UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE DETECCION DE AGRESIONES VERBALES, POR MEDIO DE ALGORITMOS DE MACHINE LEARNING

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL INDUSTRIAL

VICTOR GABRIEL BUGUEÑO SAEZ

PROFESOR GUIA:

SR. ÁNGEL JIMÉNEZ MOLINA

MIEMBROS DE LA COMISIÓN:

SR. IGNACIO CALISTO LEIVA

SR. ALBERTO CABEZAS BULLEMORE

Este trabajo ha sido financiado por el proyecto Fondecyt 11130252

SANTIAGO DE CHILE

2017

Resumen ejecutivo

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR: VICTOR BUGUEÑO SAEZ
FECHA: 28/08/2017
PROF. GUÍA: SR. ÁNGEL JIMÉNEZ

El presente trabajo tiene como objetivo detectar evidencias de agresiones verbales en archivos de audio. Al respecto se identifican dos clases: Conversaciones normales y Agresión verbal. Para lograr el objetivo, se aplican 4 métodos. El modelo utilizado por Vincenzo Carletti et.al [7], en donde se propone un enfoque para la detección de eventos de audio basados en el paradigma de "bolsa de palabras" (del inglés, Bag of Words), una variación de este modelo utilizando features de la herramienta openSMILE, Support vector machine alimentado por ambas clases y finalmente regresión lineal alimentada por ambas clases.

Al no tener bases de datos abiertas con las clases que se desean analizar, se procedió a construir una base de datos propia recolectando archivos de audio de películas y de internet. De esta manera se obtuvo 809 archivos de audio de 3 segundos de ambas clases de interés y 145 archivos de audio de pocos milisegundos para alimentar la primera etapa del modelo de Vincenzo Carletti.

Se utilizó el software Audacity para transformar los archivos de audio y para extraer el audio de registros audiovisuales de películas. Se utilizó el software Rstudio para el procesamiento general.

En este trabajo se utilizaron 11 features para realizar la clasificación en el modelo original de Vincenzo Carletti. El modelo de Vincenzo, se repitió usando 148 features provenientes de la herramienta openSMILE. Estas mismas features fueron utilizadas para ejecutar los algoritmos de Support vector machine y Regresión lineal.

Los resultados obtenidos fueron de 86.27% de exactitud para el modelo con features originales de Vincenzo Carletti, 79.32% de exactitud para el modelo de Vincenzo utilizando 148 features de openSMILE, 98.19% para algoritmo Support vector machine y 97.74% de exactitud para el algoritmo de regresión lineal.

Los resultados arrojaron resultados prometedores, respecto a la identificación de agresiones verbales, lo que puede permitir el desarrollo de aplicaciones que las puedan identificar monitoreando en tiempo real y que permitan detectar alguna situación de peligro de una persona en condición de vulnerabilidad.

A mi familia, los quiero mucho

Agradecimientos

Quiero agradecer a mi familia, a mis hermanos y a mis padres por toda la vida feliz que he compartido junto a ellos.

Gracias al laboratorio de WeSST Lab y a todos sus integrantes por su constante camaradería y buena onda.

A mi profesor guía Ángel Jiménez por su constante apoyo y soluciones que me permitieron seguir avanzando en la memoria a pesar de los obstáculos.

A mis compañeros de bachillerato especialmente al Sr. Felipe Zúñiga por las múltiples peripecias que sufrimos cuando fuimos compañeros y sus epifanías que eran un rayo de esperanza al momento de realizar las complejas tareas que eran asignadas.

A mis primos por tantos momentos de alegría.

A Dios por todo lo bueno que he recibido, en la vida.

Tabla de contenido

Índice de Tablas

Tabla de Figuras

1. Introducción

En este capítulo se desarrolla la presentación del tema del estudio propuesto. Este estudio comienza con la entrega de antecedentes generales y el contexto institucional del trabajo para entender el ambiente en que se desarrolla. Continúa con la explicación del proyecto a nivel general, se entrega el contexto en el cual se origina la necesidad de realizar este trabajo, luego se indica los objetivos específicos que hará que se cumpla el objetivo general, luego se plantea la hipótesis de investigación. Se plantea los resultados esperados, el alcance del proyecto y al final se detalla la metodología del trabajo en donde se muestran las condiciones de satisfacción de cada etapa.

1.1 Antecedentes Generales

Los medios de comunicación nos informan habitualmente de agresiones físicas entre personas, muchas en una situación de dominancia respecto al otro [1,2]. El análisis posterior de algunos de estos hechos indicarán que existió agresión verbal que al ser detectada y detenida preventivamente podría haber evitado ese hecho lamentable. Hasta ahora no existe un método eficaz universal que proteja a las personas respecto a ese tema. Por otra parte también es conocido el alto grado de estrés con que la sociedad en general desarrolla sus actividades, especialmente en las grandes urbes [3,4], situación que agudiza el problema, especialmente para personas que interactúan cara a cara con personas. Hoy se aplican varias acciones que tienden a detectar y prevenir estos hechos (e.g. alarmas, cámaras de video, vigilancia, etc.), pero esos métodos están orientados a proteger posesiones más que a individuos y tienen limitaciones (e.g. cámaras de vigilancia menos efectivas de noche).

En este contexto se hace necesario agotar los medios para detectar agresiones verbales. El desarrollo de sistemas y algoritmos de equipos móviles disponibles y usados en la actualidad en nuestra sociedad, permiten proponer soluciones avanzadas que ayuden a satisfacer el requerimiento planteado.

A la fecha y según el estado del arte realizado en este trabajo, se han desarrollado estudios que abordan esta necesidad en campos de audio específicos (disparos, roturas de vidrios, gritos), que están más bien enfocados al ámbito de la vigilancia que contra delincuencia. De todas formas, los modelos construidos y entrenados en estos ámbitos han tenido resultados positivos en su clasificación. Sin embargo, al examinar la literatura, no ha habido estudios que aborden en plenitud el tema de las agresiones verbales que no se enfoquen en el estado emocional.

.

1.2 Contexto institucional

Este trabajo de título se desarrolla en el marco del proyecto Fondecyt otorgado en elaño 2013 titulado A Cognitive Resource-Aware Mobile Service Framework to Support Human-Computer-Interactions in Ubiquitous Computing Environments a cargo delprofesor del Departamento de Ingeniería Industrial Ángel Jiménez Molina. En este proyecto se pretende desarrollar un mecanismo de ingeniería cognitiva que seleccione, componga y desarrolle funcionalidades, durante el tiempo en que se ejecute, tomando enconsideración el contexto situacional y los recursos cognitivos empleados por el usuario deacuerdo a las tareas HCI realizadas.

1.3 Descripción del proyecto

1.3.1 Consideraciones preliminares

Algunos autores han planteado que la agresión verbal es extremadamente difícil de definir ya que el habla humana es ampliamente variada [5]. Y además es difícil determinar objetivamente el deseo de dañar del hablante [6]. Para ello es necesario establecer ciertas consideraciones acerca de la definición:

- La definición de agresión verbal en este trabajo será la siguiente [6]: “Cualquier oración o frase aislada y utilizada como una reprimenda, una orden áspera, una habladuría, un insulto, un rechazo, una afirmación hostil de posesión o prioridad, declaración objetiva insensible, acusación, crítica u obscenidades".

1.3.2 Proyecto

Dos o más personas discuten en forma vociferante, con fuertes e hirientes palabras llegando a una situación en que la agresión física parece inminente. Sus volúmenes de voz se elevan y cambian repetidamente. Externamente se observa que sus rostros se contraen, la presión arterial sube cada vez que responde, sus músculos están hinchados. Esta es una situación repetitiva que inquieta a los presentes que no participan de la discusión. Esta situación puede llegar a la agresión física.

Dado el nivel de desarrollo de la tecnología que hoy se dispone, lo que propone esta tesis es detectar por medio de sonido la existencia de agresión verbal entre personas y diferenciarla de conversaciones normales.

Trabajos futuros podrían mejorar el modelo y entregar predicciones más certeras de lo que entregará el presente trabajo para, a futuro, detectar y poder inclusive prevenir agresiones cuando los tonos alcancen.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo de clasificación de agresiones verbales a través de algoritmos de machine learning.

1.4.2 Objetivos específicos

- Estado del arte para evidenciar el avance actual de los modelos de detección de eventos en audio y agresiones verbales e identificar las fuentes de datos de agresión verbal
- Diseñar un método para la extracción de datos y features, a partir de archivos de audio
- Determinar si las features extraídas son capaces de discriminar entre situaciones con agresión verbal y sin agresión verbal
- Evaluar modelos de clasificación mediante técnicas de minería de datos para agresión verbal y conversaciones sin agresiones verbales comparando con resultado de trabajos anteriores
- Discusión general y vincular el trabajo desarrollado con posibles trabajos futuros

1.5 Hipótesis de investigación

La literatura ofrece una amplia gama de características propias del ámbito del análisis de sonido para describir señales de audio y variados modelos para detectar eventos de audio.

La hipótesis consiste en comprobar que es posible detectar la existencia de una agresión verbal a través del análisis de señales de audio con la data recolectada.

1.6 Resultados esperados y alcances

Para este trabajo se esperan los siguientes resultados:

- Estado del arte de modelos de detección de eventos de audio y agresiones verbales Discusión general y vincular el trabajo desarrollado con posibles trabajos futuros
- Obtención de fuente de datos de agresiones verbales para iniciar el proceso KDD

- Resultados de análisis estadístico
- Resultados de los modelos de clasificación
- Aceptación o rechazo de la hipótesis de investigación

En cuanto a los alcances:

- No se pretende que se incluyan estrategias para mejorar el rendimiento de modelos o la implementación de nuevos modelos que no hayan sido utilizados en trabajos previos
- Solo se recolectan eventos de agresión verbal en personas de sexo masculino
- La agresión verbal en este trabajo, es detectada en cuanto a sus características de audio, sin incluir en el análisis el contenido en los mensajes transmitidos
- La agresión verbal considerada es la que terceros determinen al escuchar grabaciones de audio de interacciones humanas
- La atención de este trabajo no está enfocado en el estado emocional real de quienes participaron en los registros de audio utilizados, sino en la existencia de agresiones verbales consideradas por terceros

1.7 Metodología

1) Estudio del estado del arte: Se revisará la literatura para descubrir las metodologías y técnicas que se han utilizado recientemente en el ámbito de la detección de eventos de audio, en particular en agresiones verbales
2) Extraer datos de audio de agresiones verbales de manera correcta para ser incorporados al proceso KDD
3) Aplicación del Proceso KDD, aplicando las etapas de pre procesamiento y culminando con conocimiento nuevo
4) Resultados: Medidas estadísticas del desempeño del modelo para la detección correcta de agresión verbal con vociferaciones
5) Construcción de prototipo de la aplicación con el modelo desarrollado incorporado
6) Evaluación del prototipo con data de testeo generada previamente.
7) Discusión: Comparación con los diferentes métodos utilizados y sus resultados y posible trabajo futuro
8) Conclusiones

2. Marco Teórico

En este capítulo se profundiza en conceptos vistos en el capítulo anterior y en otros que se verán en capítulos siguientes. Se comienza describiendo el modelo, “Aural words” que utiliza Vincenzo Carletti et al. [7] para detectar eventos de interés en datos de audio, que será utilizado en este trabajo. Posteriormente definiciones fisiológicas, conceptos de análisis de señales, trabajos anteriores y finalmente el proceso KDD.

2.1 El clasificador de palabras auditivas (del inglés “aural words”)[7]

El clasificador de palabras auditivas, adopta el paradigma de la bolsa de palabras y tiene dos ventajas principales sobre otras técnicas presentes en la literatura: la capacidad de adaptarse automáticamente a sonidos cortos impulsivos y largos sostenidos y la capacidad de trabajar en entornos ruidosos donde los sonidos de interés se superponen a sonidos de fondo que posiblemente tienen características similares.

Si bien, el metodo de “bag of words” no se aplica de forma exacta en este trabajo, una variación en su concepto si es utilizada, por lo tanto una breve explicación de este metodo el punto 2.1.1 permitirá introducir de mejor manera al punto 2.1.2 en donde se explica en detalle el clasificador de palabras audibles en el cual se basará este trabajo. En dicho modelo las palabras serán análogas a fragmentos pequeños de sonido y la bolsa de palabras será análogo a una bolsa de “palabras audibles”. Los histogramas también son análogos: en la bolsa de palabras, se usa frecuencia de palabras y en “la bolsa de palabras audibles” se usa la frecuencia de “palabras audibles”.

2.1.1 El método BoW (Bag of words)

En este método, todas las palabras de un documento se tratan como términos primarios de ese documento. Además se asigna un peso a cada término en función de su importancia, determinada normalmente por su frecuencia de aparición en el documento (de este modo, no se toma en consideración el orden, la estructura, el significado, etc. de las palabras). Por lo tanto en primera instancia, cada documento queda representado por un histograma de la frecuencia de cada palabra de la bolsa de palabras presente en él. La bolsa de palabras es un conjunto que se construye unificando las palabras de todos los documentos, donde lógicamente cada palabra dentro de este conjunto tiene frecuencia igual a 1. De esta manera se pueden realizar clasificaciones en clústeres (e.g. K-means) de documentos según su similitud en su histograma y asi, de esta forma, al recibir un documento nuevo éste puede ser traducido en dicho histograma y ser asignado a las clases definidas por los núcleos del algoritmo.

2.1.2 Descripción del método de clasificación mediante palabras auditivas

Este modelo fue usado por Vincenzo Carletti et al., para la detección de eventos basados en audio en donde se propone un enfoque para la detección de eventos de audio basados en el paradigma de "bolsa de palabras" (del inglés, Bag of Words), comúnmente usado para la categorización de documentos de texto y recientemente aplicado con éxito para detección de objetos basados en video y otros problemas similares [11]. El modelo está enfocado a la detección de eventos de audio de la siguiente forma:

Dadas M clases de sonidos de interés $\{C_1...C_M\}$, cada una representada por un conjunto de muestras y una transmisión de audio, el objetivo es encontrar si existen eventos de interés dentro de la transmisión. Todos los sonidos que no sean eventos de interés son considerados sonido de fondo y se le asignan la clase C_0.

En palabras simples, en la primera etapa, este modelo calcula características de bajo nivel, en pequeños tramos de audio de cantidad fija, medidos en unidades de PCM. Una cierta cantidad de unidades L de PCMs conforma un tramo y una cierta cantidad de tramos N conforma una ventana R. La cantidad de elementos de toda ventana R es constante dado que los parámetros L y N son constantes para una configuración predefinida y como la frecuencia F_s (la frecuencia de muestreo) también lo es, la duración de cada ventana R es constante. Para cada tramo i $i \in \{1...N\}$, se construye un vector v_i cuya dimensión será el número de características (del inglés features) a utilizar. Estos vectores pueden ser utilizados para dos cosas:

En la fase de entrenamiento del primer nivel, seleccionando solo los audios de la clase de interés que se quiere detectar, se usan los vectores de los tramos para entrenar el modelo de clasificación por clústeres haciendo uso del algoritmo K-Means. Esto produce un conjunto de K centroides, que conforman el libro de código (del inglés codebook) del sistema:

$$CB = (w_1, ..., w_k) \quad (1)$$

De esta manera las características de los tramos de eventos de sonido, permitirán tener una medida de que tan cerca está el tramo de un clúster que fue asociado a la ocurrencia de una clase de interés (i.e. agresión verbal).

En la fase de entrenamiento del segundo nivel, se busca en el libro de código la palabra verbal (i.e. clúster) más cercana a v_i. Definiendo b_i como el índice de dicha palabra se tiene:

$$b_i = \arg\min_j \left| v_i - w_j \right| \text{ para } j \in \{1...K\} \quad (2)$$

Finalmente, el vector del segundo nivel de características $U = (u_1, ..., u_k)$ es definido de la siguiente manera:

$$u_j = \sum_{i=1}^{N} \delta\left(b_i, j\right) \text{ para } j \in \{1...K\} \quad (3)$$

Donde $\delta(\bullet)$ es el delta de Kronecker.

Por lo tanto el vector de características del segundo nivel es el histograma con las frecuencias de las N palabras verbales, asociadas a la agresión verbal, detectadas en la ventana R.

Con los vectores de características de segundo nivel, se utiliza el clasificador SVM (Support Vector Machine), utilizando un conjunto de entrenamiento etiquetados con la clase de interés en el intervalo de tiempo que dura una ventana R, definida por los parámetros del problema. Se usará SVM lineal ya que reportó mejores resultados en el trabajo anterior.

El sistema está organizado con varios casos SVM operando en paralelo, es decir, se tienen $M+1$ clasificadores (donde M es el número de las clases a reconocer). El clasificador i (con $i = 0,...,M$) es entrenado usando como ejemplos positivos las muestras de la clase C_i y como ejemplos negativos todas las demás clases. Los clasificadores producen una salida y una puntuación. Si al menos un clasificador produce una salida positiva, el vector se le asigna la clase con la puntuación máxima. Si todos los clasificadores dan una salida negativa, el vector se asigna a la clase C_0 (sonidos de fondo).

Para concluir, este modelo es elegido para ser usado en el presente trabajo ya que sus autores reportaron dos principales ventajas sobre otras técnicas presentes en la literatura: la habilidad de adaptarse tanto a sonidos impulsivos cortos como a sonidos prolongados largos y la habilidad de funcionar en ambientes ruidosos, donde los sonidos de interés están sometidos a intervenciones de sonidos de otras fuentes que eventualmente pueden tener características similares.

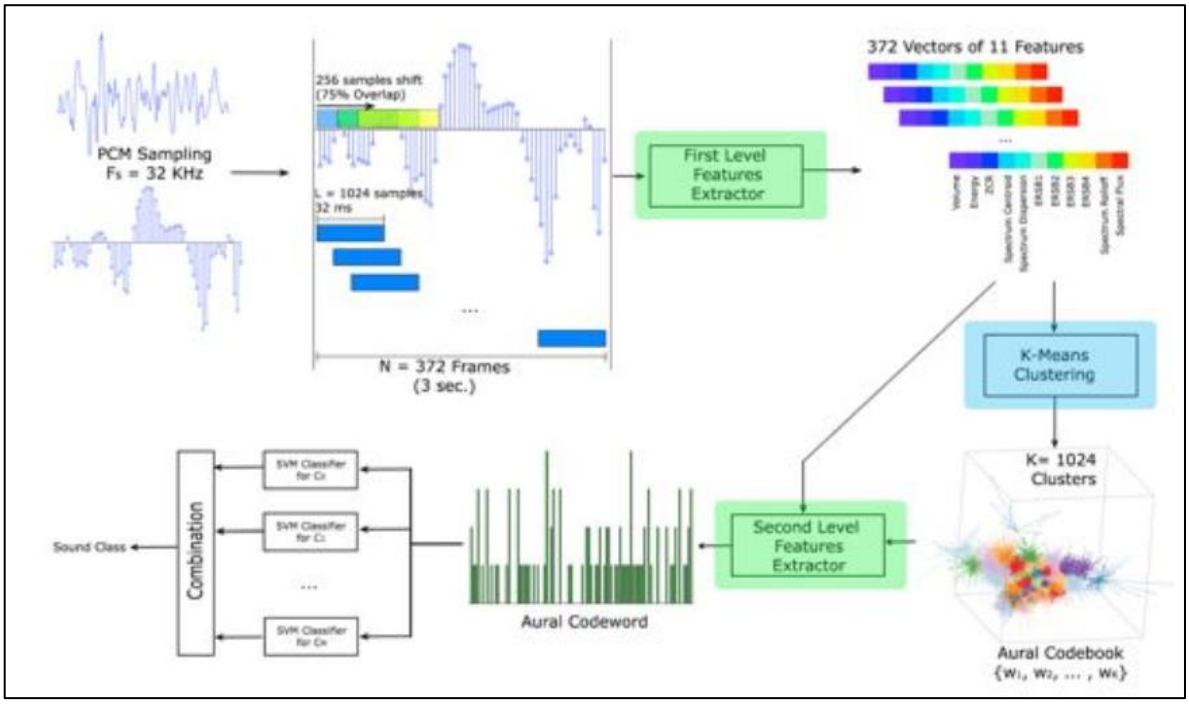


Figura 1: La arquitectura de sistema del método propuesto. Los módulos usados en fase de entrenamiento y operativa están en verde, mientras que el modulo azul es usado solo durante la fase de entrenamiento. Los valores de F_s, N, L, K *, también son reportados [7].*

2.2 Definiciones fisiológicas

2.2.1 La voz humana

Las señales de sonido como la voz, consisten en la variación de la presión del aire por ondas longitudinales. En humanos y primates, la fuente de emisión de estas ondas involucra la vibración de las cuerdas vocales de la laringe. La fuente de sonido viaja hasta el tracto vocal, donde las cavidades orales y nasales del tracto vocal actúan como filtro pasando energía acústica en algunas frecuencias y atenuando en otras de acuerdo a la función de transferencia especifica según al tamaño y forma. Por lo tanto el sonido irradiado en los labios refleja la acción combinada de una fuente y un filtro acústico [19].

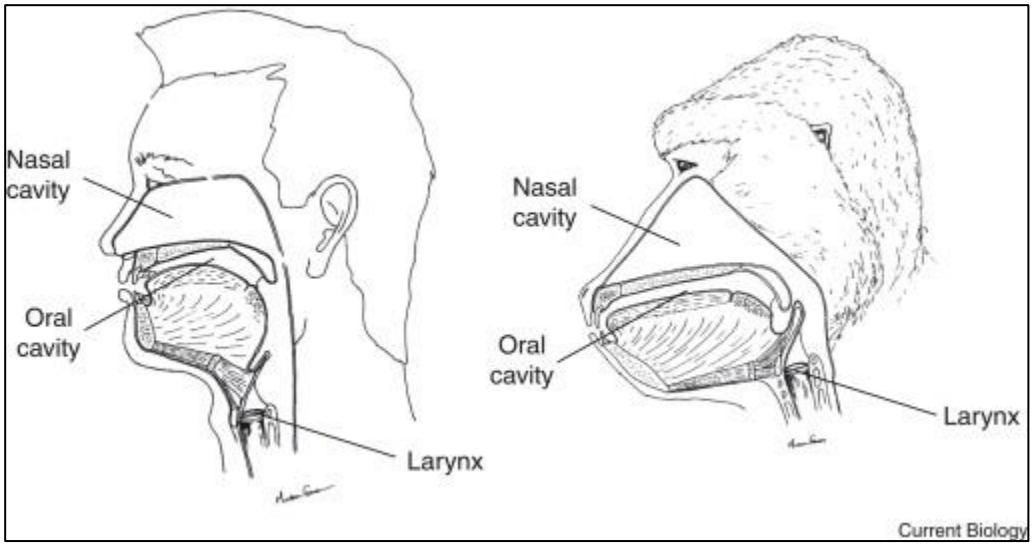


Figura 2: Anatomía del aparato vocal. Imagen extraída de [19]

Comparado con el rango completo de audición humana, la voz humana cubre un ancho de banda relativamente estrecho, aproximadamente desde 100 Hz a 6 kHz. Dentro de ese rango, la distribución de potencia el habla es fuertemente ponderado en frecuencias por debajo de aproximadamente 1 kHz con cerca de un 80% de la energía concentrada en el rango debajo de 500 Hz [20].

Mientras el habla tiene muy poco contenido de alta frecuencia, casi toda la energía de consonantes ocurre sobre 1 kHz [20].

Como un mínimo practico, el ancho de banda de un sistema de reproducción del habla, debe extenderse desde los 300 Hz hasta los 3.5 kHz (frecuencia de respuesta de un receptor telefónico), el cual puede mejorarse aumentando los decibeles entre los 2 kHz y 5 kHz [20].

2.2.2 El sistema auditivo

El oído humano consta de 3 partes separadas con diferentes funciones en el mecanismo de audición [21].

El oído externo: El oído externo está formado por la oreja y por el conducto auditivo externo, cuya misión es captar las ondas sonoras y llevarlas hasta la membrana timpánica [21].

El oído medio: Esta situado en el interior del hueso temporal y es una pequeña cámara conectada con el oído externo pro el tímpano. Cuando la membrana timpánica vibra al llegarle las ondas sonoras, transmite estas vibraciones a tres huesecillos que hay en el oído medio, llamados: martillo, yunque y estribo. Las vibraciones se transmiten desde el estribo al líquido que existe en el oído interno. La trompa de Eustaquio tiene por misión mantener una presión similar a la atmosférica en el interior del oído medio [21].

El oído interno: Contiene los órganos del equilibrio y de la audición (cóclea). En la cóclea se encuentran los receptores auditivos que se estimulan por el movimiento del líquido que existe en el interior de la cóclea. Los estímulos nerviosos van a través del nervio acústico hasta la corteza auditiva en el lóbulo temporal [21].

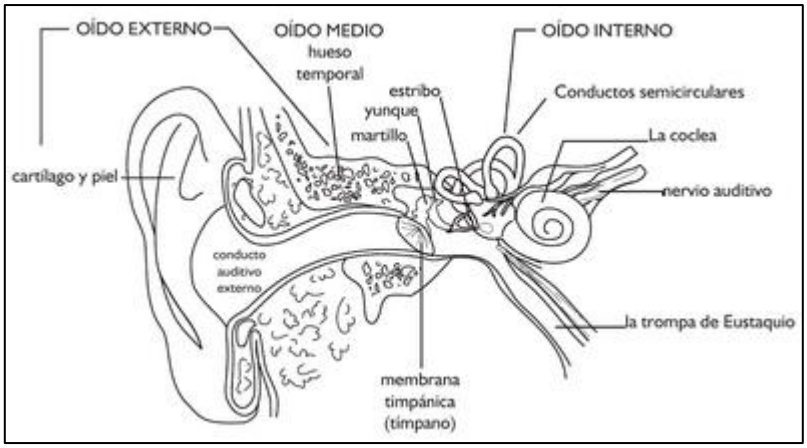


Figura 3: Esquema del oído humano. Imagen extraída de [21]

Incluso los oídos humanos más perfectos y sin daños, solo pueden detectar un pequeño rango de sonidos posibles, aproximadamente desde 20 a 20 kHz. El habla humana ocurre casi exclusivamente desde los 300 Hz hasta los 3500 Hz, con la mayoría del habla debajo de los 1000 Hz [22].

El sonido está distribuido en diferentes partes de la cóclea de acuerdo a la frecuencia. Tonos de diferentes frecuencias, producen vibraciones de máxima amplitud a lo largo de la longitud de la cóclea. La base cerca del estrubo, es rígida y estrecha y responde más a los sonidos de alta frecuencia (agudos). El ápice lejos del estribo, es ancho y responde más a los sonidos de baja frecuencia [8].

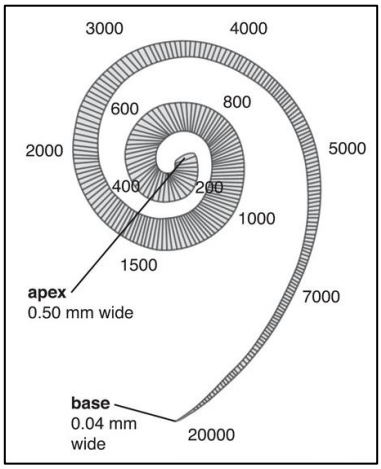


Figura 4: Frecuencias específicas, causan vibraciones de máxima amplitud en diferentes puntos a lo largo de la cóclea. Los números en el diagrama representan frecuencia en Hertz. Imagen extraída de [8].

Para abordar la capacidad auditiva, es útil entender como escuchamos en el día a día en cuanto a la frecuencia e intensidad.

Respecto a la frecuencia, los sonidos más importantes están en el rango entre 250 a 6000 Hz. El habla incluye una mezcla entre sonidos de alta y baja frecuencia. Vocales como una “o” corta en la palabra “hot” tienen bajas frecuencias (250 a 1000 Hz) y son generalmente fáciles de escuchar. Consonantes como “s”, “h” y “f”, tienen más altas frecuencias (1500-6000 Hz) y son más difíciles de escuchar. Las consonantes transmiten la mayor parte del significado de lo que se dice. Alguien que no puede oír los sonidos de alta frecuencia tendrá dificultades para entender el habla y el lenguaje [22.3].

En cuanto al volumen, una persona normal puede escuchar entre 0 y 140 dB [22.3].

Lo anterior, permite apreciar el rango de frecuencias del habla humana, ya que esta alcanza sus mayores frecuencias en consonantes alrededor de los 6000 Hz, por lo tanto la grabación de agresiones verbales deberían ser con una frecuencia de muestreo de 12000 Hz o mayor, dado el teorema de sampleo de Nyquist-Shannon, el cual indica que la frecuencia de sampleo debe ser por lo menos el doble de la frecuencia máxima de la señal para representar y/o reconstruir una señal analógica original [42].

Con esta información y la del punto 2.2.1, se puede afirmar que el habla humana tiene un ancho de banda que cae prácticamente en su totalidad dentro del rango de audición humana. Por lo tanto se puede desprender que las manifestaciones de violencia verbal que hemos escuchado, se encuentran dentro del rango de frecuencias audibles para el ser humano, es decir son menores a 20000 Hertz.

2.2.3 El estado emocional de ira

La ira puede ser definida como “un estado emocional que puede variar en intensidad de irritación leve a intensa furia y rabia. La ira tiene efectos físicos, incluyendo la aceleración de la frecuencia cardíaca y los niveles de presurización sanguínea de adrenalina y noradrenalina” [Medical Dictionary ,2004]. La ira es una emoción que se puede describir en dos formas, la ira caliente o explosiva (rabia) y la ira fría (irritación) (Scherer, 1986).

La ira es una emoción como cualquier otra emoción, que suele ser una respuesta a algo que está sucediendo o que ha sucedido y puede surgir como una respuesta a los sentimientos o la vulnerabilidad y la inquietud debido a una frustración de deseo, sentimientos de dolor, miedo, vulnerabilidad, una amenaza a sus necesidades (emocional o físico), o un reto [23].

Simplemente, la ira es una reacción no planificada a un factor de estrés. La ira es una emoción universal, pero no todos responden a la ira con la agresión o la violencia [23].

Cuando la ira se maneja de manera constructiva (por ejemplo, comunicación asertiva, razonamiento crítico), la ira puede ayudar a mantener a las personas seguras y cumplir con los fines. La ira se vuelve poco saludable si se interpone en el camino del funcionamiento de una persona, las relaciones, o pone a otros en riesgo. Cuando la ira

se deja sin control y se intensifica, los resultados a menudo conducen a formas negativas de agresión o violencia. La actuación de la ira puede satisfacer necesidades inmediatas, pero a expensas de causar daño emocional o físico a nosotros mismos o a otros [23].

La expresión de ira puede ser a través de comportamientos activos o pasivos. En el caso de la emoción "activa", la persona enojada "ataca" verbalmente o físicamente a un objetivo deseado. Cuando el enojo es una emoción "pasiva", se caracteriza por el silencio, el comportamiento pasivo-agresivo (hostilidad) y la tensión [24].

El enojo puede estar correlacionado con un F0 (frecuencia fundamental, ver glosario en anexos) de alta media (i.e., un aumento en el tono de la voz), un aumento de la energía (i.e., alta presión del habla) de la vocalización, un aumento de la energía formante de alta frecuencia, un aumento en la tasa de articulación de la emisión y es caracterizada por una calidad de la articulación “tensa” y una calidad de voz entrecortada. La "ira caliente" también se caracteriza por un aumento en la variabilidad de F0 y un aumento en el rango de F0, mientras que la "ira fría" también se caracteriza por un contorno de entonación dirigido hacia abajo (i.e., es decir, el tono de la voz disminuye a lo largo del enunciado) [52].

2.3 Análisis de señales de audio

Los sonidos reales son grabados en formato digital para su análisis. El método para su grabación y reproducción implica el conocimiento de algunos conceptos útiles para entender de mejor manera el proceso. A continuación se explican algunos necesarios para entender de mejor manera el trabajo realizado.

2.3.1 Modulación por impulsos codificados

La Modulación por impulsos codificados (PCM por sus siglas en inglés) es un esquema digital para transmitir datos analógicos. Las señales en PCM son binarias; Es decir, sólo hay dos estados posibles, representados por lógico 1 (alto) y lógico 0 (bajo).

Esto es cierto, no importa cuán compleja sea la forma de onda analógica. Usando PCM, es posible digitalizar todas las formas de datos analógicos [25].

Para obtener PCM a partir de una forma de onda analógica en la fuente (extremo del transmisor) de un circuito de comunicaciones, la amplitud de la señal analógica es muestreada (medida) a intervalos de tiempo regulares. La frecuencia de muestreo, o número de muestras por segundo, es varias veces la frecuencia máxima de la forma de onda analógica en ciclos por segundo o Hertz. La amplitud instantánea de la señal analógica en cada muestreo se redondea al más cercano de varios niveles predeterminados específicos. Este proceso se llama cuantificación. El número de niveles es siempre una potencia de 2 - por ejemplo, 8, 16, 32 o 64. Estos números pueden ser representados por tres, cuatro, cinco o seis dígitos binarios (bits), respectivamente. La salida de un modulador de códigos de impulsos es, por lo tanto, una serie de números binarios, cada uno representado por una potencia de 2 bits [25].

En el destino (extremo receptor) del circuito de comunicaciones, un demodulador de códigos de impulsos convierte los números binarios en impulsos que tienen los mismos niveles cuánticos que los del modulador. Estos impulsos se procesan adicionalmente para restaurar la forma de onda analógica original [25].

2.3.2 Códecs

Un códec comprime o descomprime archivos multimedia como canciones o videos. Aplicaciones como Windows Media Player utilizan códec para reproducir y crear archivos multimedia [26].

Un códec puede consistir de dos partes: un codificador que comprime el archivo multimedia (codificación), y un decodificador que descomprime el archivo (decodificación). Algunos códec incluyen ambas partes, y otros códecs sólo incluyen uno de ellos [26].

En palabras simples traducen señales de audio analógicas a patrones digitales para almacenamiento, y patrones digitales a señales de audio analógicas para reproducción [27].

2.3.3 Formato WAVE

Usualmente conocido como WAV (gracias a su extensión de archivo), este formato de audio soporta formatos de audio comprimidos y no comprimidos. WAV trabaja con numerosos codecs de audio ampliamente disponibles. [27].

El formato WAV se basa en la modulación por codificación de pulsos (PCM, por sus siglas en inglés) el cual se utiliza en formatos de grabación de audio digital, así como en la creación de CDs de audio. PCM es un formato sin comprimir, lo que significa que si bien los archivos pueden ser grandes, éstos contienen información detallada de alta calidad, haciendo que el formato WAV sea perfecto para aplicaciones de audio profesionales. [29].

2.3.4 Formato AAC

Es un códec de audio con pérdidas (“lossy”) y de tamaño comprimido [28]. Se recomienda una tasa de bits tan baja como 80 kbps para efectos de sonido y grabaciones de voz [30].

La tarea básica de un sistema perceptivo de codificación de audio es comprimir los datos de audio digital [31] de una manera que:

- La compresión es lo más eficiente posible, es decir, el archivo comprimido es lo más pequeño posible

- El sonido reconstruido (decodificado) suena exactamente (o lo más cerca posible) al audio original antes de la compresión

La técnica para hacer esto se llama codificación perceptual y utiliza el conocimiento de la psicoacústica para alcanzar el objetivo de la compresión eficiente pero inaudible. La codificación perceptual es una técnica de compresión con pérdidas ("lossy compression" en inglés), es decir, el archivo decodificado no es una réplica de bits exactos de los datos de audio digitales originales [31].

La siguiente figura muestra un diagrama de bloques básico del sistema de codificación perceptual:

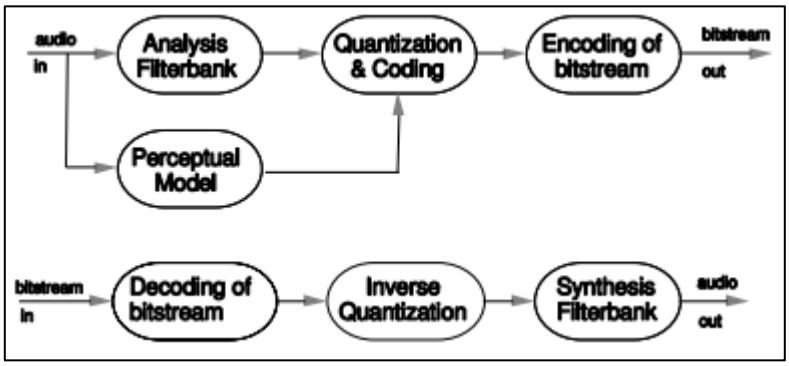


Figura 5: Diagrama de bloques del sistema de codificación/decodificación perceptual. Imagen extraída de [31]

Este consiste en los siguientes bloques [31]:

Banco de filtros (Filter Bank en inglés): Se utiliza un banco de filtros para descomponer la señal de entrada en componentes espectrales sub muestreadas (dominios tiempo/frecuencia). Junto con el banco de filtros correspondiente en el decodificador, forma un sistema de análisis / síntesis.

Modelo perceptivo (Perceptual model en inglés): Utilizando la señal de entrada del dominio de tiempo y / o la salida del banco de filtros de análisis, se calcula una estimación del umbral de enmascaramiento real (dependiente del tiempo y la frecuencia) usando reglas conocidas de la psicoacústica. Esto se llama el modelo perceptual del sistema de codificación perceptual.

Cuantización y codificación: Los componentes espectrales se cuantifican y codifican con el objetivo de mantener el ruido que se introduce por cuantificación por debajo del umbral de enmascaramiento. Dependiendo del algoritmo, este paso se realiza de maneras muy diferentes, desde la compresión de bloques simples hasta sistemas de análisis por síntesis que utilizan compresión adicional sin ruido.

Codificación de flujo de bits: Se usa un formateador de flujo de bits para montar el flujo de bits, que típicamente consiste en los coeficientes espectrales cuantificados y codificados y alguna información lateral, e.g. Información de asignación de bits.

Al utilizar archivos con formatos AAC que solo eliminan bandas de frecuencia y datos que una persona en promedio no escucha basados en técnicas de audio perceptuales, se puede tener cierta confianza en que no se eliminan frecuencias importantes para detectar eventos de agresión verbal, ya que éstas ocurren dentro del rango de frecuencias que si escucha el ser humano.

2.4 Trabajos anteriores

Para hacer un resumen de los trabajos previos, estos fueron organizados en tablas en cuyos atributos se considerarán el objetivo, las características utilizadas, clases de interés, algoritmos y medidas de desempeño estadístico. Se comienza por el trabajo de Vincenzo et al. que es adaptado y usado en el presente trabajo.

2.4.1 Estado del arte

El estado de arte fue construido buscando trabajos relacionados usando palabras clave propias del presente trabajo (e.g. “anger”, “speech”, “violence”, “audio”, “agression, etc). Además solo se consideraron trabajos superiores al año 2000. Se utilizó el motor de búsqueda Google Scholar [43].

Carletti, Vincenzo, et al. [7]. Publicado en IEEE. Año 2013, 23 citas en Google Scholar.

Objetivo	Detectar eventos basados en audio
Datos	Sonidos tomados por separado y combinados
Descriptores de bajo nivel	Spectral centroid (SC), Spectral spread (SS), Spectral rolloff, Spectral flux, Energy ratios in sub-bands (ERSB), Volume, Energy, Zero crossing rate (ZCR)
Clases de interés	Gritos, Disparos, Vidrios quebrándose, Ruido
Algoritmos/Metodos	K-Medias, Support Vector Machine
Resultados	Exactitud 95.8%

van Hengel, Peter WJ, and Tjeerd C. Andringa [9]. Publicado en IEEE, año 2007, 22 citas en Google Scholar.

Objetivos	Detectar evidencia de agresión por medio de análisis de audio en ambientes no controlados
Datos	~1400 grabaciones reales
Descriptores de bajo nivel	F0 (tono), F0 Armónicas, Forma espectral, Energía temporal y espectral, SNR (signal to noise ratio)
Clases de interés	Grito agresivo, Ruido
Algoritmos/Métodos	Modelo cóclea humana para separación de sonidos de fondo y primer plano. Detección de voz. Detección de tono. Distorsión de la voz
Resultados	Sensibilidad 100% Precisión 93%

Foggia, Pasquale, et al. [10]. Publicado en IEEE, año 2014, 8 citas en Google Scholar.

Objetivos	Detección de eventos de interés a través de análisis de audio
Datos	Sonidos tomados por separado y combinados
Descriptores de bajo nivel	Funciones Haar
Clases de interés	Gritos, Disparos, Vidrios quebrándose, Ruido
Algoritmos/Métodos	AdaBoost Banco de filtros audibles gamma-tono (Derivado de distribución gamma, tono sinusoidal y Anchos de Banda Rectangulares Equivalentes (del inglés Equivalent Rectangular Bandwith))
Resultados	95.89% de exactitud

Lecomte, Sébastien, et al. [11]. Publicado en IEEE, año 2011, 26 citas en Google Scholar.

Objetivos	Método no supervisado para detección en tiempo real de eventos anormales en el contexto de la vigilancia con audio
Datos	Sonidos tomados por separado y combinados
Descriptores de bajo nivel	Energías de banco de filtros lineal de Fourier
Clases de interés	27 eventos anormales, ambiente
Algoritmos/Métodos	OC-SVM (SVM de una clase) Coeficiente de correlación intra segmento
Resultados	0.2% < EER < 12.5% (según SNR) para cada clase

Valenzise, Giuseppe, et al. [12]. Publicado en IEEE, Año 2007, 201 citas en Google Scholar.

Objetivo	Vigilancia de video basada en sistema de audio que detecta automáticamente eventos de audio anómalos en una plaza publica
Datos	Películas e internet
Descriptores de bajo nivel	Características temporales Características de energía Características espectro Características perceptuales Distribución espectral Basados en correlación
Clases de interés	Gritos, disparos, ruido

Algoritmos/Métodos	GMM Selección de features
Resultados	Precisión 93%

Atrey, Pradeep K., Namunu C. Maddage, and Mohan S. Kankanhalli[13]. Publicado en IEEE, año 2006, 150 citas en Google Scholar.

Objetivos	Enfoque jerárquico para detección de eventos de audio basados en vigilancia
Datos	Audio por separado
Descriptores de bajo nivel	ZCR, LPC, LPCC, LFCC
Clases de interés	Hablar, gritar, golpear, pasos (corriendo y caminando), ruido
Algoritmos/Métodos	GMM
Resultados	Exactitud 89% y 90% para clases vocales - no vocales

2.4.2 Descriptores de bajo nivel en audio

2.4.2.1 Centroide espectral

El centroide espectral es el centro de masa del espectro y es calculado de la siguiente manera:

$$SC = \frac{\sum_{i=1}^{L_F} i \frac{F_s}{L_F} |X(i)|}{\sum_{i=1}^{L_F} |X(i)|} \quad (4)$$

Donde F_s es la velocidad de muestreo del flujo de audio de entrada, L_f el largo de la señal de entrada y $|X(i)|$ el módulo de la FFT de la señal de entrada

Esta correlacionado con el “Brillo” o “Agudeza” de una señal de audio de acuerdo a Zwicker and Fastl (1999). Otros estudios (e.g., Kendall and Carterette 1996) sugieren

que el radio entre F_0 (la frecuencia fundamental) y el centroide espectral esta mejor correlacionada con el brillo que el centroide espectral en solitario

2.4.2.2 Desviación estándar

La desviación estándar espectral (i.e. la raíz cuadrada del segundo momento espectral), es calculada como la dispersión de los componentes de frecuencia alrededor del centroide.

$$SS = \sqrt{\frac{\sum_{i=1}^{L_F}\left[i\frac{F_s}{L_F} - SC \right]^2 \left| X(i) \right|}{\sum_{i=1}^{L_F}\left| X(i) \right|}} \quad (5)$$

2.4.2.3 Rotación espectral

La rotación espectral es la medida de inclinación del espectro y es definido como la frecuencia f_{ro} a la cual el P% de los componentes espectrales de la señal están a menores frecuencias. En este trabajo se considera F =90 y determina el valor f_{ro} de la siguiente relación:

$$\sum_{i=1}^{f_{ro}}\left| X(i) \right| = \frac{P}{100}\sum_{i=1}^{F_{\max}}\left| X(i) \right| \quad (6)$$

2.4.2.4 Flujo espectral

El flujo espectral representa una versión cuadrática de la diferencia espectral simple. En este trabajo se usa la versión sin normalizar

$$SF = \sum_{i=1}^{L_F}\left[X_n(i) - X_{n-1}(i) \right]^2 \quad (7)$$

2.4.2.5 Ratio de energías en sub-bandas

Los ratios de energía en sub bandas ERSB (del inglés Energy Ratios in Sub-Bands) dan una aproximación tosca de la distribución de energía del espectro. La señal del espectro se divide en cuatro sub-bandas, y por cada sub banda se computa la razón entre la energía contenida en esa sub banda y la energía total del fragmento de audio.

$$ERSB_n = \frac{\sum_{i=k_{n1}}^{k_{n2}} \left| X\left(i\right) \right|^2}{\sum_{i=1}^{F_{\max}} \left| X\left(i\right) \right|^2}$$

$$\left[k_{n1}, k_{n2}\right] = \begin{cases} \left[1, 630\right], & n = 1 \\ \left[631, 1720\right], & n = 2 \\ \left[1721, 4400\right], & n = 3 \\ \left[4401, 22000\right], & n = 4 \end{cases} \quad (8)$$

2.4.2.6 Volumen

El volumen se calcula como el valor cuadrático medio RMS (del inglés Root Mean Square), de la amplitud de los valores de las muestras en los fragmentos de audio.

$$V = \sqrt{\frac{1}{L} \sum_{i=1}^{L} x\left(i\right)^2} \quad (9)$$

2.4.2.7 Energía

La energía se calcula como la suma cuadrática de los valores de las muestras de audio.

$$E = \sum_{i=1}^{L} x\left(i\right)^2 \quad (10)$$

2.4.2.8 Tasa de cruces por cero

La tasa de cruces por cero ZCR (del inglés Zero Crossing Rate) es la tasa de cambios de signo a lo largo de un fragmento y es especialmente usado para caracterizar sonidos percusivos y ruido ambiental. Para un fragmento $x(i)$ de L muestras, la tasa de cruces por cero es la siguiente:

$$ZCR = \frac{1}{2L} \sum_{i=1}^{L} \left| \mathrm{sgn}\left(\mathrm{x}\left(\mathrm{i}+1\right)\right) - \mathrm{sgn}\left(\mathrm{x}\left(\mathrm{i}\right)\right) \right| \quad (11)$$

2.5 Knowledge Discovery in Databases

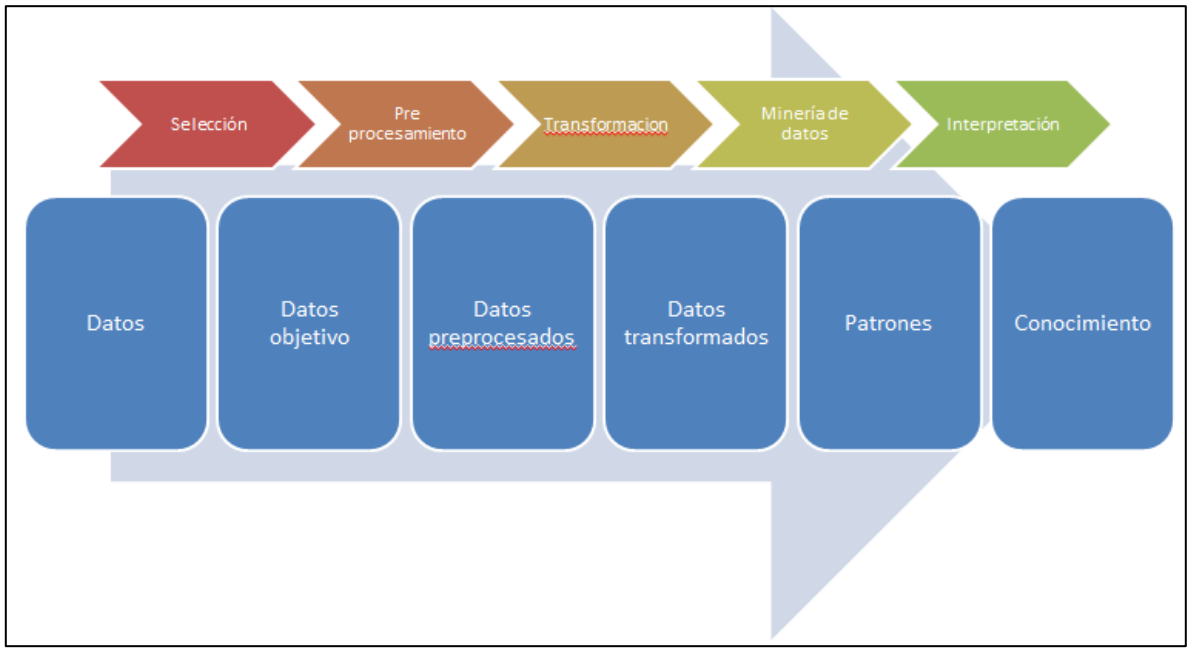


Figura 6: Etapas generales del proceso KDD

La metodología se compone de 9 pasos principales [17]:

1) Comprender el dominio de la aplicación: En esta etapa se debe obtener conocimiento previo relevante del problema que se quiere resolver y los objetivos de la aplicación. Reflexionar si lo mejor es poder predecir el valor de variables de interés en el futuro o si es mejor encontrar una caracterización que encuentre relaciones valiosas entre los datos El sonido reconstruido (decodificado) suena exactamente (o lo más cerca posible) al audio original antes de la compresión.

2) Crear un conjunto de datos objetivo: En esta etapa se debe seleccionar un conjunto de datos. Se debe evaluar si los datos disponibles tienen la calidad necesaria para tomarlos como base del trabajo a realizar. Se debe evaluar si los atributos de los datos son suficientes para explicar el problema planteado o en caso contrario, intentar una nueva realización de captura de datos, ya que son la base del desarrollo.

3) Limpieza de datos y pre procesamiento: Como sugiere el nombre se tratan los datos, corrigiendo sus errores en la medida de lo posible. Problemas típicos son: Valores perdidos, datos fuera de rango, datos mal escritos, eliminación del ruido, etc. Para errores de mayor magnitud es necesario recurrir a técnicas más sofisticadas para su tratamiento

4) Reducción de data y proyección: Consiste en encontrar parámetros útiles para representar la data de acuerdo al objetivo buscado y usar reducción de dimensión o métodos de transformación para reducir el número efectivo de

variables consideradas o encontrar una representación invariable de la data. Eliminar la data redundante es beneficioso para los pasos posteriores ya que no genera valor y ralentiza los trabajos de procesamiento.

5) Escoger la función de minería de datos: Se debe decidir el propósito del modelo derivado del algoritmo de minería de datos que pueden ser del tipo clasificación, clusterización o regresión. La decisión está altamente influida por el punto 1, en donde se reflexionó acerca de qué era lo mejor para encontrar conocimiento nuevo: predecir variables de interés o encontrar relaciones valiosas dentro de los datos.

6) Escoger algoritmos de minería de datos: Se seleccionan los métodos a usar para buscar patrones en la data. Es apropiado decidir cuál modelo se ajusta mejor al problema según sus características y evaluando que tenga los atributos necesarios para representar de buena manera el enfoque y el tipo de conocimiento que se pretende obtener. El algoritmo debe coincidir con el objetivo general del proceso KDD.

7) Mineria de datos: Encontrar patrones de interés en una forma de representación particular o en un conjunto de esas representaciones, incluyendo reglas de árboles de clasificación, regresiones, clústering, modelamiento secuencial. Es conveniente que el algoritmo sea ejecutado varias veces para calibrar mejor los parámetros utilizados.

8) Interpretación: Consiste en interpretar los patrones descubiertos y posiblemente volver a un paso previo así como la posible visualización de los patrones extraídos, remover patrones redundantes o irrelevantes y trasladar los útiles en términos entendibles por los usuarios.

9) Usar el conocimiento descubierto: Esta etapa consiste en incorporar el conocimiento al contexto del problema enmarcado inicialmente, y así comprobar y resolver conflictos con conocimiento anterior. Si todos los pasos fueron seguidos correctamente se habrá generado conocimiento nuevo que puede ser muy valioso.

2.5.1 Algoritmos de minería de datos

2.5.1.1 Support Vector Machine

El SVM es usado ya que es capaz de construir una función de decisión que da sólo un subconjunto de las características de entrada un peso no cero. De esta manera puede aprender cuáles son las palabras auditivas que son realmente discriminantes para los eventos de interés, e ignorar a los demás.

Considerando el set de entrenamiento $\left\{\left(\mathrm{x}_i, y_i\right)\right\}_{i=1}^{n}$ donde $x_i \in \mathbb{R}^p$ es el vector de features de entrada para la i-ésima muestra e $y_i \in \{1,-1\}$ es su correspondiente etiqueta indicando si la muestra es positiva $\left(y_i = +1\right)$ o negativa $\left(y_i = -1\right)$. Para empezar se asume que el set de datos positivos y negativos son linealmente separables, i.e., existe

una función $f(x) = \langle w, x \rangle + b$, donde $w \in \mathbb{R}^p$ (llamado el vector de peso) y $b \in \mathbb{R}$ (llamado bias) tales que [18]:

$$\langle w, x_i \rangle + b < 1 \quad \text{para } y_i = +1$$
$$\langle w, x_i \rangle + b > 1 \quad \text{para } y_i = -1$$

Posteriormente, el SVM busca los w y b del siguiente problema de optimización:

$$\underset{w \in \mathbb{R}^p, b \in \mathbb{R}}{\text{maximize}} \quad \underset{l \le i \le i}{\min} \quad \frac{\langle w, x_i \rangle + b}{\|w\|}, \quad s.t. \begin{cases} \langle w, x_i \rangle + b > 0 & \text{si } y_i = +1 \\ \langle w, x_i \rangle + b < 0 & \text{si } y_i = -1 \end{cases} \quad \forall \text{i}$$

2.5.1.2 K-Medias

También se utiliza K-Medias en este modelo ya que los sonidos propios de agresiones verbales son tan variados y pueden ocurrir de tantas formas diferentes que sería difícil asignarle una clase a un pequeño fragmento de audio del orden de pocos milisegundos. Por lo tanto se libera la labor humana de realizar este etiquetado realizando el algoritmo de clústering que realiza la tarea de clasificar segmentos de audio característicos (“aural words”) asociados a agresiones verbales.

Respecto al algoritmo, la idea básica es que dado un agrupamiento inicial no optimo, se traslada cada punto a su nuevo centroide más cercano, se actualiza los centroides de los clústeres calculando el promedio de los puntos de cada miembro y se repite el proceso de traslado y actualización hasta que se cumple un criterio de convergencia definido previamente [18].

Algorithm 1 *K*-means clustering algorithm

Input: K, number of clusters; D, a data set of N points
Output: A set of K clusters

1. Initialization.
2. **repeat**
3. **for** each point p in D **do**
4. find the nearest center and assign p to the corresponding cluster.
5. **end for**
6. update clusters by calculating new centers using mean of the members.
7. **until** stop-iteration criteria satisfied
8. **return** clustering result.

Figura 7: Algoritmo de K-medias [18]

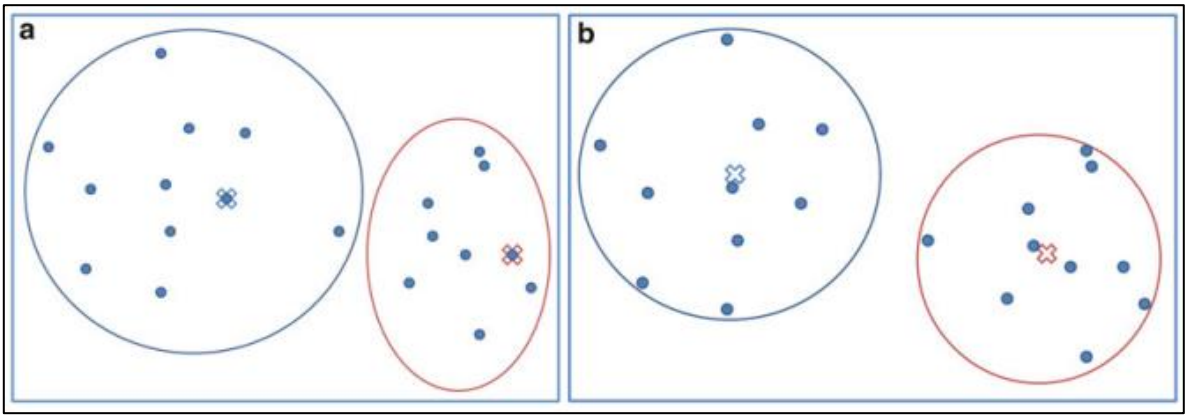


Figura 8: Ejemplo de clústering por K-Means (K=2). El centro de cada clúster está marcado por "x". (a) Iniciación. (b) Relocalización. [18]

3. Adquisición de datos

3.1 Antecedentes

Este es un capítulo importante de abordar ya que es la base para permitir desarrollar el trabajo propuesto y por lo tanto se explicará precavidamente lo trabajado.

Los datos requeridos son de agresiones verbales u conversaciones normales sin agresiones verbales manifestadas de manera audible. Como hay asuntos legales y éticos de por medio, esto dificulta tener voces reales de personas que en situaciones en que realmente experimentan emociones intensas como el llanto y la ira [34]. Se buscó sistemáticamente alguna fuente de datos abierta para acceder a estos datos, pero la búsqueda no fue fértil, no se encontraron fuentes de datos de audio de agresiones verbales. El resultado infructuoso de esta tarea se ve en parte justificado en el estudio de El Ayadi et al. [34], en cuyo segundo capítulo, dispone una tabla con las características de bases de datos comúnmente utilizadas en reconocimiento de emoción por voz (17 en total), en donde solo dos de las presentadas contaban con mediciones de ira intensa, que podrían haber servido para construir la base de datos, a saber, “Pereira” y “LDC Emotional Prosody Speech and Transcripts”, siendo aquellas de naturaleza privada y comercialmente disponible, respectivamente, pero a un costo elevado respecto al contexto en que se desarrolla este trabajo.

Además puede notarse que en todas las bases de datos mostradas en la tabla anteriormente citada, las emociones fueron articuladas artificialmente, excepto en las bases de datos “Natural” y “SUSAS” en que los datos fueron obtenidos de call centers y de estrés real respectivamente.

Por lo tanto al no poder contar con bases de datos preexistentes, se procedió a construir una base de datos propia. Al igual que varios trabajos anteriores como por ejemplo el de A Pikrakis et al. [35], De Santo, Massimo, et al. [36], Schuller, Björn, et al. [37] y Valenzise, Giuseppe, et al. [12], se procedió a construir la base de datos a partir de películas americanas y registros audiovisuales de internet.

La estructura de la base de datos construida puede ser vista con la clasificación que se muestra en la siguiente figura:

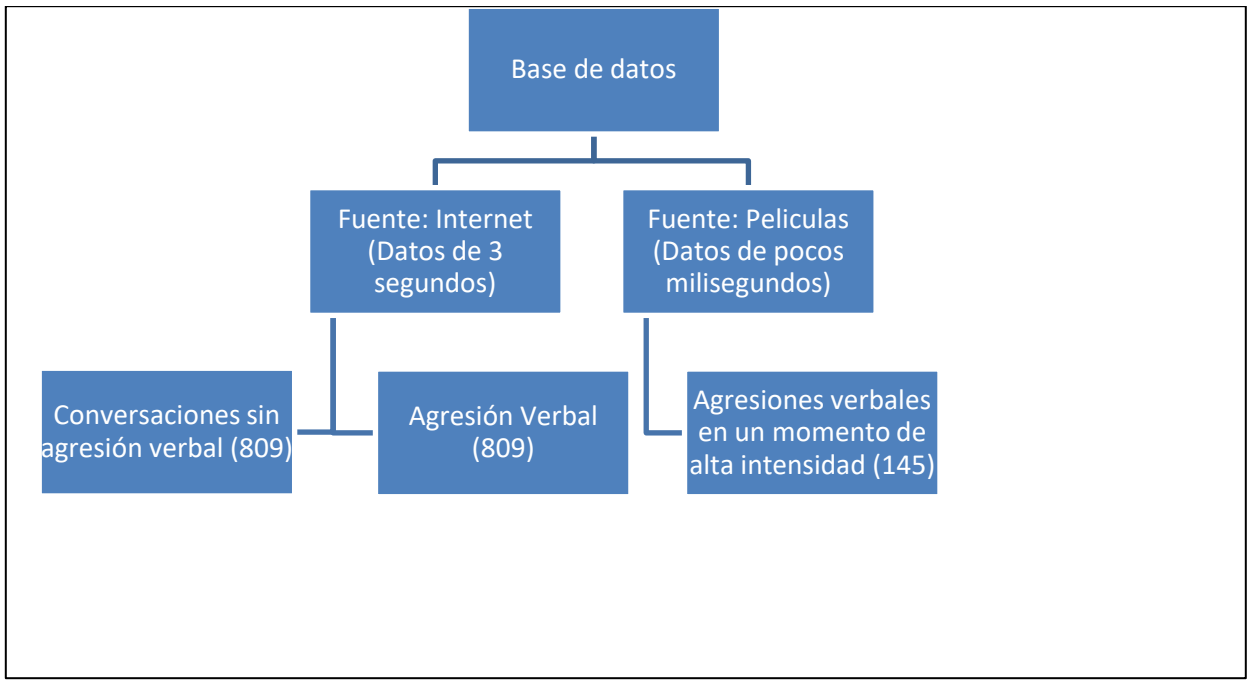


Ilustración 9: Estructura de la base de datos

Se requiere este set de datos adicional debido a la naturaleza del modelo de Carletti, Vincenzo, el cual consta de dos etapas: la primera de clusterización y la segunda de aplicación de SVM. Por lo cual, los datos de fuentes de películas solo serán usados exclusivamente para dicho modelo.

Por otro lado los datos de fuentes de internet serán usados para los demás algoritmos de machine learning que son ejecutados en esta memoria. Incluido el SVM de la segunda etapa del modelo de Carletti, Vincenzo

En definitiva, la cantidad de datos es 809 para agresión verbal y conversaciones sin agresión verbal y de 145 exclusivos para la primera etapa del modelo de Vincenzo

3.2 Software a utilizar

2.13.1 MediaInfo

MediaInfo es una visualización unificada práctica de los datos técnicos y de etiquetas más relevantes para archivos de vídeo y audio [14]. Este programa será utilizado para extraer los detalles de los archivos audiovisuales con los que se cuentan.

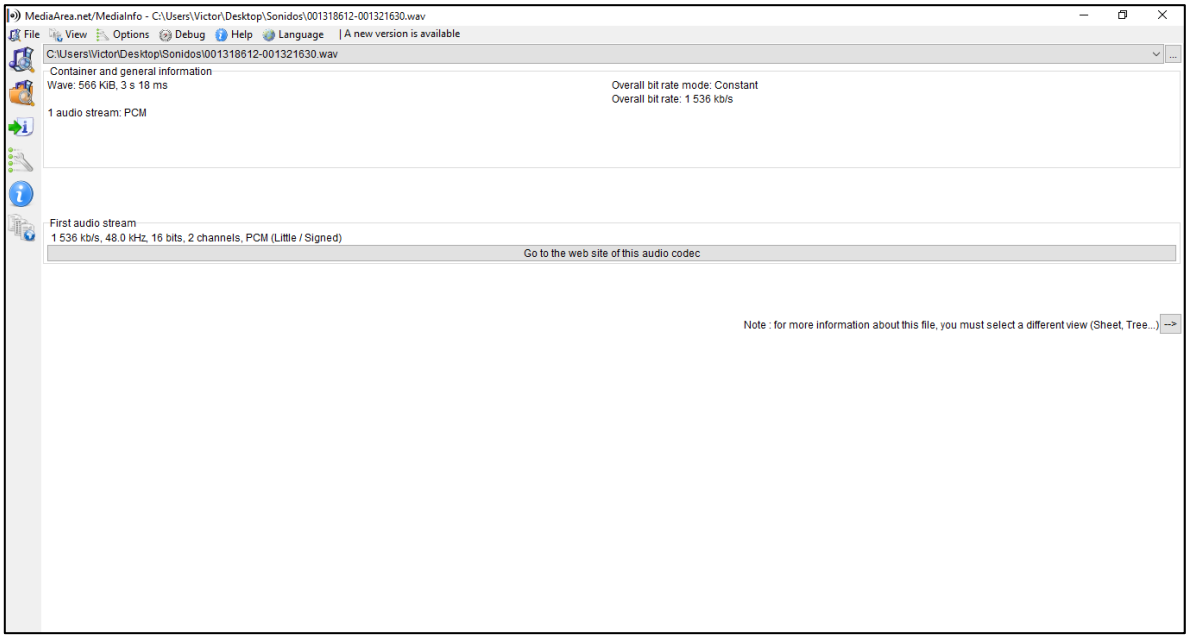


Figura 10: Interfaz MediaInfo

2.13.2 Audacity

Audacity es un software de audio de código abierto gratuito y multiplataforma para la grabación y edición múltiples pistas [15]. Este programa será utilizado para extraer el audio de un formato audiovisual, para transformar desde el formato AAC a WAVE (sin perdidas) y para exportar segmentos de audio.

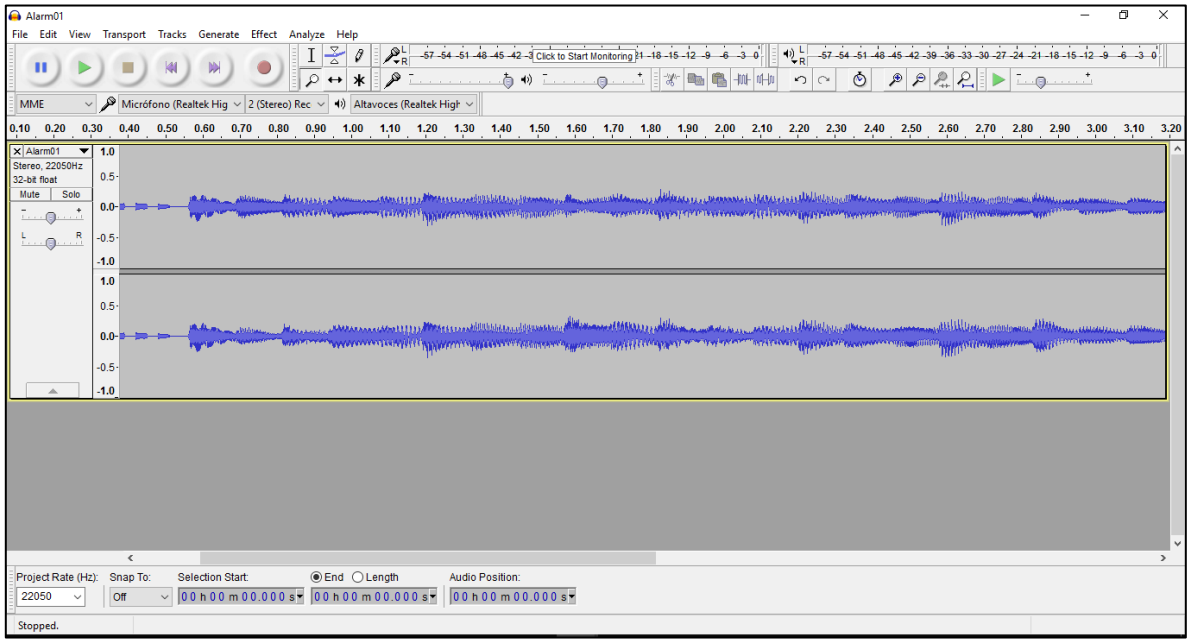


Figura 11: Interfaz Audacity. Fuente software Audacity

2.13.3 RStudio

RStudio desarrolla un software para el entorno de computación estadística R. R es un entorno de software libre para la informática estadística y gráficos. Compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS [16][51]. Este software que utiliza el lenguaje estadístico R, será fundamental para transformar y procesar todo el análisis de los archivos de audio. Lo único que se requiere es que los archivos de entrada se encuentren en formato WAVE.

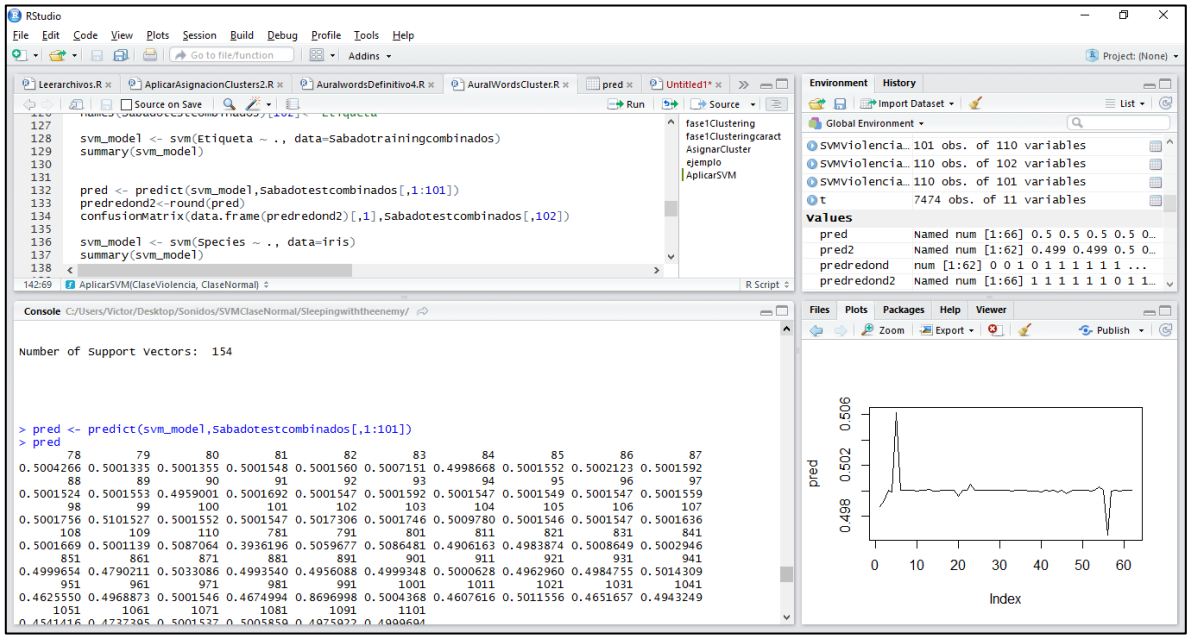


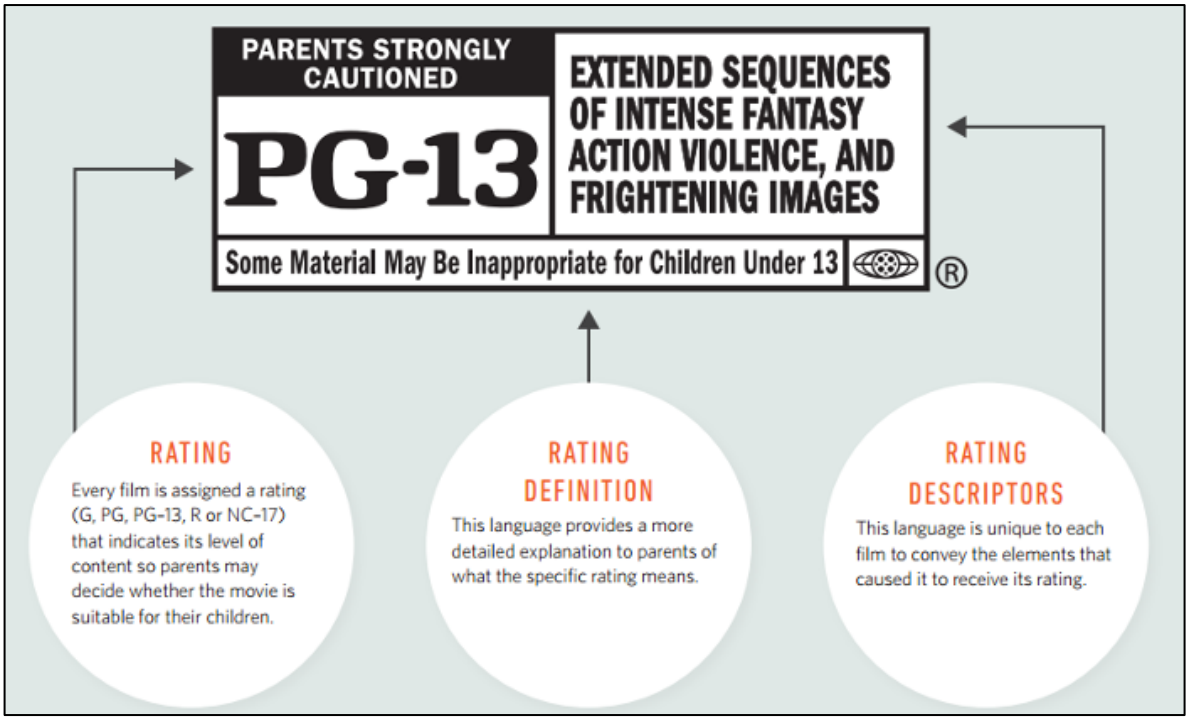
Figura 12: Interfaz RStudio. Fuente software RStudio

3.3 Extracción de datos

3.2.1 Datos originales

3.2.1.1 Datos para la primera etapa del modelo de Vincenzo Carletti

Las películas usadas fueron de rating “R” (Restricted) y PG-13 en la escala de la MPAA (“Motion Picture Association of America”) ya que en ellas puede existir contenido violento. El proceso de calificación consiste en un panel de padres que consideran factores tales como violencia, sexo, lenguaje y uso de drogas, luego asignan una calificación que creen que la mayoría de los padres estadounidenses darían a una película [38].



.

Figura 13: Rating de una película según la MPAA (Motion Picture Association of America). Extraida de [38]

En primera instancia, se contaba con una gran cantidad de archivos audiovisuales de películas. La primera tarea fue ejecutar una etapa de filtro para poder descartar películas con pocas posibilidades de contener agresiones verbales. Para hacer esto, se utilizó el motor de búsqueda online QuoDB [44], el cual busca palabras en una base de datos que contiene diálogos de millones de películas. Por lo cual se introdujeron palabras clave asociadas a la definición de agresión verbal señalada en la sección 1.3.1 de este trabajo.

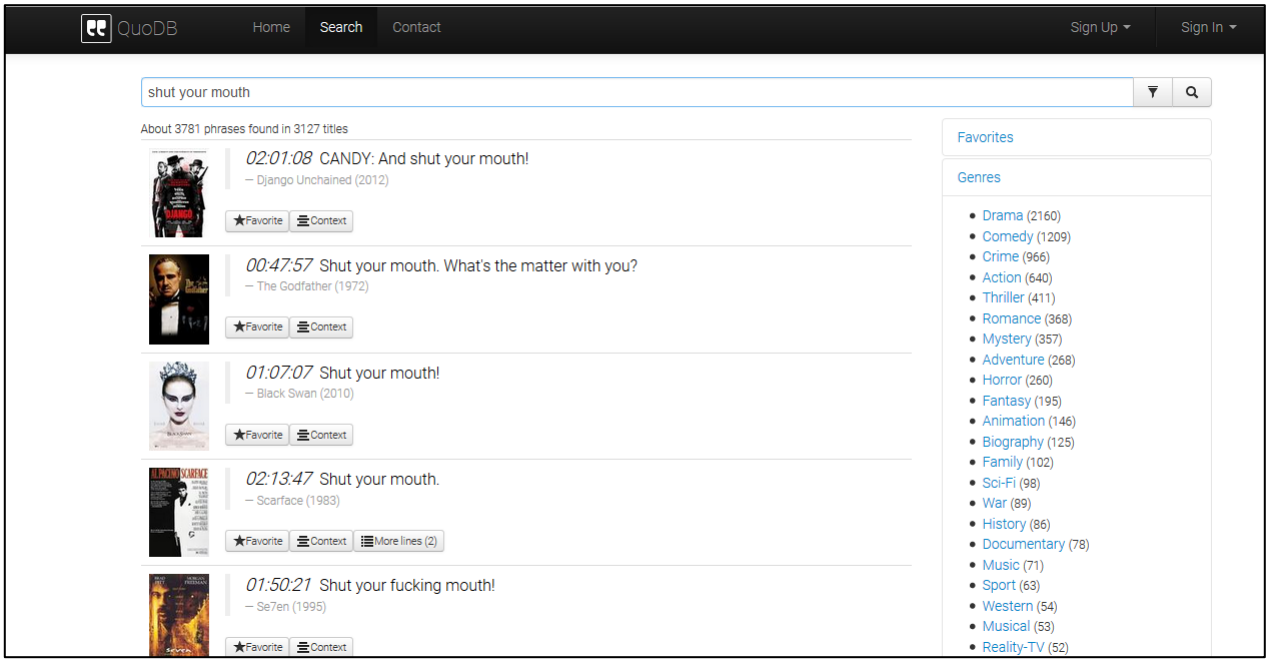


Ilustración 14: Interfaz del sitio web QuoDB y su interfaz tras realizar una búsqueda [44]

Además, junto con desplegar los resultados, el sitio web también muestra el tiempo en que ocurre dicho dialogo en las películas, lo cual permitió acelerar en cierta manera, el proceso de búsqueda de fragmentos de agresiones verbales.

Cabe hacer notar que no necesariamente los fragmentos de películas asociados a los diálogos, encontrados por este método, fueron útiles para ser incluidos a la base de datos, puesto que la escena podía contener excesivo ruido, o la agresión podía decirse en una voz muy baja, entre otras cosas. En definitiva los fragmentos siempre debieron ser vistos y filtrados manualmente para poder ser finalmente agregados a la base de datos

A pesar de la utilidad, el motor de búsqueda del sitio web, no funcionaba de una manera ideal y el tiempo en que mostraba la ocurrencia del dialogo, no calzaba con lo que realmente ocurría en una película. Esto probablemente se debe a la diferente cantidad de formatos de una misma película, o escenas censuradas en algunos países, entre otras causas. En definitiva, la regla general fue que el dialogo a menudo ocurría unos minutos antes o después de lo señalado, por lo cual fue necesario buscar manualmente entre esos intervalos.

Para acelerar algo más el proceso, se recurrió a los archivos de subtítulos de las películas, en donde se encuentran los diálogos con su marca temporal específicos para esa película. Por lo cual se buscaba el dialogo en el archivo de los subtítulos, se obtenía la marca temporal y se veía y escuchaba directamente el fragmento de audio

Dicho todo lo anterior, cabe hacer notar, que en una película que en general varía entre una hora y media y tres horas, en general, solo fueron encontrados pequeños fragmentos de agresión verbal, los cuales eran identificados manualmente, y

extraídos por el software Audacity, por lo que la recolección de datos fue un proceso muy lento y agotador.

En la Tabla 12, en anexos, se muestran algunas de las películas que fueron analizadas y sus características generales. Las que tienen sufijo “S” en primera instancia fueron pensadas para entrenar el SVM del segundo nivel del modelo de Vincenzo, pero finalmente fueron descartadas porque la expresión de agresiones verbales fueron muy sutiles según el criterio del autor. Por otro lado las películas con sufijo “E” fueron descartadas por dos razones: poseer una calificación MPAA G y PG que no cuentan con escenas de violencia ya que son aptas para ser vistas por niños o por no poseer el formato AAC.

De dicho conjunto de películas, solo 7 de las 8 con sufijo “C” fueron utilizadas para entrenar los clústeres en el primer nivel del modelo de Vincenzo

Nombre	Año de estreno	Calificación MPAA	Tasa de bits (kb/s)	Frecuencia (kHz)	Canales	Formato Audio
Pelicula2C	2008	PG-13	93.9	48	2	AAC(LC)
Pelicula3C	1994	R	64	22.05	2	AAC(HE-AAC/LC)
Pelicula4C	2002	PG-13	96	48	2	AAC(LC)
Pelicula5C	2006	R	94	44.1	2	AAC
Pelicula6C	2013	R	93.8	48	2	AAC(LC)
Pelicula7C	1997	PG-13	64	48	2	AAC(HE-AAC/LC)

Tabla 1: Películas finalmente seleccionadas

Se puede notar de la imagen anterior, que las frecuencias de todos los archivos son mayores a 22 kHz. Considerando que el contenido del habla humana de mayor frecuencia son las consonantes que se encuentran cerca de los 6 kHz, y además considerando que un grito femenino que es más agudo que el masculino tiene su máximo de energía cerca de los 1.5 kHz, según el trabajo de Begault, Durand R. [40], se puede afirmar que no existirán frecuencias importantes del habla humana que se pierdan en este formato AAC.

Con estas 6 películas se formaron 145 pequeños fragmentos en donde solo se seleccionaron intervalos del orden de decenas de milisegundos que capturaban un momento de alta intensidad de agresión verbal. Por ejemplo en la siguiente figura, se puede apreciar en plomo oscuro el intervalo seleccionado en un fragmento de agresión verbal.

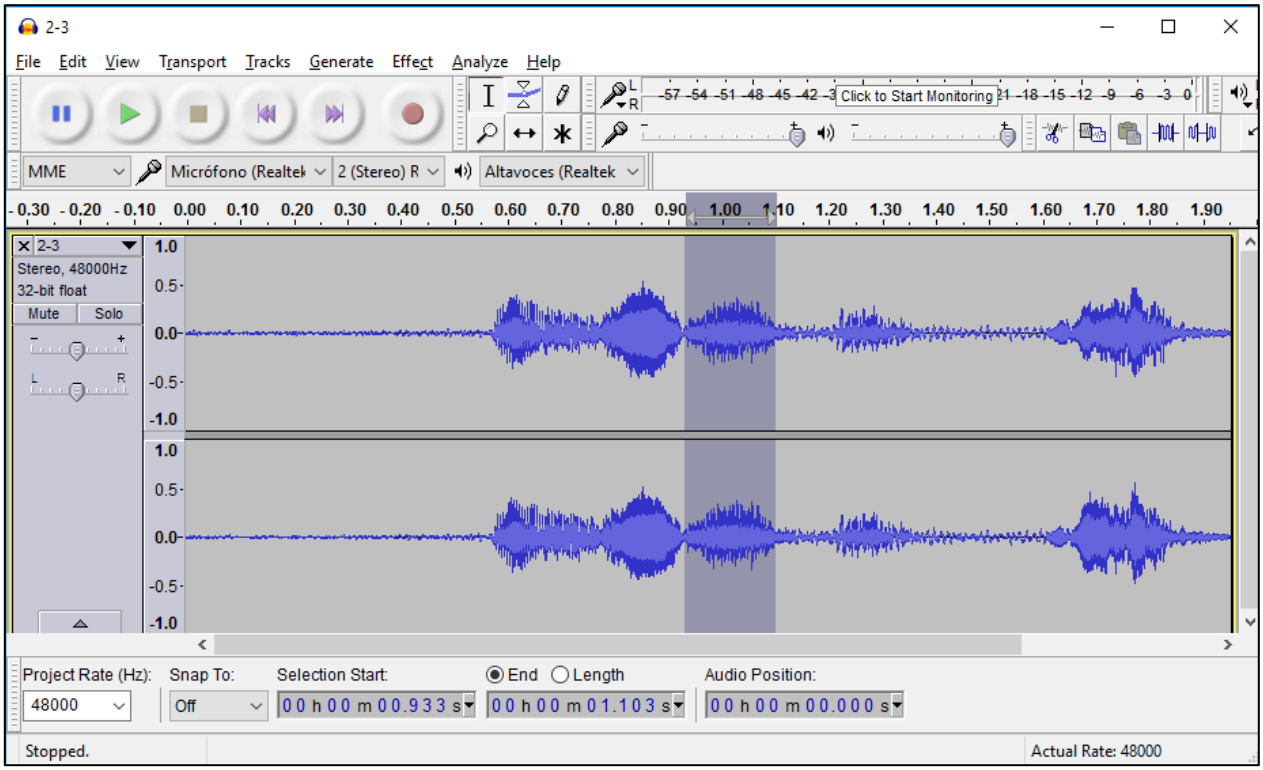


Figura 15 : Selección de un intervalo muy corto de audio para ser utilizado en la primera etapa del modelo de Carletti, Vincenzo. Fuente: Software Audacity

Con estos 145 fragmentos de audio son entrenados los clústeres de la primera etapa del modelo de Carletti, Vincenzo.

3.2.1.2 Datos para los algoritmos de machine learning

Debido a los problemas para extraer datos de la fuente películas donde un factor relevante era el gran gasto de tiempo, se optó por buscar registros audiovisuales en internet. Como estos registros en general duran pocos minutos y tienen un nombre que hace referencia al contenido, es más fácil encontrar y descartar registros. Se buscaron videos en diferentes motores de búsqueda usando palabras clave que se reflejan en la definición de agresión verbal establecida en el punto 1.3.1 de este trabajo.

Al encontrar un registro útil para cualquiera de las dos clases de la base de datos, se procedía a extraer el audio del video. Para esto se usó la herramienta OnlineVideoConverter [45], que permite extraer audio de registros audiovisuales de diversas fuentes.

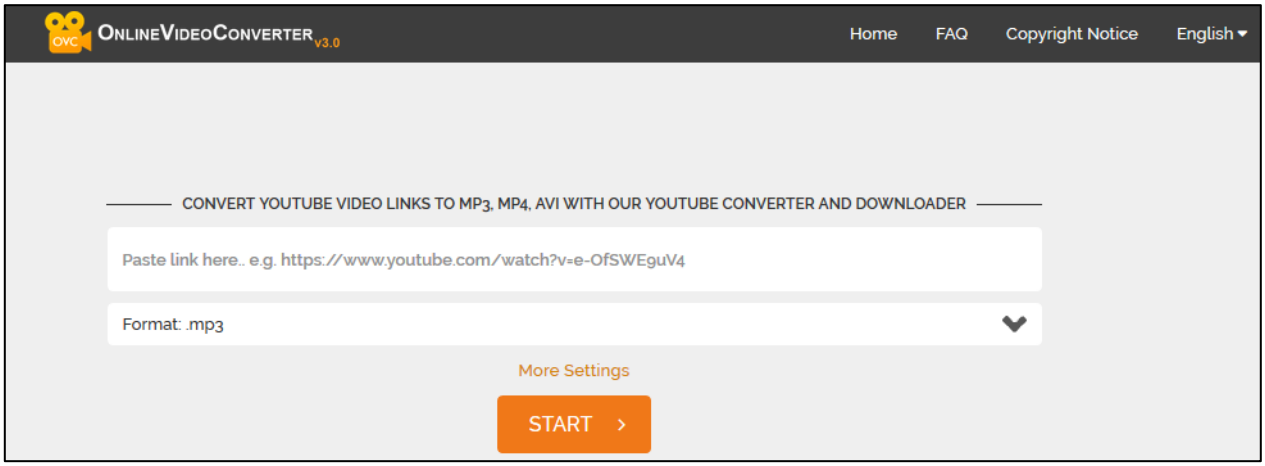


Ilustración 16: Interfaz de "OnlineVideoConverter", herramienta para extraer audio de variadas fuentes de registros audiovisuales de internet [45].

Para descargar los videos se seleccionó el formato wav. De esta manera se construyó la base datos. Se consideraron en primera instancia 31 videos con contenido de agresión verbal y 20 videos con contenido de conversaciones normales

Los datos para la clase "conversaciones normales", fueron obtenidos de registros audiovisuales de un matinal estadounidense. En este programa, se traen invitados, a los cuales se les hace una entrevista de manera relajada. Se eligió esta fuente porque hay largos intervalos de tiempo de permanente conversación. Como la situación es relajada, el entrevistador y el invitado pueden comportarse de manera seria o distendida según ellos estimen conveniente y pueden hablar fuerte, despacio, reír, etc. Es decir se tiene un espectro amplio del sonido que se emite al entablar una conversación. Además se consideraron entrevistas a actores ya que sus conversaciones giran en torno a sus trabajos en películas, y dado que este es un tema relacionado con la recreación, la naturaleza de las preguntas y la conversación en general es más de camaradería. Cabe mencionar que en algunos de estos registros, se trajo a un grupo de invitados, por lo cual se cubre aún más el espectro de posibles sonidos del habla en conversaciones.

Por otro lado, los datos para la clase "agresión verbal", fue el punto más dificultoso, dado que no es fácil encontrar tantos videos en que ocurra el caso. Se utilizó el método de buscar palabras clave relacionadas con la definición de agresión verbal del punto 1.3.1 del trabajo, en motores de búsqueda. Al encontrar un video, este era visto y si la manifestación de agresión verbal de manera audible se podía escuchar relativamente bien, entonces era incluido en la base de datos, si no, era desechado. Factores identificados que podían deteriorar el audio de la agresión verbal son: la presencia de ruido, saturación del micrófono, caídas del aparato con el cual se estaba grabando, etc.

3.2.2 Transformación de datos

3.2.2.1 Datos para la primera etapa del modelo de Vincenzo Carletti

El proceso para obtener los datos es el siguientes:

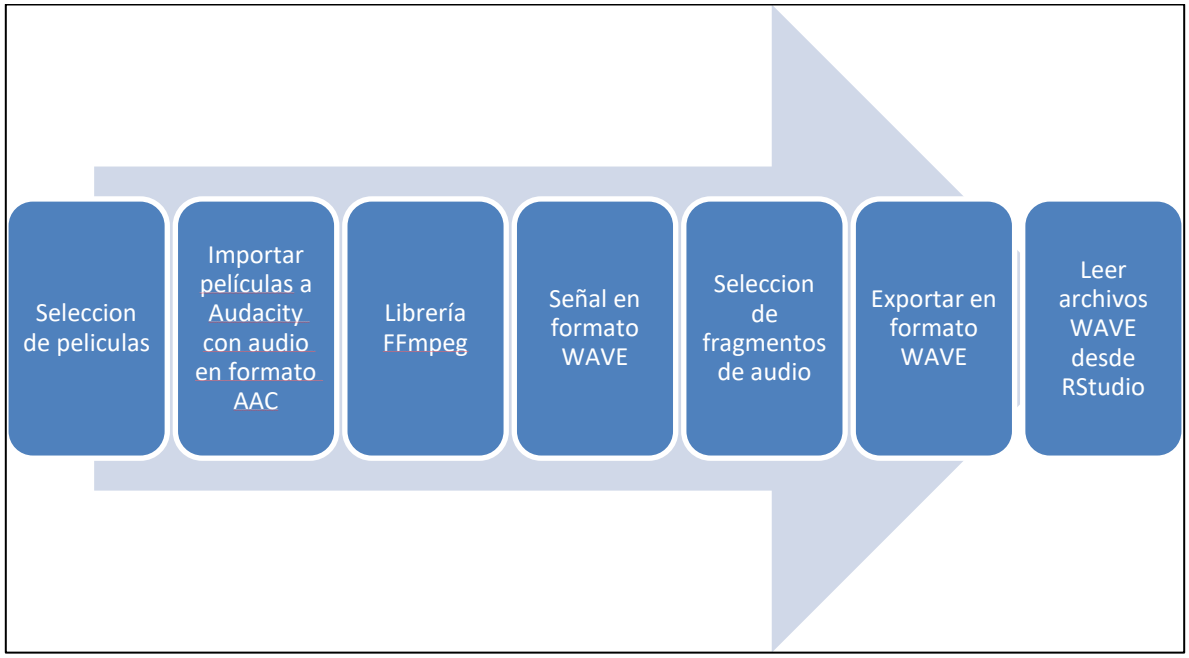


Figura 17: Diagrama de las acciones para procesar los datos

El primer paso consiste en seleccionar las películas. Entre los criterios descritos anteriormente, primero, se obtiene la información sobre la calificación según la MPAA de las películas y solo se consideran aquellas que estén calificadas con R o PG-13. Además se seleccionaron aquellas que estaban codificadas en formato AAC para mantener el mismo formato entre los datos. Se procesaron 6 películas en total.

El segundo paso consiste en importar las películas en el software Audacity, en donde se transforma el archivo audiovisual de formato mp4 a formato de audio WAVE a través de la librería FFmpeg.

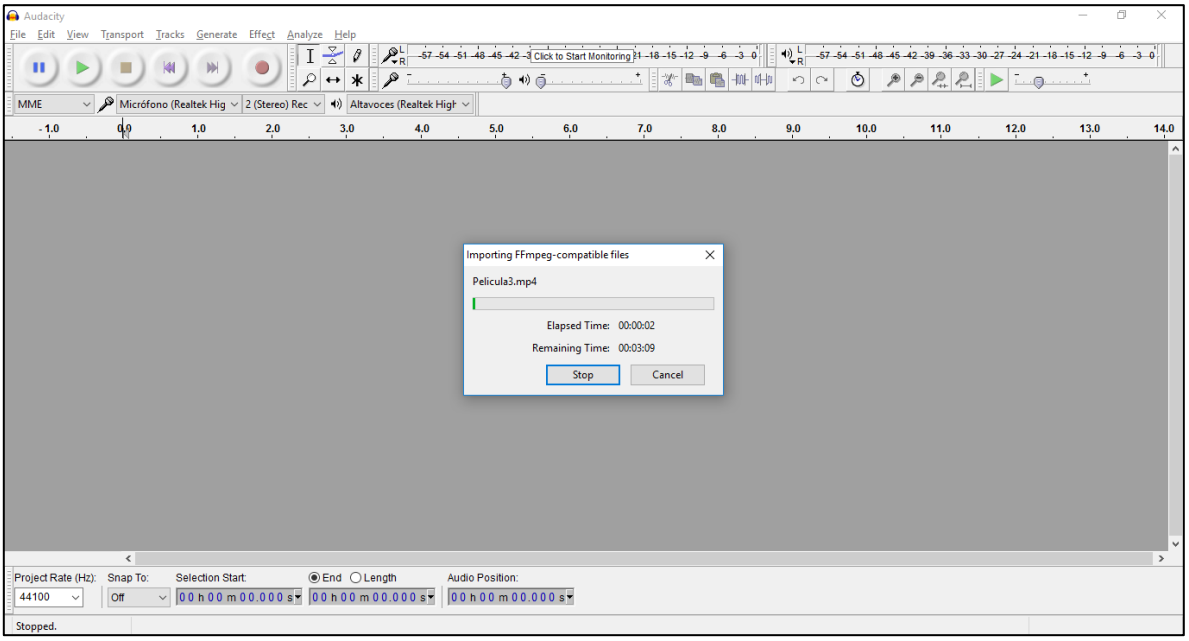


Figura 18: Pantalla al importar película en formato mp4 a Audacity

El tercer paso consiste en seleccionar y exportar fragmentos de audio en donde se encuentren fragmentos de violencia verbal. Los archivos se exportan con la misma frecuencia original en formato wav.

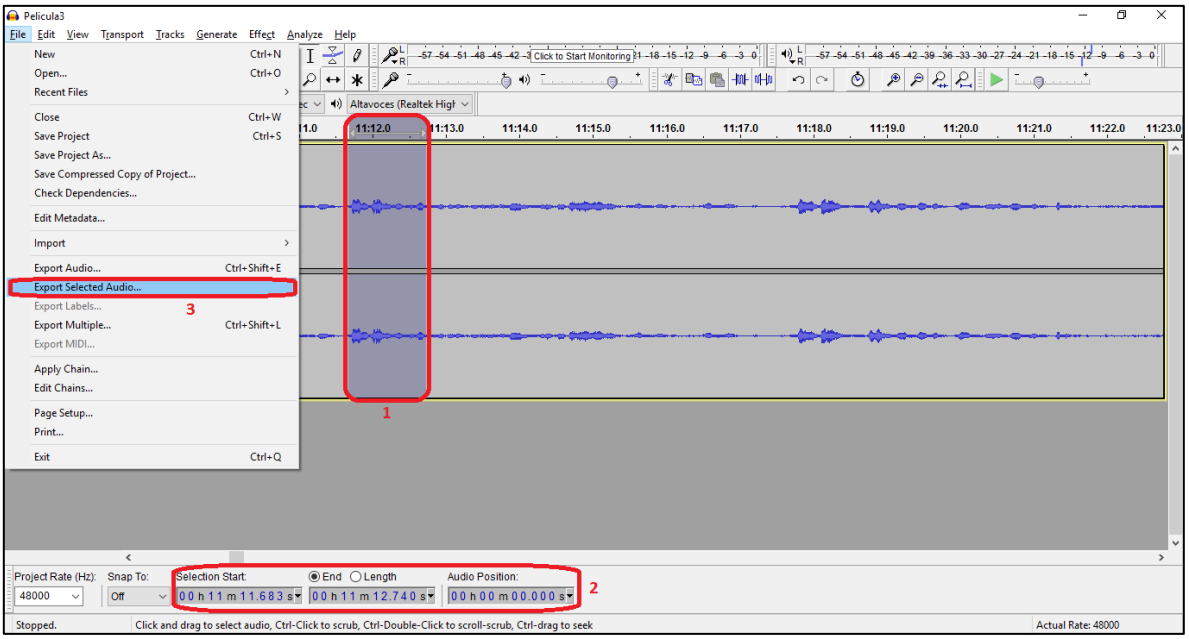


Figura 19: Pantalla para exportar fragmentos de audio. En 1 se puede seleccionar el audio con el cursos. En 2 se puede seleccionar el audio con el tablero. En 3 se selecciona el nombre y destino del archivo

Finalmente se leen los archivos del cuarto paso desde el software RStudio con la librería “TuneR” con la función “readWave()” para comenzar el análisis de las señales de audio. También son leídos a través del archivo ejecutable “SMILExtract_release.exe” de openSMILE que se revisará posteriormente.

Todos los archivos contaban con dos canales, en este trabajo solo se utilizó el canal izquierdo para simplificar el procesamiento.

3.2.2.2 Datos para los algoritmos de machine learning

El proceso para obtener los datos es el siguiente:

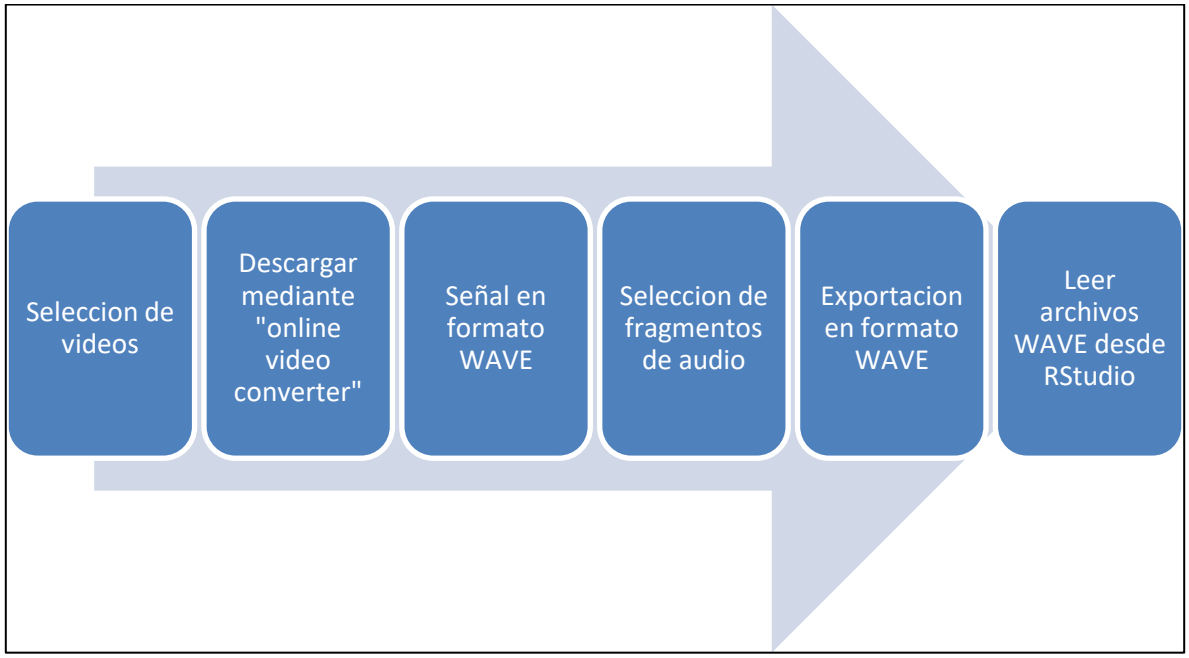


Ilustración 20: Diagrama de las acciones para procesar los datos

El primer paso consiste en buscar según palabras clave en motores de búsqueda de videos. Posteriormente al haber algún video con contenido de agresión verbal, se procedió a descargarlo mediante la herramienta “Online Video Converter” en formato wav.

Al tener un audio de varios minutos al extraer el audio de videos de internet, solía ocurrir que se mantenía una constante manifestación audible de las clases de interés que se necesitaban extraer. Es decir en el caso de agresiones verbales, existían videos en que se discutía permanentemente durante cierto intervalo de tiempo no menor. Y en el caso de conversaciones normales, es decir en el matinal, se conversaba continuamente durante varios segundos.

Para seguir con los demás pasos, era necesario, a priori, repetir el método de la figura 19, es decir digitar el tiempo de inicio, sumar 3.1 segundos y digitar el tiempo de fin y luego exportar el audio reiterativamente, lo cual se convierte en una tarea muy larga y monótona.

Para abordar este problema se recurrió a la opción de Audacity de agregar etiquetas cada cierto intervalo, como se aprecia en la siguiente figura.

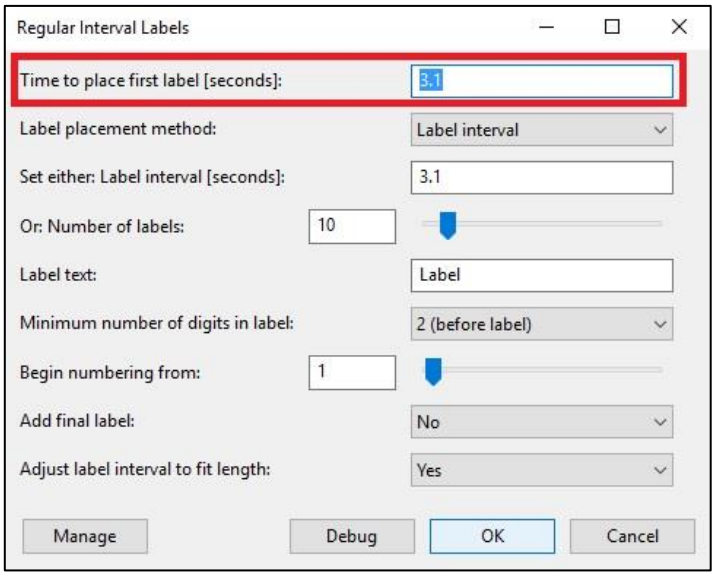


Figura 21: Opción para agregar etiquetas en software Audacity. Fuente: Software Audacity

Después de agregar las etiquetas se procede posteriormente a exportar los fragmentos de audio de la manera en que se muestra en la siguiente figura.

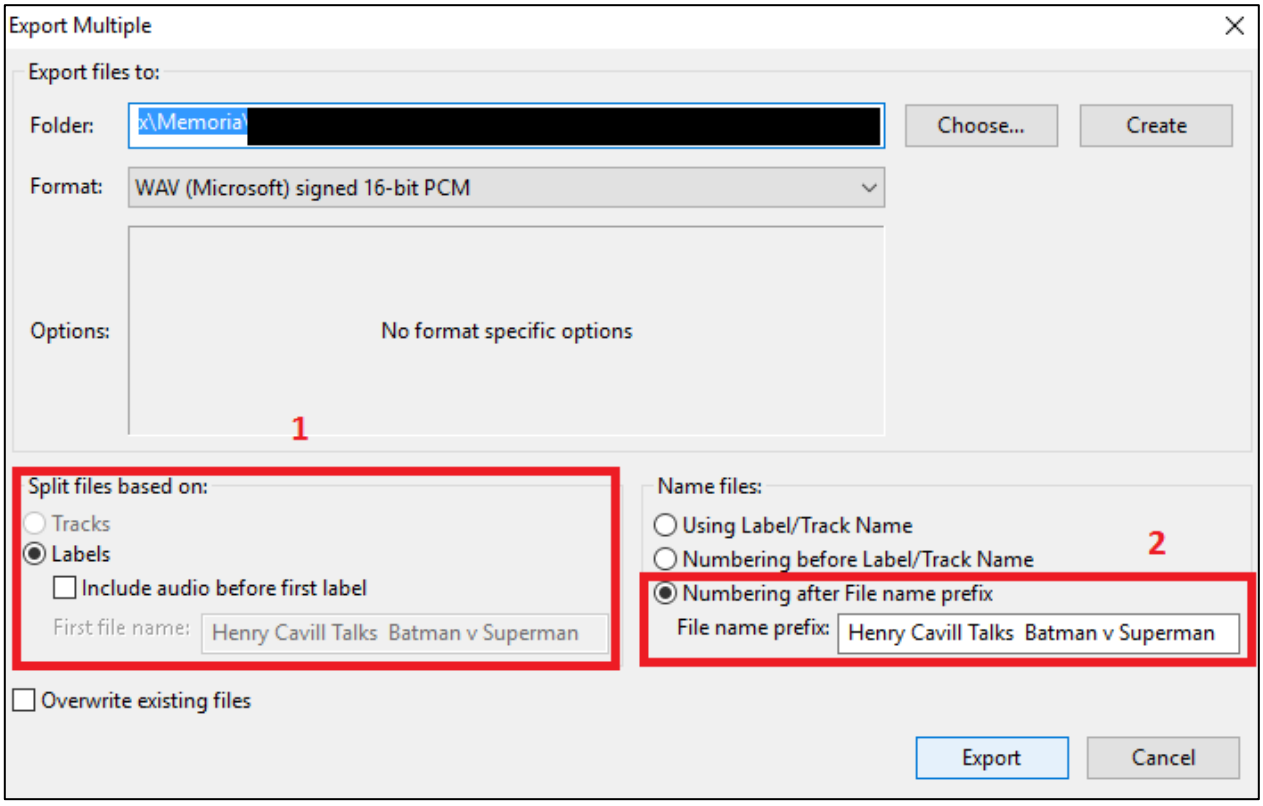


Figura 22: Exportar múltiples archivos de acuerdo a etiquetas en software Audacity. Los números 1 y 2 muestran las opciones seleccionadas en el presente trabajo. Fuente Software Audacity.

De esta manera se transformó cada archivo de audio que originalmente tenía una duración del orden de algunos minutos, a muchos archivos pequeños de audio de 3 segundos. Finalmente se creaba una lista de reproducción en el software VLC con los múltiples archivos de cada video, se escucharon y se escogieron si pertenecían a la clase de interés buscada.

En el caso de audios provenientes de videos con contenido de agresión verbal, era más dificultoso el proceso porque a menudo, solo una pequeña fracción del video contenía la clase agresión. En el caso de los audios provenientes del matinal, el proceso fue mucho más sencillo, dado que la entrevista duraba varios minutos, por lo cual se podían seleccionar estos audios manualmente de manera mas cómoda

Es necesario realizar todo este procedimiento ya que la agresión verbal es una reacción impulsiva y puede durar corto tiempo, por lo que si se consideran ventanas de tiempo más largas, es posible que la reacción pase inadvertida.

4 Aplicación de la metodología

En el presente trabajo se utilizaron 4 modelos, los cuales se enumeran a continuación

- Modelo de Carletti, Vincenzo con features originales. Primera etapa con clustering y segunda etapa con SVM aplicado a las clases “agresión verbal” y “conversaciones normales”
- Modelo de Carletti, Vincenzo con features de openSMILE. Primera etapa con clustering y segunda etapa con SVM aplicado a las clases “agresión verbal” y “conversaciones normales”
- SVM aplicado a las clases “agresión verbal” y “conversaciones normales”
- Regresión lineal aplicado a las clases “agresión verbal” y “conversaciones normales”

4.1 Consideraciones preliminares

A pesar de que según cada frecuencia de muestreo existe una cantidad de muestras determinadas según la resolución de frecuencia que se desea mantener constante (ver ecuación (12) en glosario), esto produce que cada fragmento tenga la misma duración en tiempo que otro fragmento proveniente de otra frecuencia al mantener constante la resolución de frecuencia, ya que la proporción se mantiene.

Por ejemplo comparando un archivo cuya frecuencia es de 48000 Hz con otro que es de 44000 Hz, se tiene:

$$\frac{48000}{x} = 31.25 \Rightarrow x = 1536$$

$$\frac{44000}{x} = 31.25 \Rightarrow x = 1408$$

$$\frac{1536}{48000} = \frac{1408}{44000} = 0.032[\text{s}]$$

Es decir, cada fragmento de audio a pesar de que tiene distinta cantidad de puntos (1536 o 1408 en el ejemplo) según su frecuencia, tiene la misma duración en tiempo, por lo tanto, se necesitan la misma cantidad de estos fragmentos para recrear un segundo de audio.

Si se considera que cada fragmento de audio dura 0.032[s] y que como hay 0.75% de traslape cada fragmento adicional, añadirá 0.032/4=0.008 segundos. Se tiene:

$$0.032 + \frac{0.032}{4} \cdot x = 3 \Rightarrow x = 372$$

Es decir que 3 segundos de datos se construirán con 372 fragmentos independientemente de su frecuencia C.

4.2 Modelo de Carletti, Vincenzo con features originales

4.2.1 Features

Se obtiene cada set de features: Centroide Espectral, Desviación Estándar, Rotación Espectral, Flujo Espectral, Ratio de energía en Sub bandas (4 rangos), Volumen, Energía y Tasa de cruces por cero. Se usó un traslape de 0.75 al igual que el trabajo de Carletti, Vincenzo, et al. [7]. Además se utiliza una resolución de frecuencia de 31.25 al igual que el trabajo original, esto determina el largo del segmento (o fragmento en este trabajo) al cual se le extraen features (la cantidad de puntos), según la frecuencia de muestreo del archivo de audio.

4.2.2 Primera etapa – Clústeres

El procesamiento general de esta etapa se ejemplifica en la siguiente figura:

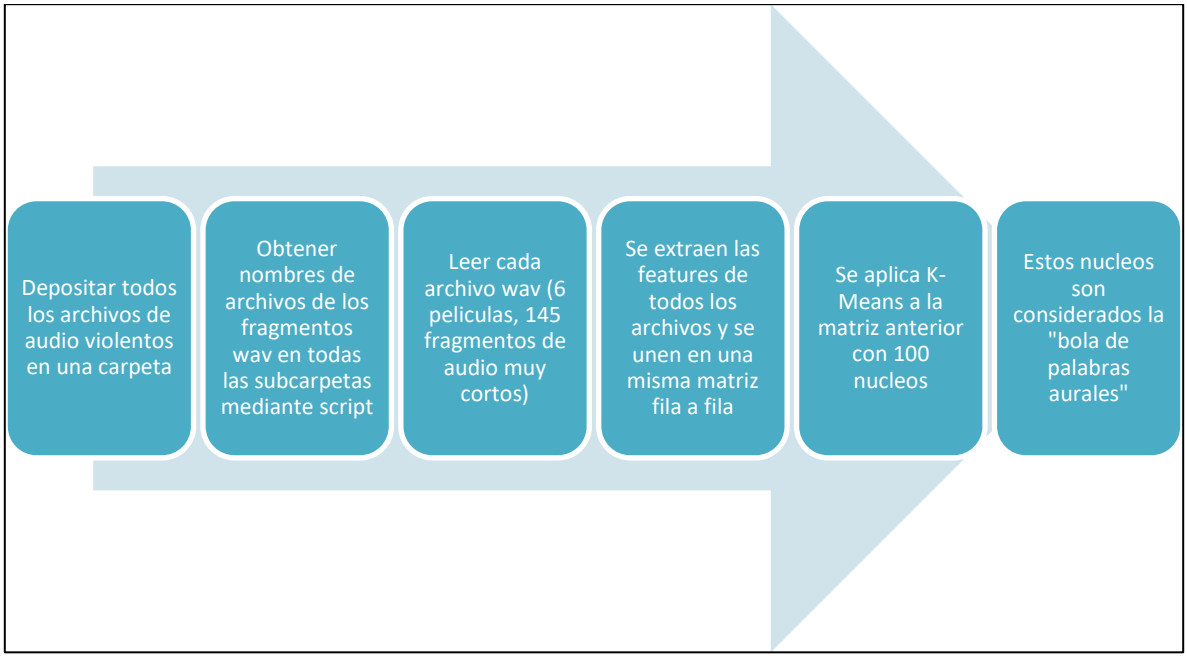


Figura 23 Pasos generales para el entrenamiento del primer nivel del algoritmo

En primer lugar se deben entrenar los 100 clústeres del primer nivel del modelo con archivos de audio muy cortos provenientes de la fuente de datos “películas”, los cuales contienen agresión verbal a lo largo de toda su corta duración. El resultado de esta proceso es una matriz de 100 filas y 11 columnas (las 11 features) a la cual se le llama “codebook” que es la llamada “bolsa de palabras aurales” y que contiene en cada fila las features de cada núcleo.

Estos clústeres servirán para generar nuevas features en el segundo nivel del modelo para el clasificador SVM.

4.2.3 Etapa intermedia

La matriz “codebook” descrita anteriormente contiene 100 núcleos. En esta etapa se procesan los 809 audios de agresión verbal y los 809 audios de conversaciones sin agresión verbal, extraídos de internet. Cada uno de estos 1618 audios se divide en 372 unidades iguales, a cada unidad se le calculan las 11 features y se le aplica un algoritmo para asignar el clúster más cercano de la matriz “codebook”, basado en la distancia entre sus features. En definitiva cada audio de 3 segundos queda fragmentado en 372 unidades pequeñas que son asignadas a cada uno de los 100 clústeres. En la siguiente tabla se puede apreciar la estructura de los datos que se obtiene, aplicando lo anterior:

Unidad 1	Unidad 2	Unidad 3	…	…	…	Unidad 372	Etiqueta
Clúster 1	Clúster 30	Clúster 90				Clúster 40	“Violencia”

Tabla 2: Ejemplo de estructura de datos de fragmento de 3 segundos, después de aplicar el algoritmo del clúster más cercano a cada una de sus 372 unidades. Esta fila de información corresponde a sólo 1 archivo de audio de 3 segundos

Finalmente, la última sub etapa, es aplicar un histograma a los 100 clústeres de la tabla anterior, obteniendo lo que se muestra en la siguiente tabla:

Clúster 1	Clúster 2	Clúster 3	…	…	…	Clúster 100	Etiqueta
50	20	11				0	“Violencia”

Tabla 3: Ejemplo de formato de datos de entrada para SVM. Esta fila de información corresponde a solo 1 archivo de audio de 3 segundos

Se puede apreciar en a tabla anterior que cada archivo de audio de 3 segundos dispone de una estructura lista para ser procesada con el algoritmo de SVM, dado que cuenta con las features y su correspondiente etiqueta.

A continuación se muestra un histograma generado en una iteración de este trabajo, en el paso 6 descrito anteriormente.

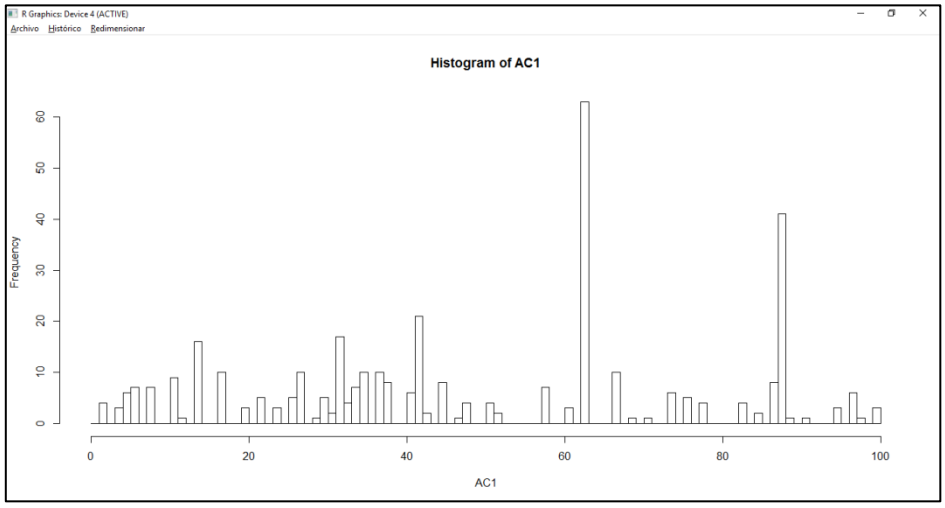


Figura 24: Histograma de un archivo de 3 segundos. En general cada sonido tiene un histograma particular en base al “codebook” construido y el SVM de la segunda etapa debe encargarse de la clasificación. Se espera que las dos clases tengan histogramas característicos para que el SVM pueda discriminar correctamente

4.2.4 Segunda etapa – SVM

Fue utilizada la librería “e1071” de Rstudio aplicando la función svm() del paquete

El resultado del SVM se redondeó a 0 o a 1, según el caso y en base a aquello se le asignó la clase de violencia o conversación normal.

4.3 Modelo de Carletti, Vincenzo con features de openSMILE

El procesamiento es exactamente igual al anterior, solo varia la extracción de features

En este caso fue utilizada la herramienta de extracción de features openSMILE, que permite extraer una amplia cantidad de features de audio en tiempo real [47]. La herramienta puede ser ejecutada a través del Simbolo de Sistema de Windows a través de líneas de comando. Un ejemplo de su funcionamiento puede apreciarse en la siguiente figura:

Figura 25: Ejecución de OpenSmile para obtener features del audio “Alarm01.wav” dadas ciertas opciones.

Con la línea de código de la imagen anterior se ejecuta SMILExtract_release.exe. La opción –C especifica la configuración que se utilizará, -I especifica el archivo de audio para el input, -csvoutput indica el nombre del archivo de salida en formato csv, -start indica desde que segundo del audio comienza la extracción de features y –end indica hasta qué segundo se desea aplicar la extracción de features

La herramienta viene con archivos de configuración preliminares que han sido utilizados para participar en ciertas competencias internacionales [48]. Cada

configuración difiere en ciertos parámetros, como por ejemplo: la selección de funcionales (e.g. media, desviación estándar, kurtosis, etc), la selección de descriptores de bajo nivel (e.g. frecuencia fundamental, tono, tasa de cruces por cero, etc), ventanas, traslape de ventanas, etc. Cabe señalar que es posible crear nuevos archivos de configuración si un usuario así lo desea.

En este trabajo se utilizó la configuración “MediaEval2012 TUM”, que fue utilizada para la detección de escenas violentas en películas populares de Hollywood [48] en la competencia Mediaeval [49]. En esa competencia fueron evaluadas escenas incluyendo su contenido visual además del audio. Para su entrenamiento los participantes contaron con 24 películas etiquetadas por los organizadores según presencia de violencia en intervalos de cuadros (frames en inglés), según siete conceptos visuales y 3 auditivos (“presencia de gritos”, “disparos” y explosiones). Por dichas razones se escoge esta configuración [50].

Se modificó la duración de la ventana original estableciéndola en 0.1 segundos, y el largo del paso se dejó en 0.032 ms, al igual que en el trabajo de Carletti, Vincenzo. Aparte de las features, los pasos son idénticos al punto 4.2

4.3.1 Features

Se modificó la configuración original de “MediaEval2012 TUM”, dejando solo los funcionales “Media Aritmética” y “Media aritmética de valores absolutos”. Además de aplicar los dos funcionales descritos anteriormente, también se le aplica la derivada a cada uno de los 37 descriptores de bajo nivel, obteniendo 148 features ($37 \cdot 4$features). Estos descriptores consideran:

- Coeficientes espectrales de frecuencia de Mel [1-16]
- Tasa de cruces por cero
- Logaritmo de la energía
- Espectrograma auditivo
- Espectrograma auditivo filtrado con Transformada Espectral Relativa (RASTA por sus siglas en inglés)
- Rotación espectral [25,50,75,90]
- Flujo espectral
- Centroide espectral
- Entropía espectral
- Varianza espectral
- Coeficiente de asimetría espectral
- Curtosis espectral
- Pendiente espectral
- Agudeza psicoacústica
- Armónicas
- Ratio de energía en sub-bandas[150-650-4000-15000]

4.4 SVM aplicado a las clases “agresión verbal” y “conversaciones normales”

Las features usadas fueron las extraidas a través del software Opensmile descritas en la sección 4.3.1.

Se tienen dos grandes grupos de 809 audio de cada clase, están etiquetados y con las 148 features descritas anteriormente extraídas con openSMILE

Es utilizada la librería “e1071” de Rstudio aplicando la función svm() del paquete. El resultado del SVM se redondeó a 0 o a 1, según el caso y en base a aquello se le asignó la clase de violencia o conversación normal.

4.5 Regresión lineal aplicada a las clases “agresión verbal” y “conversaciones normales”

Las features usadas fueron las extraídas a través del software openSMILE descritas en la sección 4.3.1.

Se tienen dos grandes grupos de 809 audio de cada clase, están etiquetados y con las 148 features descritas anteriormente extraídas con openSMILE

Es utilizada la función lm() de RStudio. El resultado de la regresión se redondeó a 0 o a 1, según el caso y en base a aquello se le asignó la clase de violencia o conversación normal.

5. Análisis y resultados

5.1 Análisis estadístico

Con el objetivo de determinar si las features extraídas desde las señales de audio son capaces de discriminar entre la clase de agresión verbal y no agresión se pretende realizar un test de hipótesis entre las dos clases independientes.

Para realizar esto se crea un conjunto de datos adicional para contrastar la agresión verbal proveniente de películas. Es decir, se extraen 145 nuevos fragmentos de audio sin agresión verbal, para poder comparar.

En primer lugar, se realiza la prueba de Shapiro cuya hipótesis nula asume que las clases comparadas poseen distribución independiente con el fin de determinar la normalidad de cada set de datos. Se tiene un set de datos, el que se determinan los clústeres.

Para recordar las 11 features se numeran de la siguiente manera: 1) Centroide espectral, 2) Desviación estándar, 3) Flujo espectral, 4) Ratio de energía en sub-bandas [1-630]Hz, 5) Ratio de energía en sub-bandas [631-1720]Hz, 6) Ratio de energía en sub-bandas [1721-4400]Hz, 7) Ratio de energía en sub-bandas [4401-22000]Hz, 8) Volumen, 9) Energía y 10) Tasa de cruces por cero.

Los resultados de la prueba Shapiro son los siguientes:

Features Clústeres

Feature	1	2	3	4	5	6	7	8	9	10	11
W	0.956	0.966	0.971	0.912	0.865	0.945	0.951	0.938	0.969	0.921	0.964
p-value	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16

Como se aprecia para cada feature, se rechaza la hipótesis nula. Esto implica que la data analizada no tiene distribución normal con lo que no es posible aplicar un test de hipótesis t.

Debido a lo anterior, es posible aplicar el test de wilcoxon considerando dos variables nominales y una variable de medida, el factor (ausencia o presencia de agresión) y el registro correspondiente a los archivos extraídos desde películas.

Las hipótesis del test indican lo siguiente:

- $H_o: \mu_1 - \mu_2 = 0$; No existe diferencia significativa entre las medianas de la clase no violencia y la clase violencia
- $H_1: \mu_1 - \mu_2 > 0$; Existe diferencia significativa entre las medianas de la clase no violencia y la clase violencia

A continuación, se muestran los resultados del test:

Feature	1	2	3	4	5	6	7	8	9	10	11
W[1]	152220	205770	195340	24776	288220	47793	86229	166590	21422	20477	122000
p-value	1.3e-10	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	0.018	<2e-16	<2e-16	<2.2e-16

Por lo tanto, se desprende de los resultados que en cada caso se rechaza la hipótesis nula con lo que las features extraídas pueden discriminar entre las dos clases planteadas.

5.2 Resultados modelo de Carletti, Vincenzo con features originales

Se aplicó la regla 70/30 para el entrenamiento y el testeo. Los resultados son los siguientes:

		Referencia	
		0	1
Prediccion	0	226.5	47.03
	1	16.5	195,97

Tabla 4: Matriz de confusión del modelo de Carletti, Vincenzo con features originales. Se iteró 100 veces

Accuracy	86.27%	Sensitivity	80.64%	Recall	80.64%
Kappa	73.85%	Specificity	93.20%	F1	86.03%
AccuracyLower	83.60%	Pos Pred Value	92.26%	Prevalence	50.0%
AccuracyUpper	89.79%	Neg Pred Value	82.84%	Detection Rate	40.32%
AccuracyNull	50%	Precision	92.26%	Detection Prevalence	43.71%
AccuracyPValue	0			Balanced Accuracy	86.92%
McnemarPValue	37.04%				

Tabla 5: Resultados a partir de la matriz de confusión de la tabla 4

[1] Cada valor fue dividido por 100

5.3 Resultados modelo de Carletti, Vincenzo con features de openSMILE

Se aplicó la regla 70/30 para el entrenamiento y el testeo. Los resultados son los siguientes:

		Referencia	
		0	1
Prediccion	0	190.53	48.59
	1	51.47	193,41

Tabla 6: Matriz de confusión del modelo de Carletti, Vincenzo con features extraídas con openSMILE. Se iteró 100 veces

Accuracy	79.32%	Sensitivity	79.92%	Recall	79.92%
Kappa	58.65%	Specificity	78.73%	F1	79.44%
AccuracyLower	75.44%	Pos Pred Value	79.03%	Prevalence	50.0%
AccuracyUpper	82.84%	Neg Pred Value	79.72%	Detection Rate	39.96%
AccuracyNull	50%	Precision	79.03%	Detection Prevalence	50.59%
AccuracyPValue	0			Balanced Accuracy	79.32%
McnemarPValue	54.25%				

Tabla 7: Resultados a partir de la matriz de confusión de la tabla 9

5.4 Resultados SVM aplicado a las clases “agresión verbal” y “conversaciones normales”

Se aplicó la regla 70/30 para el entrenamiento y el testeo. Los resultados son los siguientes:

		Referencia	
Prediccion	0	238.69	4.45
	1	4.31	238.55

Tabla 8: Matriz de confusión del algoritmo SVM aplicado en las dos clases “Agresión verbal” y “Conversaciones normales”. Se iteró 100 veces

Accuracy	98.19%	Sensitivity	98.16%	Recall	98.16%
Kappa	96.39%	Specificity	98.22%	F1	98.19%
AccuracyLower	96.58%	Pos Pred Value	98.23%	Prevalence	50.0%
AccuracyUpper	99.17%	Neg Pred Value	98.17%	Detection Rate	49.08%
AccuracyNull	50%	Precision	98.23%	Detection Prevalence	49.97%
AccuracyPValue	0			Balanced Accuracy	98.19%
McnemarPValue	68.85%				

Tabla 9: Resultados a partir de la matriz de confusión de la tabla 11

5.5 Resultados regresión lineal aplicada a las clases “agresión verbal” y “conversaciones normales”

Se aplicó la regla 70/30 para el entrenamiento y el testeo. Los resultados son los siguientes:

		Referencia	
		0	1
Prediccion	0	238.72	6.69
	1	4.28	236.31

Tabla 10: Matriz de confusión del algoritmo regresión lineal aplicado en las dos clases “Agresión verbal” y “Conversaciones normales. Se itero 100 veces

Accuracy	97.74%	Sensitivity	97.24%	Recall	97.24%
Kappa	95.48%	Specificity	98.23%	F1	97.73%
AccuracyLower	96.00%	Pos Pred Value	98.22%	Prevalence	50.0%
AccuracyUpper	98.85%	Neg Pred Value	97.28%	Detection Rate	48.62%
AccuracyNull	50%	Precision	98.22%	Detection Prevalence	49.51%
AccuracyPValue	0			Balanced Accuracy	97.74%
McnemarPValue	52.78%				

Tabla 11: Resultados a partir de la matriz de confusión de la tabla 13

6. Discusión

Lo primero que se puede apreciar fue la alta exactitud de los modelos construidos: Especificamente: Modelo de Carletti con features originales 86.27%, Modelo de Carletti con features de openSMILE 79.32%, SVM 98.19% y Regresión lineal 97.74%

El modelo presenta una buena sensibilidad lo que es muy favorable ya que en el marco de este trabajo en que la salud está en juego, es deseable que no se pasen por alto verdaderos positivos.

Vincenzo tuvo datos suficientes en su trabajo como para poder calcular 1024 clústeres que simbolizaban su bolsa de palabras. En este trabajo solo se usaron 100 por la poca cantidad de datos. El modelo eventualmente tendría mejores resultados si se aumentara el número de clústeres, pero se necesitaría entrenarlo con mayor cantidad de archivos de audio.

La hipótesis de investigación se considera cumplida con los datos usados en este trabajo, dado que se obtienen exactitudes superiores al 79% con cada uno de los 4 algoritmos.

Una razón de las exactitudes altas puede ser la fuente de datos original, una clase (“la de conversaciones normales”) fue extraída de un solo matinal estadounidense. En cambio, los audios de la clase “agresiones verbales” fueron extraídos de multiples fuentes. Esta diferencia puede haber ayudado al model a clasificar mejor las clases

Los resultados de SVM y Regresión Lineal aplicado a ambas clases producen una exactitud, similar a las que otros autores han logrado con otras clases de interés por ejemplo Vincenzo logro una exactitud de 95.8% al diferenciar 4 clases. Por otro lado van Hengel, Peter WJ, y Tjeerd C. Andringa lograron una sensibilidad de 100% de detección de gritos en ambientes no controlados. Tambien Foggia, Pasquale, et al., logró una exactitud de 95.89% al detectar 4 clases. Por lo tanto el resultado obtenido en este trabajo puede ser genuinamente alto, dado que en la literatura se han logrado altas exactitudes.

Sin embargo el modelo de Carletti, obtuvo una menor exactitud al ser replicado en este trabajo, la exactitud es buena (86.27%), pero dista del 95.8% de exactitud reportado en su trabajo. Una de las razones de esta baja en la exactitud es muy posible que sea la reducción de clústeres en la primera etapa de su modelo, ya que él ocupo más de 1000 clústeres, sin embargo dada la escasez de datos, en el caso de este trabajo, solo se establecieron 100.

Respecto a la gran cantidad de features que se usa con la herramienta openSMILE, es posible que haya un sobreajuste, por lo cual es necesario implementar algoritmos que lo controlen para asegurar una cifra mas cercana a la real.

Por otra parte, el análisis estadístico realizado permite afirmar que el trabajo posterior realizado en base a los datos crudos tiene validez ya que es posible discriminar las clases analizadas en base a las 11 features utilizadas.

Al realizar el estado del arte en este trabajo, se navegó en motores de búsqueda de artículos académicos, usando palabras claves asociados a detección de agresión verbal y por lo que respecta al autor, no se encontró algún estudio que aborde el problema con el enfoque que se dio en este trabajo, es decir detectar agresiones verbales sin poner el foco en el estado emocional de quienes participan en el dialogo. Sí se pudo evidenciar una vasta literatura de detección en cuanto emociones por medio de audio, pero pocas contaban con la emoción “hot anger” o ira caliente dentro de sus estudios, que se relaciona muy de cerca con lo que siente una persona al momento de agredir verbalmente.

El trabajo ocupó conversaciones comunes que fueron en su totalidad tranquilas para realizar las comparaciones. Se podría subir el nivel de exigencia del modelo comparándolo con conversaciones más intensas y con otros ruidos humanos y no humanos que puedan tener features similares a los audios de agresión verbal.

7. Conclusiones y trabajo futuro

En este trabajo se propuso como objetivo encontrar algún método para poder detectar la violencia verbal que en este caso se contrasto con la clase conversaciones normales. Para esto se tuvo que realizar un estado del arte usando motores de búsqueda de artículos académicos como Google Scholar para examinar los diversos trabajos de diferentes autores. En esta búsqueda se pudo apreciar los avances en el ámbito de la vigilancia, en donde el problema común es detectar correctamente la ocurrencia de un evento de peligro respecto a otros sonidos a través de audio y también de video. En estos trabajos se usaron diversos métodos y modelos y también diferentes clases de interés en donde detectar percusiones de disparos y vidrios quebrándose eran los más recurrentes. Por lo tanto se pensó que uno de esos modelos podría ser útil para aplicar en la problemática que se estaba intentando resolver en este trabajo. El modelo de Carletti, Vincenzo, et al. fue elegido por ser considerado por el autor como un modelo intuitivo y práctico que puede ser entrenado fácilmente para propósitos similares, teniendo como principal beneficio la construcción de la bolsa de palabras (“aural words”), que permite considerar variadas formas de la clase que se desea detectar, en este caso se computaron las diversas formas de cada actor de manifestar su violencia verbal, teniendo una bolsa de palabras bastante variada. Además su autor indica que permite detectar sonidos cortos y largos durante los 3 segundos de cada ventana, lo cual es muy favorable dado que la violencia verbal se puede comportar de esa forma.

Además de esto se extrajeron los audios de internet correspondientes a las clases de “violencia verbal” y “conversaciones normales”. En total se contó con 1618 fragmentos de audio de 3 segundos, lo que es bastante, aunque para el ámbito del machine learning puede resultar muy escaso.

Por otro lado se considera cumplida la hipótesis de investigación sujeto al set de datos extraído ya que se logra diferenciar una agresión verbal de variadas conversaciones con una buena exactitud. Además se computaron archivos con agresión verbal de diferente duración en el segundo nivel del modelo (en la fase de SVM), pudiendo existir agresión verbal en los 3 segundos completos de los archivos o en pequeñas fracciones de segundo. Por lo tanto lo que decía Carletti, Vincenzo en su trabajo, se corroboró en cierta forma en este trabajo.

Como trabajo futuro se pueden realizar 3 acciones principales:

- Incrementar la base de datos procesando mayor cantidad de archivos variados de audio de violencia verbal en cuanto a diferentes personas, intensidad, contexto, etc.
- Se deben evaluar las features que están siendo consideradas ya que se usaron las mismas que uso Carletti, Vincenzo, que estaban pensadas para detectar otras clases. En este trabajo sirvieron las originales, pero quizás es posible incrementar el desempeño del algoritmo agregando o reemplazando por otras features, por ejemplo agregando la feature F0 que se modifica cuando una persona presenta la emoción de ira según lo explicado en el marco teórico.

- Optimizar los scripts utilizados, pudiendo de esta manera acercar lo máximo posible el procesamiento a monitoreo en tiempo real, para detectar si hay algún episodio de violencia en presencia de personas que no tienen facilidades de defenderse como los niños
- Introducir más clases en la base de datos permitir que el modelo se adapte a otros ruidos que suceden en la vida cotidiana y que podrían ser eventualmente confundidos con agresiones verbales.
- Controlar el sobreajuste de los modelos.
- Probar una mayor variedad de algoritmos de machine learning (e.g. redes neuronales, markov oculto, etc.)

Se espera que este trabajo pueda dar el pie para que futuros avances logren desarrollar productos que incluyan este tipo de algoritmos para detectar violencia verbal en tiempo real en donde existen situaciones de riesgo que puedan ser evitadas oportunamente, permitiendo intervenir a tiempo para que no se perpetúe un mal mayor, teniendo en consideración que es muy frecuente que en los casos en que existió violencia física, esta fue generalmente precedida por violencia verbal. Algunas situaciones en donde los descubrimientos de este trabajo pueden ser útiles son, por ejemplo, ambientes en donde se convive con niños, en situaciones de violencia doméstica, como seguridad en el hogar, entre otras aplicaciones.

8 Glosario

Canales: Un camino para una señal. Por ejemplo, un micrófono mono conectado a una entrada mono tiene un canal de entrada. Una señal estéreo conectada a dos altavoces tiene dos canales de reproducción [32].

Codificación con pérdida (“Lossy coding”): Usa un modelo perceptual para codificar niveles y arroja información basada en la incapacidad del oído para oír sonidos de bajo nivel en presencia de otros ruidos en el mismo rango de frecuencia [33].

Codificación sin pérdida (“Loseless coding”): Un método de codificación de audio que reduce su bitrate y tamaño de archivo sin perder información audible [33].

Frecuencia de muestreo: El número de veces por segundo del convertidor analógico a digital muestrea la señal analógica. La frecuencia de muestreo determina el rango de frecuencia de la grabación. Teóricamente, es posible una representación digital perfecta de una señal de audio analógica cuando la frecuencia de muestreo es por lo menos dos veces la frecuencia más alta de la señal. El mejor oído humano puede escuchar hasta 20-24 kHz, por lo que una tasa de muestreo de 40-48hz puede (teóricamente) reproducir toda la gama de la audiencia humana [32].

Frecuencia Fundamental (F0): Se relaciona con el tono de voz que se percibe. Es el reflejo de las características biomecánicas de las cuerdas vocales en tanto interaccionan con las presiones su glóticas y en tanto se modifican por la estructura laríngea y la fuerza muscular aplicada.

Profundidad de bits (Bit depth): Controla el rango dinámico, la relación señal-ruido y fidelidad y precisión general. Un rango dinámico más amplio da como resultado una mayor relación señal-ruido. Una mayor profundidad de bits resulta en una conversión más exacta y fiel desde una fuente analógica [32].

Resolución de frecuencia: El tamaño de la FFT (Fast Fourier transform) define el número de compartimientos utilizados para dividir la ventana. Por lo tanto, un compartimento es una rango de espectro, y define la resolución de frecuencia de la ventana.

$$frecuencybinrange = \frac{SampleFreq}{num(DFT\text{points})} \quad (12)$$

Tasa de bits (Bit rate en inglés): Consiste en el número computado de bits transmitidos o procesados por unidad de tiempo. Normalmente esta expresado en kilobits por segundo (kbps). Para un archivo PCM sin comprimir (“uncompressed”), la velocidad de bits, kbps, es la frecuencia de muestreo multiplicada por la profundidad de bits (“bit depth”) multiplicado por el número de canales. La tasa de bits son mucho más bajos para los formatos comprimidos (“compressed”) o con pérdida (“lossy”) [28].

Tasa de bits constante (CBR por sus siglas en inglés): En este formato, se usa el mismo número de bits para grabar la misma duración de sonido.

Tasa de bits variable (VBR por sus siglas en inglés): Es un método para comprimir audio que no siempre usa el mismo número de bits para grabar la misma duración de sonido [28].

9. Bibliografía

[1] http://www.24horas.cl/nacional/menor-de-solo-10-anos-denuncia-multiples-agresiones-propinadas-por-su-padre-1568570. visitada en Mayo de 2017.

[2] http://impresa.elmercurio.com/Pages/NewsDetail.aspx?dt=2017-03-01&dtB=01-03-2017%200:00:00&PaginaId=10&bodyid=3.visitada en Mayo de 2017.

[3] https://www.publimetro.cl/cl/nacional/2016/10/27/80-chilenos-viven-alto-nivel-estres.html. visitada en Mayo de 2017.

[4] http://edition.cnn.com/2016/10/11/health/reducing-stress-in-the-city-can-improve-mental-health/. visitada en Mayo de 2017.

[5] Marion K Underwood, “Social Aggresion Among Girls”, 2003.

[6] McCabe, Allyssa, and Thomas J. Lipscomb. "Sex differences in children's verbal aggression." Merrill-Palmer Quarterly (1982-) (1988): 389-401.

[7] Carletti, Vincenzo, et al. "Audio surveillance using a bag of aural words classifier." Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on. IEEE, 2013.

[8] Study, Biological Sciences Curriculum, and National Institutes of Health. "Information about Hearing, Communication, and Understanding." (2007).

[9] van Hengel, Peter WJ, and Tjeerd C. Andringa. "Verbal aggression detection in complex social environments." Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007.

[10] Foggia, Pasquale, et al. "Cascade classifiers trained on gammatonegrams for reliably detecting audio events." Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on. IEEE, 2014.

[11] Lecomte, Sébastien, et al. "Abnormal events detection using unsupervised One-Class SVM-Application to audio surveillance and evaluation." Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on. IEEE, 2011.

[12] Valenzise, Giuseppe, et al. "Scream and gunshot detection and localization for audio-surveillance systems." Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007.

[13] Atrey, Pradeep K., Namunu C. Maddage, and Mohan S. Kankanhalli. "Audio based event detection for multimedia surveillance." Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 5. IEEE, 2006.

[14] https://mediaarea.net/en/MediaInfo. Visitado Julio 2017.

[15] http://www.audacityteam.org/. Visitado Julio 2017.

[16] https://www.RStudio.com/products/. Visitado Julio 2017.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases [Internet]. 1996.

[18] Sammut, Claude, and Geoffrey I. Webb, eds. *Encyclopedia of machine learning and data mining*. Springer, 2016.

[19] Ghazanfar, Asif A., and Drew Rendall. "Evolution of human vocal production." Current Biology 18.11 (2008): R457-R460.

[20] Davis, Gary, and Gary D. Davis. The sound reinforcement handbook. Hal Leonard Corporation, 1989.

[21] Concepción Fernandez Gonzalez. Auxiliar Cuidador. Temario Y Test. E-book. MAD-Eduforma. 1997.

[22] Richard Pak, Anne McLaughlin. Designing Displays for Older Adults. CRC Press, 2010.

[22.3] https://www.cdc.gov/ncbddd/hearingloss/sound.html visitada en Julio 2017.

[23] Elizabeth M. Varcarolis. Essentials of Psychiatric Mental Health Nursing: A Communication Approach to Evidence-Based Care. Elsevier Health Sciences. 2012.

[24] Elana I. Clausen. Psychology of Anger. Nova Science Publishers. 2007.

[25] Shashi Banzal. Data and Computer Network Communication. Firewall Media. 2007.

[26] https://support.microsoft.com/en-us/help/15070/windows-media-player-codecs-frequently-asked-questions visitada en Julio de 2017.

[27] Ed Tittel, Chris Minnick. Beginning HTML5 and CSS3 For Dummies. John Wiley & Sons,2013.

[29] Cliff Truesdell. Mastering Digital Audio Production: The Professional Music Workflow with Mac OS X. John Wiley & Sons, 2007.

[28] http://manual.audacityteam.org/man/glossary.html visitada en Julio 2017.

[30] Tony Bove. iPod & iTunes For Dummies. John Wiley & Sons, 2010.

[31] Brandenburg, Karlheinz. "MP3 and AAC explained." Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding. Audio Engineering Society, 1999.

[32] Carla Schroder. The Book of Audacity: Record, Edit, Mix, and Master with the Free Audio Editor. No Starch Press, 2011.

[33] Bob Katz. ITunes Music: Mastering High Resolution Audio Delivery: Produce Great Sounding Music with Mastered for ITunes. CRC Press, 2013.

[34] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3 (2011): 572-587.

[35] Pikrakis, Aggelos, Theodoros Giannakopoulos, and Sergios Theodoridis. "Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks." Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008.

[36] De Santo, Massimo, et al. "Classifying audio of movies by a multi-expert system." Image Analysis and Processing, 2001. Proceedings. 11th International Conference on. IEEE, 2001.

[37] Schuller, Björn, et al. "Speaker independent speech emotion recognition by ensemble classification." Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005.

[38] http://filmratings.com/RatingsGuide visitada en Julio 2017.

[39] http://www.imdb.com/ visitada en Julio 2017.

[40] Begault, Durand R. "Forensic analysis of the audibility of female screams." Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice. Audio Engineering Society, 2008.

[41] Faustino Núñez Batalla, Carlos Suárez Nieto. “Manual de evaluación y diagnóstico de la voz.” Universidad de Oviedo, 1998.

[42] Charles Stephen Lessard. “Signal Processing of Random Physiological Signals.”. Morgan & Claypool Publishers, 2006.

[43] https://scholar.google.es/intl/es/scholar/about.html visitada en Julio 2017

[44] http://www.quodb.com/ visitada en Julio 2017.

[45] https://www.onlinevideoconverter.com/video-converter visitada en Julio 2017

[46] https://www.videolan.org/vlc/index.es.html visitada en Julio 2017

[47] http://audeering.com/technology/opensmile/ visitada en Julio 2017

[48] Eyben, F., M. Woellmer, and B. Schuller. "The openSMILE book-openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor.",2010.

[49] http://www.multimediaeval.org/about/ visitada en Julio 2017

[50] Sjöberg, Mats, et al. "The MediaEval 2014 Affect Task: Violent Scenes Detection." MediaEval. 2014.

[51] https://www.r-project.org/ visitada en Julio 2017.

[52] Anita V. Clark. Psychology of Moods. Nova Publishers, 2005.

10 Anexos

Nombre	Año de estreno	Calificación MPAA	Tasa de bits (kb/s)	Frecuencia (kHz)	Canales	Formato Audio
Pelicula1C	2012	R	96	48	2	AAC(LC)
Pelicula2C	2008	PG-13	93.9	48	2	AAC(LC)
Pelicula3C	1994	R	64	22.05	2	AAC(HE-AAC/LC)
Pelicula4C	2002	PG-13	96	48	2	AAC(LC)
Pelicula5C	2006	R	94	44.1	2	AAC
Pelicula6C	2013	R	93.8	48	2	AAC(LC)
Pelicula7C	1997	PG-13	64	48	2	AAC(HE-AAC/LC)
Pelicula1S	2005	R	93.8	48	2	AAC(LC)
Pelicula2S	2012	R	96	48	2	AAC(LC)
Pelicula3S	2008	R	113	48	2	AAC(LC)
Pelicula4S	2014	R	93.8	48	2	AAC(LC)
Pelicula5S	2008	R	94	44,1	2	AAC(LC)
Pelicula6S	2006	R	160	48	2	AAC(LC)
Pelicula7S	2005	R	96	48	2	AAC(LC)
Pelicula8S	2003	R	93.8	48	2	AAC(LC)
Pelicula9S	2000	R	93.7	48	2	AAC(LC)
Pelicula10S	1991	R	93.7	48	2	AAC(LC)
Pelicula1E	2006	G	32	48	2	AAC(HE-AAC/LC)
Pelicula2E	2001	G	96	48	2	AAC(LC)
Pelicula3E	2001	PG	64	48	2	AAC(LC)
Pelicula4E	1995	G	96	48	2	AAC(LC)
Pelicula5E	2009	PG	93.7	48	2	AAC(LC)
Pelicula6E	1980	R	32	48	2	MP3

Tabla 12: Datos originales previo procesamiento. Los archivos audiovisuales se encontraban en formato mp4. Esta información se obtuvo con el software MediaInfo y el sitio web IMDb [39].

Película 2C	Película 3C (1/2)	Película3C (2/2)	Película4C	Película5C (1/2)	Película5C (2/2)	Película6C (1/2)	Película6C (1/2)	Película7C
198	226	147	202	159	315	225	415	123
216	438	155	260	61	301	108	144	154
244	287	204	144	125	356	108	171	100
190	166	200	144	142	106	189	207	154
263	204	132	219	90	246	189	162	651
192	264	196	143	346	182	243	207	296
212	1428	143		130	116	180	234	332
228	204	105		132	149	180	171	186
268	188	98		187	142	153	216	227
238	143	196		287	61	135	207	
157	151	665		180	19	153	216	
238	257	166		83		126	234	
152	143	204		159		424	108	
623	136	166		59		180	171	
265	234	317		80		225	433	
531	302	272		394		180	126	
	109	317		197		288	207	
	215	120		230		135	261	
	211	204		341		189	216	
	102			189		1317	830	
	298			135		126	204	

Tabla 13: Duración de fragmentos de audio según cada película para entrenar los clústeres en la primera etapa del método de Carletti, Vincenzo. Datos en milisegundos