Detección acústica de situaciones violentas en transporte público[1]

García Gómez, J.
Departamento de Teoría de la Señal y Comunicaciones, Universidad de Alcalá de Henares. Madrid. España
joaquin.garciagomez@uah.es

PACS: 46.60.c

Premio SEA: Trabajo Fin de Máster.75 Aniversario del edificio "Torres Quevedo"

Resumen

La detección de violencia es un importante problema a tener en cuenta en el diseño de algoritmos para entornos inteligentes. Además, el transporte público constituye un escenario susceptible de sufrir situaciones violentas debido a su naturaleza dinámica. Este trabajo propone un sistema capaz de detectar acústicamente este tipo de situaciones de manera eficiente. En nuestra solución se utilizan algoritmos genéticos para seleccionar el subconjunto de características que mejor funciona. Tras realizar un estudio del sistema, se demuestra la viabilidad del mismo gracias al bajo coste que requieren ciertas características. Esto hace factible su implementación en un microprocesador actual.

Palabras clave: detección acústica, situaciones violentas, transporte público, procesado de audio, inteligencia artificial, machine learning, reconocimiento de patrones, extracción de características, selección de características, algoritmos genéticos, clasificación.

Abstract

Violence detection represents an important issue to take into account in the design of algorithms for smart environments. In addition, public transport constitutes a scenario where violent situations can take place because of its dynamic nature. This project proposes a system capable of efficiently detecting these kinds of situations through audio source. In our solution, genetic algorithms are used to select the best subset of features. After studying the system, the viability of it is demonstrated thanks to the low cost that certain features require. It makes feasible an implementation in a nowadays microprocessor.

Keywords: acoustic detection, violent situations, public transport, audio processing, artificial intelligence, machine learning, pattern recognition, feature extraction, features selection, evolutionary algorithms, classification.

1. Introducción

La violencia es un problema aún presente en la sociedad hoy en día. Prácticamente todo el mundo ha presenciado o sido víctima alguna vez de una situación violenta, en cualquiera de sus vertientes (física o verbal), y no parece avistarse un horizonte cercano en que este tipo de situaciones haya sido mitigado por completo. La violencia se define, según la Organización Mundial de la Salud (OMS), como *el uso intencional de la fuerza física, amenazas contra*

[1] El proyecto "Detección acústica de situaciones violentas en transporte público" fue galardonado el pasado 22 de junio de 2018 en la XXXVIII Edición de los Premios del Colegio Oficial de Ingenieros de Telecomunicación (COIT) y Asociación Española de Ingenieros de Telecomunicación (AEIT) con el **Premio FERMAX a Mejor Trabajo Fin de Máster en Técnicas de Comunicaciones en Entornos Residenciales.** Se trata de unos premios (https://goo.gl/J882BJ) que contaron con el apoyo de 20 empresas e instituciones colaboradoras, y en el que se concedieron 29 galardones entre Trabajos Fin de Máster, Tesis Doctorales y Trayectoria Académica, habiéndose presentado 115 propuestas de un total de 34 centros participantes.

El proyecto ha aparecido en varios medios de comunicación, incluyendo el programa Esto *me suena de* ***Radio Nacional,*** y el espacio de Ciencia y Tecnología del programa Hoy *por Hoy de* ***Cadena Ser Henares,*** donde se entrevistó al autor acerca del trabajo.

uno mismo, otra persona, un grupo o una comunidad que tiene como consecuencia o es muy probable que tenga como consecuencia un traumatismo, daños psicológicos, problemas de desarrollo o la muerte [1]. Para poner en perspectiva la necesidad de mitigación de la violencia a nivel mundial y dar una visión general a la preocupante situación actual, se van a citar algunas estadísticas proporcionadas por la OMS en un informe de mayo de 2017 [2]:

- La violencia produce 1,4 millones de muertes al año, lo que equivale a más de 3.800 muertes diarias. Este dato evidencia que se trata de un importante problema que afecta a la salud pública, los derechos y el desarrollo humano.
- De entre los fallecidos por violencia, un 56% se debe a suicidios, un 33% a lesiones ocasionadas por terceros y un 11% a guerras u otras formas de violencia colectiva.
- El 90% de este tipo de muertes suele ocurrir en países de ingresos bajos y medios. Los países con mayores niveles de desigualdad suelen presentar mayores tasas de mortalidad por violencia.
- Por cada persona que muere hay muchas más con lesiones y problemas relacionados con la salud física, sexual, reproductiva y mental. Concretamente, por cada joven fallecido debido a la violencia se calcula que entre 20 y 40 sufren lesiones que requieren tratamiento hospitalario.
- El impacto sanitario de la violencia no se limita a las lesiones físicas, sino que puede acarrear efectos tales como trastornos mentales, depresión, intentos de suicidio, síndromes de dolor crónico, embarazos no deseados, etc. Además, los niños que sufren violencia son más proclives a abusar del alcohol, las drogas, el tabaco o comportamientos sexuales de alto riesgo en el futuro. Como consecuencia, estas generaciones pueden desarrollar pasados los años enfermedades como cardiopatías, cánceres e infecciones de transmisión sexual.
- La violencia supone una enorme carga para la economía de los países, resultando en un coste nacional medio de miles de millones de dólares dedicados a atención sanitaria, vigilancia del cumplimiento de la ley y pérdida de productividad.

La violencia se encuentra presente en casi la totalidad de ámbitos de la vida. Sin embargo, se ha decidido centrar el estudio en los medios de transporte público. Concretamente la finalidad de este trabajo será el desarrollo de un sistema inteligente susceptible de ser implementado en un autobús, tren o metro, con el objetivo de mejorar la seguridad de los viajeros gracias a una monitorización acústica permanente del escenario mediante micrófonos. De esta forma, una determinada alarma podría dar aviso a las fuerzas de seguridad pertinentes para que intervengan y eviten que una pelea o discusión tenga consecuencias fatales sobre la víctima o víctimas. Además la lógica nos dice que si un sistema de este tipo se encuentra implementado, el número de situaciones violentas se verá reducido, ya que los agresores tendrán más temor a ser arrestados debido a sus acciones.

1.1. La violencia en el transporte público

Centrándonos en el mundo del transporte público, existen numerosas razones por las que la prevención de la violencia en este tipo de entornos es un tema de interés. En primer lugar, el transporte público es un elemento que ayuda a reducir la exclusión social, ya que proporciona acceso al trabajo, la sanidad o el ocio a millones de personas alrededor del mundo. Además, proporciona beneficios medioambientales, ya que se trata del medio de transporte más sostenible. Por esta razón, el número de personas que viaja a través del transporte público debería crecer en el tiempo en vez de disminuir.

Diversas encuestas han identificado que los factores por los que la gente deja de usar el transporte público son el miedo a la delincuencia y la seguridad personal, seguido de la accesibilidad a los mismos. De hecho, una encuesta de Reino Unido elaborada por el Departamento de Medio Ambiente, Transporte y Región reveló que la población estaría un 3% más dispuesta a coger el transporte público en las horas punta y hasta un 10% en el resto de horas si perdieran el miedo a la delincuencia. [3] De este tipo de informes se puede deducir la importancia de conseguir que los pasajeros se sientan totalmente seguros cuando viajan. Todos los sistemas de transporte deberían asegurar este hecho desde hace mucho tiempo. [4]

Una reducción de los niveles actuales de delincuencia no tiene por qué traducirse necesariamente en una reducción de la percepción y el miedo que tiene la gente a la delincuencia. Este miedo relacionado con la seguridad personal no dista mucho del que tiene la gente cuando viaja fuera de la red de transporte y depende de cuestiones personales como género, etnia, edad o frecuencia de uso de estos medios. Sin embargo, el transporte público posee determinadas características distintivas. Una de las más destacables es su naturaleza dinámica, que le hace recibir continuamente entradas y salidas de personas a lo largo del tiempo, de manera que las posibles víctimas viajan a través de zonas con diferente riesgo de delincuencia. Esto crea un ambiente en el que la probabilidad de ocurrencia de un delito se ve aumentada con respecto a otros ámbitos.

La información disponible acerca de la delincuencia en el transporte público es muy limitada y existen discrepancias en los datos. Esto se debe a que hay múltiples agencias responsables del mantenimiento y la seguridad de los sistemas, hay una falta de estándares en cuanto a los informes y registros de las situaciones violentas y el análisis de los datos disponibles es complejo. [3] A continuación citaremos algunas estadísticas disponibles acerca del tema. Por conocer la perspectiva del viajero, una encuesta de Victimización y Eficacia Institucional realizada en 2007 en el área metropolitana de Ciudad de México mostró que aproximadamente un 36% de los usuarios que usan el transporte público se sienten inseguros o muy poco seguros al viajar. [5] Además, el miedo a la delincuencia aumentaba considerablemente si el trayecto era superior a 30 minutos. Por otro lado, un estudio del Departamento de Transporte de Reino Unido (2010) mostró que entre 2008 y 2009 se produjeron 12 acciones criminales por millón de pasajeros en el autobús, mientras que en el caso del metro fueron 13 por millón. Sin embargo, en un estudio de Los Ángeles de 2002 [6] esta cifra asciende a 1.55 casos de delincuencia por cada 100 viajes. Estudios más antiguos (1990) revelan que el 5% de los viajeros de Inglaterra fueron víctimas de violencia al viajar mientras que el 4% sufrieron algún tipo de robo.

Existen muchas dificultades a la hora de identificar el número de situaciones de violencia y delincuencia en el transporte público. Se cree que estos niveles habrán descendido como lo han hecho los niveles de delincuencia en todos los ámbitos a nivel mundial. No obstante, siguen apareciendo casos de forma diaria en este tipo de escenarios, por lo que la detección y mitigación de la misma sigue constituyendo un tema de especial interés.

2. Estado del arte

En esta sección se va a investigar acerca del estado del arte existente relacionado con la detección de violencia. Nos vamos a basar fundamentalmente en la definición de violencia considerada, la fuente de información empleada en el estudio, la base de datos, el tipo de violencia considerada y los métodos propuestos.

Una de las diferencias respecto a otros estudios realizados en este ámbito radica en la definición de violencia que se tiene en cuenta. Mientras en este estudio se ha considerado la definición general que proporciona la OMS, otros estudios consideran únicamente la violencia física: *accidente que causa lesiones o dolor* [7], *conjunto de acciones humanas acompañadas de sangrado* [8]; o puramente centradas en el cine: *escenas que no deberían ser vistas por un niño de 8 años en una película* [9]. Esta consideración tiene una clara influencia a la hora de enfocar el estudio. De hecho, algunas de ellas no serían válidas en nuestro estudio: no seríamos capaces de detectar sangre con información puramente acústica.

La detección de violencia puede llevarse a cabo a través de sistemas basados en audio y/o vídeo. Algunas investigaciones presentes en la literatura tratan de resolver el problema con sistemas que procesan las dos fuentes de información de forma simultánea [8,10,11]. Los resultados obtenidos en este caso son buenos, por lo que la combinación de ambas es eficiente. Sin embargo, el vídeo presenta algunas desventajas respecto al audio. Por un lado el coste de un sistema basado en videocámaras, tanto en términos de implementación como a nivel computacional, es superior al de otro basado únicamente en micrófonos. Por otro lado, la intrusión llevada a cabo por los sistemas de vídeo puede acarrear problemas de privacidad de los usuarios del sistema. Además, la relación área de cobertura-coste que presentan los micrófonos es superior a la de las videocámaras.

Algunos trabajos han testeado el rendimiento de los algoritmos utilizando las fuentes de audio y vídeo de forma separada y de forma conjunta [12], concluyendo que el sistema funciona correctamente y es solvente empleando únicamente la información del audio. Si se introduce el vídeo el rendimiento mejora ligeramente, pero el coste computacional también es superior. En la literatura existen otros estudios que emplean solamente el audio para tratar de detectar violencia [13], ya que este tipo de situaciones suele caracterizarse por venir acompañada de discusiones, gritos, aumento en el volumen de la conversación, etc. Los resultados obtenidos también son destacables, al alcanzarse un 90% de detecciones correctas.

No existe variedad de bases de datos en los estudios realizados hasta ahora. Las más destacadas son las empleadas entre los años 2012 y 2014 en el dataset de MediaEval [14]. Se trata fundamentalmente de un etiquetado de escenas violentas presentes en diversos vídeos. A pesar de ser extensa, la mayoría de estos vídeos se corresponde con violencia fingida extraída de películas de Hollywood, por lo que no se trata de una base de datos adecuada para el trabajo presente. Otros estudios elaboran sus propias bases de datos, pero o bien se centran en violencia no real [13], o bien se trata de una base de datos no liberada [15]. Por todas estas razones, y como se comentará en detalle en secciones posteriores, en este estudio se ha elaborado una base de datos centrada en violencia real.

No existe variedad de bases de datos en los estudios realizados hasta ahora. Las más destacadas son las empleadas entre los años 2012 y 2014 en el dataset de MediaEval [14]. Se trata fundamentalmente de un etiquetado

de escenas violentas presentes en diversos vídeos. A pesar de ser extensa, la mayoría de estos vídeos se corresponde con violencia fingida extraída de películas de Hollywood, por lo que no se trata de una base de datos adecuada para el trabajo presente. Otros estudios elaboran sus propias bases de datos, pero o bien se centran en violencia no real [13], o bien se trata de una base de datos no liberada [15]. Por todas estas razones, y como se comentará en detalle en secciones posteriores, en este estudio se ha elaborado una base de datos centrada en violencia real.

3. Originalidad del trabajo

En esta sección se van a detallar los elementos que dotan de originalidad al trabajo respecto al estado del arte existente:

- Como ya se ha comentado en la sección anterior, son novedad tanto la detección de violencia considerada como la base de datos utilizada en los experimentos. Se trata de dos consideraciones ambiciosas, que tratan de generalizar al máximo posible los resultados obtenidos.
- El hecho de centrar el estudio en los medios de transporte público no se había planteado en trabajos anteriores.
- Se ha llevado a cabo un estudio exhaustivo del coste computacional que conlleva la implementación de este algoritmo en un microprocesador, centrando el análisis en el número de FLOPS (Floating Operations Per Second) requerido por las diferentes etapas de cálculo del sistema. Como consecuencia de este análisis se han obtenido una serie de ecuaciones asociadas a cada una de las características propuestas, tanto en el dominio del tiempo como en el de la frecuencia, que permite conocer el número de operaciones en base a una serie de parámetros (número de muestras, número de tramas, solapamiento, etc.). De esta manera, otro autor puede calcular de forma sencilla el coste asociado a cualquier sistema basado en estas características, muy comunes en estudios de reconocimiento de voz.
- Se ha propuesto como novedad una característica (Tasa de Tramas Sordas, ó Ratio of Unvoiced Frames) basada en el número de tramas sonoras respecto al total de tramas del segmento de audio, algo que queda determinado a través del cálculo del pitch. Además, esta característica fue seleccionada con una tasa de selección bastante alta (casi el 90%, con clasificador lineal) en los resultados de los experimentos.
- Los algoritmos evolutivos o genéticos empleados para la selección de un subconjunto de características han sido implementados con una versión basada en torneo eliminatorio, una técnica novedosa que permite una reducción en el coste computacional de los mismos, así como un aumento en la convergencia al evitar los mínimos locales que caracterizan a los problemas de optimización.

4. Resultados

En esta sección se van a explicar los distintos experimentos que se han llevado a cabo en el trabajo, así como los resultados obtenidos en cada uno de ellos. En primer lugar, se va a describir la base de datos utilizada, incluyendo su desarrollo y las características que presenta. A continuación se va a exponer el sistema que se ha implementado. Por último, se llevarán a cabo distintos experimentos en que se estudiará el efecto que tiene cada uno de los parámetros de diseño en el sistema y en los resultados obtenidos.

4.1. Base de datos

Como ya se ha ido comentando, en este TFM se ha desarrollado una base de datos nueva. Los pasos que se han seguido para el desarrollo de la base de datos se enumeran y desarrollan a continuación:

- Obtención de vídeos. El primer paso para la elaboración de la base de datos consiste en la descarga de vídeos de la popular plataforma *YouTube*. Se ha realizado una búsqueda de vídeos grabados en distintos medios de transporte público (tren, metro, autobús), donde tengan o no lugar situaciones violentas, como es el caso de discusiones, peleas, subidas de tono, etc. entre dos o más personas.
- Extracción, validación y remuestreo del audio. A continuación el audio ha sido extraído a través del software *Audacity* (en formato *wav*). Se ha impuesto una frecuencia de muestreo mínima de 22,050 Hz, de manera que todo audio con una frecuencia inferior a dicho umbral ha sido descartado. Una vez hecho esto, todos los audios han sido remuestreados a dicha frecuencia.
- Etiquetado de los audios. Tras la normalización de los audios, se ha procedido al proceso de etiquetado de los mismos. Este paso consiste en escuchar los diferentes audios y decidir de manera binaria (1 - violencia, 0 - no violencia) cuándo está teniendo lugar una situación violenta y cuándo no. Dado que este etiquetado puede resultar confuso y es fuertemente dependiente de la persona encargada de valorarlo, dos personas distintas han lleva-

do a cabo el proceso. De esta manera, la comparación de opiniones hace que el etiquetado de los datos resulte más objetivo y eficiente.

La base de datos está compuesta por 109 audios, que equivalen a una duración de casi ocho horas, de las cuales aproximadamente una hora corresponde a momentos en los que se están produciendo situaciones violentas (aproximadamente 11% del total).

4.2. Sistema propuesto

El sistema parte de la base de datos de violencia comentada en la sección anterior, en la cual todos los archivos se encuentran muestreados a una frecuencia fs=22,050 Hz. A continuación, se aplica un enventanado que da lugar a tramas de L=512 muestras, con un solapamiento entre tramas de S=50%. Una vez acomodados los audios, se aplica a cada una de las tramas una serie de características, que son unas medidas o cualidades que nos permiten extraer información útil de la señal de audio. Las características que se han empleado forman parte de la literatura y se han aplicado exitosamente en estudios de reconocimiento de voz. Se nombran a continuación, así como el grupo de características al que pertenecen.

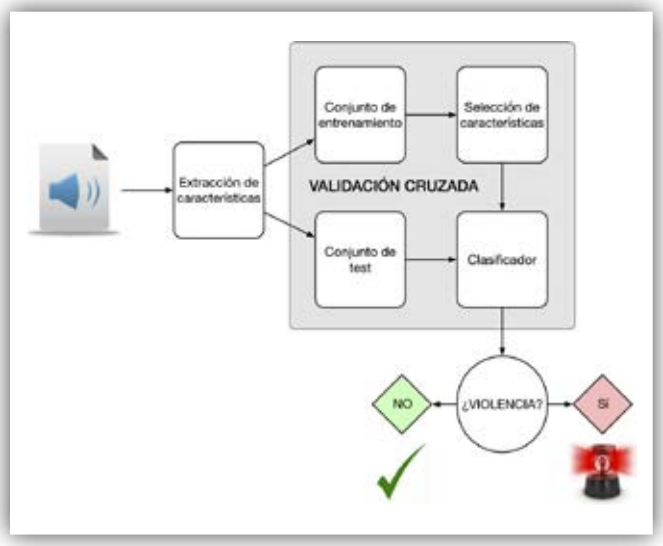


Figura 1. Esquema del sistema propuesto.

- Coeficientes Cepstrales en Frecuencia de Mel (MFCCs) y ΔMFCCs — G1
- Pitch y Tasa de Tramas Sordas (RUF). Esta última se propuso por 1ª vez en este TFM. — G2
- Tasa de Ruido Armónico (HNR) — G2
- Energía a Corto Plazo (STE) — G3
- Entropía de la Energía (EE) — G4
- Tasa de Cruces por Cero (ZCR) — G5
- Spectral Rolloff (SR) — G6
- Spectral Centroid (SC) — G7
- Spectral Flux (SF) — G8

A continuación se aplican los estadísticos correspondientes (media, desviación estándar, etc.). Este paso depende de cada cuánto tiempo se desee tomar una decisión, acerca de la presencia o no presencia de violencia. Considerando un tiempo de decisión T=1 segundo, los estadísticos deberán aplicarse cada 86 tramas de L=512 muestras con un solape de S=50%. Se obtendrán un total de N=121 características que modelan cada segundo de la base de datos.

Una vez extraídas las características se aplica la técnica de validación cruzada k-fold para obtener el mejor subconjunto de características para cada segmento, y a través de los clasificadores se determinará si existe violencia o no en cada instante de tiempo. En caso positivo, un sistema real crearía un aviso de alarma, que provocaría la intervención de la autoridad pertinente en el medio de transporte en que se encuentre instalado.

4.3. Experimentos

Los experimentos van a tratar de maximizar la Probabilidad de Detección Pd obtenida para una Probabilidad de Falsa Alarma Pfa=10%. Se ha seleccionado un valor bajo a optimizar porque así el algoritmo garantiza la máxima detección posible sin que el sistema cree excesivas falsas alertas. En una implementación real de un sistema de este tipo no interesa que se esté alertando de situaciones violentas constantemente cuando realmente no están teniendo lugar.

Los parámetros que se van a variar en los diferentes experimentos son: número de características seleccionadas Ncarar (posibles valores: 10, 20, 30, 40, 50); tiempo de observación Tobs (1, 2, 4, 8 seg), que es la duración considerada desde el momento presente hacia atrás para tomar la decisión; y coste computacional, medido en Mega Operaciones Flotantes Por Segundo (MFLOPS). Para los experimentos se han empleado algoritmos genéticos basados en torneo eliminatorio y los clasificadores empleados para la selección de características mediante algoritmos genéticos han sido el Discriminante Lineal de Mínimos Cuadrados (LSLD) y Discriminante Cuadrático de Mínimos Cuadrados (LSQD), mientras que la evaluación del rendimiento se ha llevado a cabo mediante LSLD, LSQD y Perceptrón Multicapa (MLPs).

En las Tablas 1.1, 1.2 y 1.3 se muestran las Pd obtenidas para una Pfa de 10%, variando los parámetros ya indicados. De izquierda a derecha, la evaluación se

Tablas 1.1, 1.2 y 1.3. Probabilidad de Detección para Probabilidad de Falsa Alarma 10%. LSLD para selección de características y evaluación (izquierda), LSQD para selección de características y evaluación (centro); y LSLD para selección de características y MLP con 5 neuronas para evaluación (derecha).

P_d (%)	N_{car}				
Tobs	10	20	30	40	50
1 s	67.6	67.3	71.2	70.9	69.9
2 s	72.7	71.0	**74.3**	71.5	71.3
4 s	68.3	71.4	70.5	71.5	70.8
8 s	68.8	68.6	68.1	69.2	66.4

P_d (%)	N_{car}				
Tobs	10	20	30	40	50
1 s	67.9	71.6	70.9	72.4	69.9
2 s	70.9	71.2	74.9	**77.8**	72.3
4 s	72.4	66.0	73.4	72.0	70.8
8 s	67.2	65.9	65.3	69.0	66.4

P_d (%)	N_{car}				
Tobs	10	20	30	40	50
1 s	68.7	67.9	69.7	68.8	69.1
2 s	67.4	71.7	71.1	71.6	71.6
4 s	64.4	71.3	70.6	**73.7**	70.3
8 s	62.4	58.2	54.2	62.8	48.3

ha realizado con el clasificador LSLD, a continuación LSQD y por último MLPs. LSLD obtiene el mejor resultado con 3 características y 2 segundos de observación, alcanzando un 74.3% de detección, mientras que el peor caso es el de 50 características y 8 segundos de observación (66.4%). Por su parte, el detector cuadrático (LSQD) obtiene un destacado 77.8% cuando limita el número de características a 40 y el tiempo de observación a 2 segundos. El peor resultado aparece de nuevo al ampliar el tiempo de observación a 8 segundos y tomar 30 características (65.3%). En cuanto al uso de MLPs, el mejor resultado se obtiene con 40 características y 4 segundos de observación (74.3%), mientras que el peor se alcanza con 50 características y 8 segundos de observación, al alcanzar un pobre 48.3%.

Tratando de realizar un enfoque general de los resultados obtenidos, parece claro que los mejores resultados se obtienen cuando las características se limitan a valores en torno a 30-40. Los resultados empeoran bastante cuando nos acercamos a los extremos y, o bien tomamos muy pocas características (10-20), o bien tomamos demasiadas (50). En cuanto a los tiempos de observación, los valores óptimos aparecen cuando se emplean entre 2 y 4 segundos para tomar la decisión en el momento actual. Especialmente destacable de forma negativa es el caso de 8 segundos, donde el rendimiento sufre una gran caída. Era de esperar que un valor de observación tan grande produjera este efecto, pues las muestras que aparecían hace tanto tiempo en el audio no tienen especial relevancia en lo que esté ocurriendo actualmente. Acerca de los cla-

Tablas 2.1 y 2.2. Características más seleccionadas con LSLD (izquierda) y con LSQD (derecha).

Número	Medida	Estadístico	Aparición (%)
1	MFCC 1	Media	100.0
2	MFCC 2	**Media**	100.0
3	MFCC 13	Media	100.0
4	MFCC 3	SD	100.0
5	MFCC 6	SD	100.0
6	MFCC 1	SD	99.1
7	MFCC 8	Media	97.2
8	MFCC 11	Media	96.3
9	ΔMFCC 11	SD	95.3
10	MFCC 8	SD	91.6
11	MFCC 20	SD	88.8
12	RUF	–	88.8
13	ΔMFCC 3	SD	87.9
14	MFCC 6	Media	86.0
15	MFCC 10	Media	85.1
16	SR	SD	83.2
17	ΔMFCC 14	SD	76.6
18	MFCC 18	Media	72.9
19	SC	SD	60.8
20	MFCC 4	Media	54.2

Número	Medida	Estadístico	Aparición (%)
1	MFCC 1	Media	100.0
2	MFCC 6	**Media**	100.0
3	MFCC 8	Media	100.0
4	MFCC 1	SD	100.0
5	MFCC 3	SD	100.0
6	ΔMFCC 5	SD	100.0
7	EE	Media	100.0
8	MFCC 2	Media	99.1
9	MFCC 10	Media	99.1
10	MFCC 6	SD	99.1
11	MFCC 8	SD	99.1
12	ΔMFCC 1	SD	98.1
13	MFCC 25	SD	96.3
14	MFCC 5	Media	95.3
15	MFCC 13	Media	95.3
16	ΔMFCC 6	SD	93.5
17	ΔMFCC 11	SD	92.5
18	SR	SD	92.5
19	ΔMFCC 25	SD	87.9
20	MFCC 3	Media	86.0

sificadores aplicados, funciona mejor el Discriminante Cuadrático de Mínimos Cuadrados (LSQD). Por su parte, las Redes Neuronales (MLPs) obtienen resultados bastante dispares e inferiores a los de los otros dos clasificadores. Esto se debe a que probablemente se esté empleando un clasificador demasiado complejo para una base de datos tan limitada, por lo que puede estar acarreando problemas de sobreentrenamiento.

A continuación nos vamos a centrar en las características más seleccionadas en los distintos clasificadores. Para ello, tomaremos el mejor resultado obtenido en LSLD (30 características y 2 segundos) y LSQD (40 características y 2 segundos). En las Tablas 2.1 y 2.2 se muestran las 20 características más seleccionadas, ordenadas por porcentaje de aparición en los distintos segmentos de audio (LSLD izquierda, LSQD derecha).

En la tabla izquierda (LSLD) se puede observar la gran importancia que cobran los coeficientes MFCCs para la detección de violencia. De las 20 características más seleccionadas, 17 se corresponden con estos coeficientes. Además, es especialmente reseñable que las 11 primeras posiciones las ocupan éstos coeficientes, obteniendo 5 de ellos un porcentaje de aparición del 100% en todos los audios de la base de datos. También es conveniente señalar que los coeficientes que más aparecen son los primeros de los 25 calculados (1 al 6 aparecen en la lista), ya que es en ellos donde se concentra la mayor parte de la energía de la señal. Al margen de estos coeficientes, aparece la medida propuesta (RUF) que evalúa el número de tramas sordas respecto del total en cerca del 90% de los audios. Completan la lista el Spectral Rolloff (SR) y el Spectral Centroid (SC), ambas calculadas en el dominio de la frecuencia.

En cuanto a la tabla derecha (LSQD) los MFCC ocupan igualmente gran parte de la lista, de manera que 18 de las 20 características más seleccionadas corresponden a dichos coeficientes. Aparece en una posición destacada con un 100% de selección la medida de EE, la cual no aparecía en el clasificador lineal, y que resulta útil como ya se indicó en secciones anteriores para detectar cambios bruscos en el nivel de la señal, como puede ocurrir en una discusión o pelea, que es lo que se está tratando de detectar en este sistema. Por otro lado, en ambas tablas las filas sombreadas muestran aquellas características que aparecen en ambos clasificadores. De las 20 más seleccionadas, 12 de ellas se repiten en ambos clasificadores, por lo que se podría asegurar que funcionarán correctamente para resolver este problema de detección de violencia con mayor probabilidad que el resto de ellas.

A continuación se han aplicado diferentes umbrales de coste medidos en máximo número de MFLOPS (MaxMFLOPS = {1,3,5,10,15}). Esta restricción implica que la suma de costes de las características seleccionadas por el sistema debe estar por debajo de dichos valores. Una vez seleccionadas las mejores características, se aplica un clasificador que toma la decisión final respecto a la presencia o no de violencia en el ambiente. Se ha realizado una comparativa de coste de los ocho grupos de características. Esta comparativa se observa en la Figura 2, donde el coste que requiere la Short Time Fourier Transform (STFT) se ha representado en azul, mientras que el coste adicional de cada grupo de características aparece en naranja. A la vista de los resultados, se puede apreciar que el grupo G2 (pitch, HNR y RUF) es el más caro en términos computacionales, ya que supera los

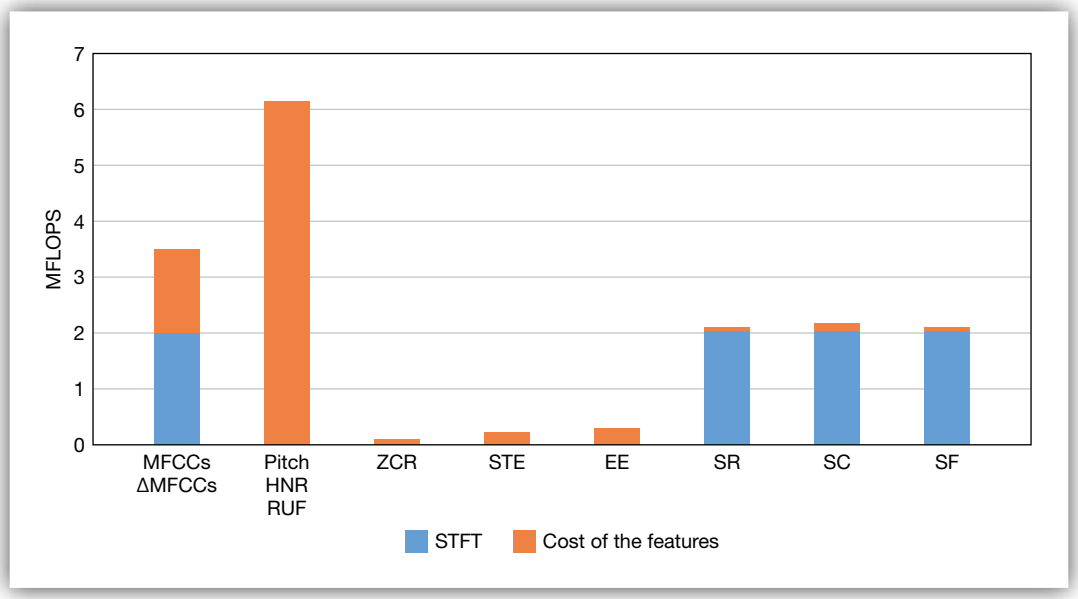


Figura 2. Coste de los diferentes grupos de características.

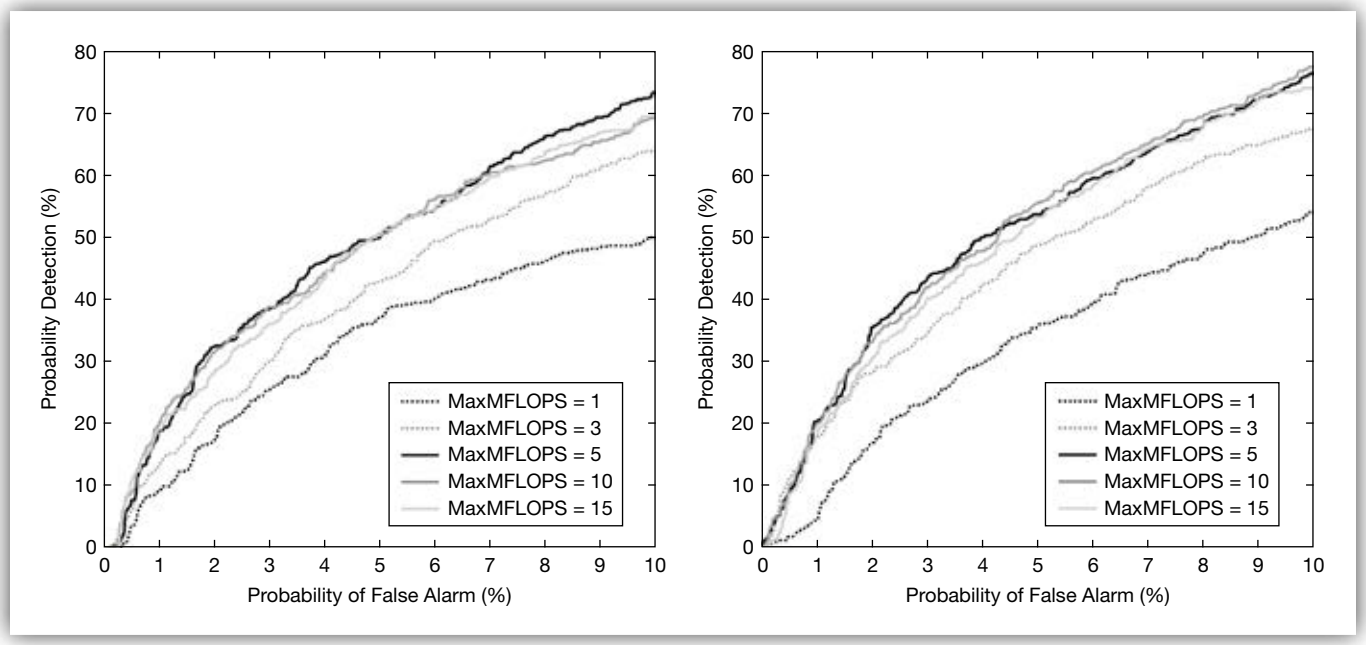
6 millones de FLOPS. El grupo G1 también es muy caro, alcanzando algo más de 3 millones, pero proporciona 100 características a los experimentos e incluye el cálculo de la STFT, empleada por otros grupos.

A continuación se procede a evaluar el efecto que tienen los umbrales de coste en el rendimiento del sistema. Las Figuras 3.1 y 3.2 muestran las curvas ROC (Pd en función de Pfa) para valores bajos de Pfa (inferiores al 10%), aplicando los 5 umbrales de coste y el detector lineal (a la izquierda) o detector cuadrático (derecha).

Los resultados obtenidos son similares con ambos clasificadores. Al aplicar umbrales muy restrictivos (1 MaxMFLOPS), la probabilidad de detección obtenida es muy pobre (alrededor del 50-55% para Pfa del 10%). Conforme aumentamos el umbral los resultados mejoran considerablemente, alcanzando alrededor de un 75-80% de detección con un umbral de 5 MaxMFLOPS. Sin embargo, esta mejora no continúa al aplicar umbrales menos restrictivos (10-15 MaxFLOPS), por lo que en este sistema carece de sentido aumentar los recursos del sistema por encima de dicho valor.

A continuación se va a estudiar qué grupos de características son los más seleccionados y útiles para resolver el problema. La Tabla 3 muestra el coste medio empleado, la probabilidad de detección alcanzada para probabilidades de falsa alarma de 10% y los porcentajes de aparición o ratios de selección de los diferentes grupos. Se ha considerado aparición de un grupo como la selección de una o más características de dicho grupo. Al principio, con el umbral más restrictivo (1 MaxFLOPS), el algoritmo selecciona los grupos G3, G4 y G5 debido a los pocos recursos con los que cuenta el sistema. Cuando aumentamos el valor a 3 MaxFLOPS aparecen las características espectrales. Por su parte, los MFCCs aparecen cuando el umbral aumenta a un valor igual o superior a 5 MaxMFLOPS, y el pitch con 10 MaxMFLOPs. El caso de 15 MaxMFLOPS permite al algoritmo seleccionar las características que necesite sin restricciones, ya que la suma total de coste s es inferior a dicho valor. Por tanto, este umbral equivaldría a no aplicar restricciones de coste.

Como puede observarse, algunas características funcionan mejor en el detector cuadrático que en el detector lineal, como en el caso del grupo G8 (SF), donde la diferencia de aparición entre los clasificadores es considerable. En los grupos G3 y G5 ocurre lo contrario. De hecho, el grupo G3 tiene una aparición en LSQD del 0%. En la tabla también queda reflejada la importancia y relevancia de ciertas características. Por ejemplo, cuando el grupo G1 (MFCCs y ΔMFCCs) aparece (desde 5 MaxMFLOPS en adelante), su aparición es del 100% tanto en clasificador lineal como cuadrático, a costa de reducirse la aparición de grupos de características que estaban presentes en umbrales inferiores, como ocurre con los grupos G4 y G5. De esta forma, y viendo cómo mejoran los resultados (+10% de detección en lineal y +9% en cuadrático) al ser seleccionado, queda demostrado que los MFCCs es el mejor grupo para la detección de violencia. Esto no ocurre en el caso del grupo G2 (pitch, HNR y RUF), el otro grupo del sistema costoso en términos computacionales, ya que no mejora los resultados cuando es seleccionado por el sistema (umbrales de 10-15 MaxMFLOPS).



Figuras 3.1 y 3.2. Curva ROC con LSLD (izquierda) y con LSQD (derecha).

Tabla 3. Coste, Pd y tasa de selección de los grupos de características.

MaxMFLOPS		1		3		5		10		15	
Clasificador		Lin.	Cua.	Lin.	Cua.	Lin.	Cua.	Lin.	Cua.	Lin.	Cua.
Coste Medio (MFLOPS)		0.4	0.4	2.6	2.6	3.9	3.7	9.8	8.3	10.0	8.8
P_d (P_{fa} = 10%) (%)		50	54	64	67	74	76	68	78	70	74
Tasa de selección (%)	G_1	0	0	0	0	100	100	100	100	100	100
	G_2	0	0	0	0	0	0	99	76	100	82
	G_3	93	0	19	0	63	0	80	0	63	0
	G_4	100	100	100	100	97	92	73	89	95	96
	G_5	100	100	100	100	80	31	25	12	80	35
	G_6	0	0	100	100	98	98	93	98	97	99
	G_7	0	0	100	9	82	77	99	78	97	86
	G_8	0	0	6	98	49	57	41	63	41	70

5. Desarrollo futuro

La detección de violencia en general, y centrada en transporte público en particular, es un campo de estudio amplio e incompleto, en el que se pueden abrir numerosas líneas de trabajo de cara al futuro. Teniendo en cuenta el punto en el que se deja el trabajo realizado hasta ahora en este trabajo, se podría avanzar en numerosas vertientes:

1. En primer lugar, es importante llevar a cabo una ampliación de la base de datos existente. En este trabajo se han recopilado casi 8 horas de audio, pero con un porcentaje de violencia de alrededor del 11% no estamos ante una base de datos amplia y generalizable. Este es un paso primordial si se desean aplicar técnicas más complejas de detección.
2. Relacionado con el punto anterior, sería conveniente introducir nuevos idiomas a la base de datos. Actualmente todos los audios son en inglés. Sin embargo, la forma en que una persona expresa actitud de ira, violencia, enfado, etc. es fuertemente dependiente del idioma, por lo que esta mejora la haría más generalizable.
3. Revisión de la base de datos actual y futura. El etiquetado de una base de datos no es algo totalmente objetivo, pues a la hora de etiquetar hay una fuerte dependencia de la persona que esté etiquetando, su estado de ánimo, edad, etc. En este estudio dos personas han etiquetado la base de datos y posteriormente se ha contrastado el trabajo. Sería conveniente que este trabajo fuera realizado por un mayor número de personas, para tratar de alcanzar la mayor objetividad posible.
4. Aplicación de otros clasificadores. En este estudio se han probado clasificadores sencillos para evitar problemas de entrenamiento, pues al introducir alguno más complejo (como redes neuronales), el rendimiento empeoraba. Con una base de datos adecuada sería posible testear el sistema con clasificadores tales como Máquinas de Vector Soporte (SVMs) o árboles de decisión.
5. Aplicación de técnicas más avanzadas y novedosas que con la base de datos actual producen sobreentrenamiento. Se podría plantear la introducción de técnicas de *Deep Learning*, que están funcionando en la actualidad.
6. Introducción de un diccionario de palabras clave y detector de las mismas. Para mejorar la detección sería interesante controlar la aparición de determinadas palabras que la gente pronuncia cuando se encuentra enfadada o en una situación de este tipo.
7. Implementación en un microprocesador con procesamiento de audio en tiempo real.

6. Bibliografía

[1] "Organización Mundial de la Salud (OMS)." Available at http://www.who.int/topics/violence/es/

[2] "10 datos sobre la prevención de la violencia, Organización Mundial de la Salud." Available at http://www.who.int/features/factfiles/violence/es/

[3] A. D. Newton, Crime and disorder on buses: towards an evidence base for effective crime prevention. PhD thesis, University of Liverpool, 2004.

[4] M. J. Smith and R. V. Clarke, “Crime and public transport”, Crime and Justice, vol. 27, pp. 169–233, 2000.

[5] C. J. Vilalta, “Fear of crime in public transport: Research in Mexico city”, *Crime Prevention and Community Safety*, vol. 13, no. 3, pp. 171–186, 2011.

[6] A. Loukaitou-Sideris, R. Liggett, and H. Iseki, “The geography of transit crime: Documentation and evaluation of crime incidence on and around the green line stations in Los Angeles”, Journal of Planning Education and Research, vol. 22, no. 2, pp. 135–151, 2002.

[7] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, “The mediaeval 2012 affect task: violent scenes detection”, in *Working Notes Proceedings of the MediaEval 2012 Workshop*, 2012.

[8] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, “Violence detection in movies”, in *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pp. 119–124, IEEE, 2011

[9] B. do Nascimento Teixeira, “Mtm at mediaeval 2014 violence detection task”, in *Working Notes Proceedings of the MediaEval Workshop*, 2014.

[10] M. Schedi, M. Sjöberg, I. Mironica, B. Ionescu, V. L. Quang, Y.-G. Jiang, and C.-H. Demarty, “VSD2014: a dataset for violent scenes detection in hollywood movies and web videos”, in *Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on*, pp. 1–6, IEEE, 2015.

[11] E. Acar, F. Hopfgartner, and S. Albayrak, “Breaking down violence detection: combining divide-et-impera and coarse-to-fine strategies”, *Neurocomputing*, vol. 208, pp. 225–237, 2016.

[12] L. J. G. Dias et al., *Detecting violent excerpts in movies using audio*. PhD thesis, 2016.

[13] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, “Violence content classification using audio features,” in *SETN*, pp. 502–507, Springer, 2006.

[14] “Mediaeval benchmarking initiative for multimedia evaluation.” Available at http://www.multimediaeval.org/mediaeval2017/

[15] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, “A multimodal approach to violence detection in video sharing sites,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3244–3247, IEEE, 2010.

decustik®

Paneles acústicos personalizados
la acústica sin límites

P.I. Mas les Vinyes - C/ Llevant, 2 - Torelló - Barcelona | T 93 859 08 38 | comercial@decustik.com | www.decustik.com