cantidad de

procesamiento que requerirá la aplicación, una aplicación para un dispositivo con bajos recursos necesitaría extraer solo características de bajo procesamiento para ser funcional.

El área de reconocimiento de voz necesita más estudios en idiomas como el español latinoamericano que no han sido explorados en vista de que hacer una base de datos para dicho idioma resulta costoso y laborioso, en futuros trabajos sería bueno realizar una base de datos mexicana para expandir la aplicación del reconocimiento de emociones de la voz en este país. Así como también sería bueno explorar nuevas bases de datos en más idiomas, nuevas características y más clasificadores para abrir nuevas áreas de aplicación para el reconocimiento automático de las emociones de la voz ya que al ser capaz de identificar más emociones y matices entre ellas (como propone el modelo dimensional) se podrían abrir nuevas aplicaciones para dicha tecnología garantizando un alto nivel de precisión.

## REFERENCIAS

- [1] H. P. Espinosa y C. A. R. García, «Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo», p. 42, 2010.
- [2] M. B. Akçay y K. Oğuz, «Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers», *Speech Commun.*, vol. 116, pp. 56-76, ene. 2020, doi: 10.1016/j.specom.2019.12.001.
- [3] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, y Lian Li, «Speech Emotion Recognition Using Fourier Parameters», *IEEE Trans. Affect. Comput.*, vol. 6, n.º 1, pp. 69-75, ene. 2015, doi: 10.1109/TAFFC.2015.2392101.
- [4] L. Devillers y L. Vidrascu, «Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs», p. 4, 2006.
- [5] Y. Hernández, L. E. Sucar, y C. Conati, «An Affective Behavior Model for Intelligent Tutors Intelligent Tutoring Systems (ITS) LNCS», vol. 5091, pp. 819-821, 2008.
- [6] «Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in chicanos/latinos.», University of Michigan, 1999.
- [7] Department of Electronics and Communication Engineering, S. G. Balekundri Institute of Technology, Belagavi-India, V. Kanabur, S. S. Harakannanavar, y D. Torse, «An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition», *Int. J. Image Graph. Signal Process.*, vol. 11, n.º 5, pp. 1-12, may 2019, doi: 10.5815/ijigsp.2019.05.01.
- [8] S. Selva Nidhyanthan, R. Shantha Selva Kumari, y T. Senthur Selvi, «Noise Robust Speaker Identification Using RASTA-MFCC Feature with Quadrilateral Filter Bank Structure», *Wirel. Pers. Commun.*, vol. 91, n.º 3, pp. 1321-1333, dic. 2016, doi: 10.1007/s11277-016-3530-3.
- [9] F. Haider, S. Pollak, P. Albert, y S. Luz, «Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods», *Comput. Speech Lang.*, vol. 65, p. 101119, ene. 2020, doi: 10.1016/j.csl.2020.101119.
- [10] S. Ying y Z. Xue-Ying, «Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition», *Future Gener. Comput. Syst.*, vol. 81, pp. 291-296, abr. 2018, doi: 10.1016/j.future.2017.10.002.
- [11] N. Semwal, A. Kumar, y S. Narayanan, «Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models», en *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, New Delhi, India, feb. 2017, pp. 1-6, doi: 10.1109/ISBA.2017.7947681.
- [12] T. Özseven, «Investigation of the Relation between Emotional State and Acoustic Parameters in the Context of Language», *Eur. J. Sci. Technol.*, pp. 241-244, dic. 2018, doi: 10.31590/ejosat.448095.

- [13] S. Latif, R. Rana, S. Younis, J. Qadir, y J. Epps, «Cross Corpus Speech Emotion Classification - An Effective Transfer Learning Technique», *ArXiv180106353 Cs*, mar. 2018, Accedido: jul. 02, 2020. [En línea]. Disponible en: <http://arxiv.org/abs/1801.06353>.
- [14] H. Sagha, J. Deng, M. Gavryukova, J. Han, y B. Schuller, «Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace», en *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, mar. 2016, pp. 5800-5804, doi: 10.1109/ICASSP.2016.7472789.
- [15] O. Scharenborg, S. Kakouros, y J. Koemans, «The Effect of Noise on Emotion Perception in an Unknown Language», en *9th International Conference on Speech Prosody 2018*, jun. 2018, pp. 364-368, doi: 10.21437/SpeechProsody.2018-74.
- [16] T. Özseven, «A novel feature selection method for speech emotion recognition», *Appl. Acoust.*, vol. 146, pp. 320-326, mar. 2019, doi: 10.1016/j.apacoust.2018.11.028.
- [17] K. Wang, G. Su, L. Liu, y S. Wang, «Wavelet packet analysis for speaker-independent emotion recognition», *Neurocomputing*, vol. 398, pp. 257-264, jul. 2020, doi: 10.1016/j.neucom.2020.02.085.
- [18] A. Vijayalakshmi, J. Midhun, y N. Moksha, «A study on Automatic Speech Recognition Techniques», *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, n.º 3, pp. 614-617, 2015.
- [19] L. Deng y L. Xiao, «Machine learning paradigms for speech recognition», vol. 21, n.º 5, pp. 1060-1089, 2013.
- [20] P. S. Pukhraj, R. D. Ratnadeep, y M. W. Vishal, «Indian Language Speech Database A Review», *Intern. J. Comput. Appl.*, vol. 47, n.º 5, pp. 17-21, 2012.
- [21] M. S. Swathy y K. R. Mahesh, «Review on Feature Extraction and Classification Techniques in Speaker Recognition», *Int. J. Eng. Res. Gen. Sci.*, vol. 5, n.º 2, pp. 78-83, 2017.
- [22] I. Nobuyasu, N. Masafumi, K. Gakuto, y T. Ryuki, «A Metric for Evaluating Speech Recognition Accuracy based on Human Perception», *Int. J. Inf. Process. Soc. Jpn.*, vol. 104, n.º 11, pp. 1-7, 2014.
- [23] R. Plutchik, «The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice.», 2001.
- [24] P. Ekman y H. Oster, «Facial expressions of emotion.», *Psychol.*, pp. 527-554, 1979.
- [25] P. Ekman, W. V. Friesen, y P. Ellsworth, «Emotion in the Human Face: Guidelines for Research and an Integration of Findings.», 2013.
- [26] P. Ekman, «Universals and cultural differences in facial expressions of emotion.. Nebraska symposium on motivation.», *University of Nebraska Press.*, 1971.
- [27] D. Watson, L. A. Clark, y A. Tellegen, «Development and validation of brief measures of positive and negative affect: the panas scales.», *J. Personal. Soc. Psychol.*, p. 1063, 1988.

- [28] J. A. Russell y A. Mehrabian, «Evidence for a three-factor theory of emotions.», *J. Res. Personal.*, pp. 273-294, 1977.
- [29] M. A. Nicolaou, H. Gunes, y M. Pantic, «Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space.», *IEEE Trans.*, pp. 92-105, 2011.
- [30] «Primitives-based evaluation and estimation of emotions in speech.», pp. 787–800, 2007.
- [31] Z. Zeng, M. Pantic, G. I. Roisman, y T. S. Huang, «A survey of affect recognition methods: audio, visual, and spontaneous expressions.», *IEEE Trans.*, pp. 39-58, 2009.
- [32] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, y B. Weiss, «A Database of German Emotional Speech», p. 4, 2005.
- [33] G. Costantini, I. Iadarola, A. Paoloni, y M. Todisco, «EMOVO Corpus: an Italian Emotional Speech Database», p. 4.
- [34] B. Logan, «Mel frequency cepstral coefficients for music modeling.», *IEEE Proc. Int. Symp. Music Inf. Retr.*, 2000.
- [35] «Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music.», 7th International Conference on Music Information Retrieval ., 2006.
- [36] C. E. Shannon, «A mathematical theory of communication.», *Bell Syst. Tech. J.* 27, pp. 379-423, 1948.
- [37] S. Jia-Lin, H. Jeih-Weih, y L. Lin-Shan, «Robust entropy-based endpoint detection for speech recognition in noisy environments», *5th Int. Conf. Spok. Lang. Process.* 98, pp. 232-235, 1998.
- [38] H. Misra, S. Ikbal, H. Bourland, y H. Hermansky, «Spectral entropy based feature for robust asr», *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 193-196, 2004.
- [39] H. Misra, S. Ikbal, S. Sivadas, y H. Bourland, «Multi-resolution spectral entropy feature for robust asr.», *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 253-256, 2005.
- [40] A. Camarena-Ibarrola, F. Luque, y E. Chavez, «Speaker identification through spectral entropy analysis.», *IEEE Int. Autumn Meet. Power Electron. Comput. ROPEC*, 2017.
- [41] A. Camarena-Ibarrola y E. Chavez, «On musical performances identification, entropy and string matching», *Mex. Int. Conf. Artif. Intell.*, pp. 952-962, 2006.
- [42] A. Camarena-Ibarrola, E. Chavez, y E. Sadit-Tellez, «Robust radio broadcast monitoring using a multi-band spectral entropy signature», *Iberoam. Congr. Pattern Recognit.*, pp. 587-594, 2009.
- [43] A. Camarena-Ibarrola, K. Figueiroa, y H. Tejeda-Villela, «Entropy per chroma for cover song identification.», *IEEE Int. Autumn Meet. Power Electron. Comput. ROPEC*, 2016.
- [44] A. Manzo-Martinez y A. Camarena-Ibarrola, «Use of the entropy of a random process in audio matching tasks.», *38th Int. Conf. Telecommun. Signal Process. TSP*, 2015.

- [45] A. Mohammad, «Entropy in signal processing. Traitement du Signal», pp. 87-116, 1994.
- [46] J. O. Smith y J. S. Abel, «Bark and erb bilinear transforms.», *IEEE Trans. Speech Audio Process.*, pp. 697-708, 1999.
- [47] H. Traunmueller, «Analytical expressions for the tonotopic sensory scale.», *J. Acoust. Soc. Am.* 88, pp. 97-100, 1990.
- [48] A. Manzo-Martínez, F. Gaxiola, G. Ramírez-Alonso, R. Cornejo, y L. C. González-Gurrola, «Kitchen Acoustic Event Identification based on the Entropy of a Random Process», p. 25, 2020.
- [49] «Análisis de codificaciones perceptuales de la voz y su comparación en el reconocimiento de comandos», Universidad Nacional Autónoma de México, México, DF, 2012.
- [50] H. Hermansky y N. Morgan, «RASTA processing of speech.», *IEEE Trans. Speech Audio Process*, pp. 578-589, 1994.
- [51] National Engineering School of Tunis (ENIT) Laboratory of Systems and Signal Processing (LSTS) BP 37, Le Belvédère, 1002 Tunis, Tunisie, H. Rahali, Z. Hajaiej, y N. Ellouze, «Robust Features for Speech Recognition using Temporal Filtering Technique in the Presence of Impulsive Noise», *Int. J. Image Graph. Signal Process.*, vol. 6, n.º 11, pp. 17-24, oct. 2014, doi: 10.5815/ijigsp.2014.11.03.
- [52] N. D. Gaubitch, M. Brookes, y P. A. Naylor, «Blind channel magnitude response estimation in speech using spectrum classification.», *IEEE Trans. Audio Speech Lang. Process.*, p. 2013.
- [53] Y. Jie y W. Zhenli, «Noise robust speech recognition by combining speech enhancement in the wavelet domain and Lin-log RASTA», en *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, Sanya, China, ago. 2009, pp. 415-418, doi: 10.1109/CCCM.2009.5267457.
- [54] N. Morgan y H. Hermansky, «RASTA extensions, Robustness to additive and convolutional noise», *Workshop Speech Process. Adverse Environ.*, pp. 115-118, 1992.
- [55] A. Rodríguez Miranda, «Modelización y análisis de la calidad del aire en la ciudad de oviedo (norte de España), mediante los enfoques pso-svm, red neuronal mlp y árbol de regresión m5», Universidad de león, Leon, 2018.
- [56] J. I. Alonso, J. Gomez, I. García, y J. Martínez, «Autolocalización inicial para robots móviles usando el método de K-NN.», 2007.
- [57] F. Moreno, «Clasificadores eficaces basados en algoritmos rápidos de búsqueda del vecino más cercano.», 2004.
- [58] I. Witten, E. Frank, y M. Hall, «Data Mining Practical Machine Learning Tools and Techniques», 2011.

- [59] G. Morales, J. Mora, y H. Vargas, «Estrategia de regresión basada en el método de los k vecinos más cercanos para la estimación de distancia de falla en sistemas radiales.», pp. 100-108, 2008.
- [60] J. E. Rodríguez, H. A. Barrera, y S. P. Bautista, «Software para el filtrado de paginas web pornograficas basado en clasificador KNN - UDWEBPORN.», 2010.
- [61] J. Mora, G. Morales, y R. Barrera, «Evaluación del clasificador basado en los k vecinos más cercanos para la localización de la zona en falla en los sistemas de potencia.», 2008.
- [62] J. R. De la O, «Interfaz Cerebro-Computadora para el control de un cursor Basado en Ondas Cerebrales.», 2007.
- [63] J. Tudela, «Agrupamiento via clasificacion», Universidad Autonoma del Estado de Mexico, Tianguistenco, Estado de Mexico, 2016.
- [64] L. Primitivo, «Analisis de la complejidad de los datos y su efecto en las redes neuro artificiales.», 2011.
- [65] E. Hernández, «Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto.», 2006.
- [66] C. E. Bedregal, «Agrupamiento de Datos utilizando técnicas MAM-SOM.», 2008.
- [67] V. Vapnik, «Statistical Learning Theory», 1998.
- [68] V. Vapnik, «Estimation of dependences based on empirical data», 1982.
- [69] V. Vapnik, «The nature of statistical learning theory.», 1995.
- [70] R. W. Haas, V. F. Maturana, J. M. Pino, y P. A. Rey, «Utilización de support vector machines no lineal y selección de atributos para credit scoring tesis para optar al grado de magíster en gestión de operaciones memoria para optar al título de ingeniero civil industrial sebastián maldonado alarcón», p. 118.
- [71] N. Cristianini y R. Holloway, «Support Vector and Kernel Methods.», 2003.
- [72] B. Schölkopf, K. Sung, C. Burgues, J. C. Girosi, P. Niyogi, y V. Vapnik, «Comparing Support Vector Machine with Gaussian Kernels to Radial Basis Function classifiers.», *IEEE*, pp. 2758–2765.
- [73] D. Pyle, «Data preparation for data mining.», *Morgan Kaufmann Publishers*, 1999.
- [74] C. Cortes y V. Vapnik, «Support vector networks.», pp. 273–297, 1995.
- [75] C. Burgues, «A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery», pp. 121-167, 1998.
- [76] J. Elgueta, «Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español», Universidad del Bío-Bío Facultad de Ciencias Empresariales, Concepción, Chile, 2017.
- [77] M. Smales, «Classifying Urban sounds using Deep Learning», p. 34, 2018.

- [78] L. Sun, B. Zou, S. Fu, J. Chen, y F. Wang, «Speech emotion recognition based on DNN-decision tree SVM model», *Speech Commun.*, vol. 115, pp. 29-37, dic. 2019, doi: 10.1016/j.specom.2019.10.004.
- [79] B. W. Schuller y A. M. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Chichester, UK: John Wiley & Sons Ltd, 2013.
- [80] A. L.-C. Wang, «An Industrial-Strength Audio Search Algorithm», p. 7.
- [81] «Support vector regression: propiedades y aplicaciones», Universida de Sevilla.