Estudio de alternativas de subtitulado accesible de estímulos sonoros no verbales para discapacidad auditiva

Doctoranda: Maria José Lucia Mulas

Tesis depositada en cumplimiento parcial de los requisitos para el grado de Doctor en

Ciencia y tecnología Informática

Universidad Carlos III de Madrid

Director/a (es/as):

Maria Belén Ruiz Mezcua

[Diciembre 2021/Enero 2022]

A Belén Ruiz

AGRADECIMIENTOS

Estoy profundamente agradecida a Belén Ruiz Mezcua por haberme dado tan generosamente la oportunidad de desarrollar esta tesis, por haberla dirigido, y por darme a conocer el gran proyecto que es Cesya. Ha sido un privilegio haber sido dirigida por una persona con tanta valía y ganas de mejorar el mundo.

También a todo el grupo de investigación de "Insubtitulables", a su generosa ayuda, y a ese espíritu alegre y distendido. A José Manuel Sánchez por su cariñosa acogida en el Cesya y en los "Insubtitulables", a Tomas Ortiz por haber dirigido la parte "cerebral" de esta tesis con tanto conocimiento y buen humor, y por tantas horas de formación y registros, a Pablo Revuelta, por iniciarme en esta investigación y compartir conocimientos, investigación y escritos, y por ser como es, a Álvaro García por su magnífico guante y por su ayuda y compañerismo incondicionales, a Víctor por esos vídeos profesionales y su infinita paciencia con las mil peticiones, a Ricardo Vergaz por sus ayudas inmediatas y esas preciosas cartas en inglés, a Elena Ortiz por su cariñoso apoyo con los EEG y sus ánimos. También al grupo HULAT, encabezado por Paloma Martinez, y a todo el CESYA por formar ese equipazo, por su ayuda siempre que lo he necesitado, y por acogerme: Mónica (¡gracias!), Blanca, José Luis, Israel, Luis, Juanma, Pachi, Leonardo y todos los que se han ido incorporando. ¡Gracias también por esos AMADIS!

Gracias finalmente a las participantes con discapacidad auditiva que asistieron a los experimentos y a Luisa Jiménez por su entusiasmo y sus consejos.

A nivel personal, quiero dar las gracias a mi familia, siempre, y a mi amigo Moncho, en el origen de todo.

CONTENIDOS PUBLICADOS Y PRESENTADOS

- Publicación 1:
  - Revuelta P., Ortiz T., Lucía M. J., Ruiz-Mezcua B., Sánchez-Pena J.M. (2020*). Limitations of standard accessible captioning of sounds and music for deaf and hard of hearing people: An EEG study*. Frontiers in Integrative Neuroscience, vol. 14, Feb. 2020.
  - doi: 0.3389/fnint.2020.00001
  - Frontiers in Integrative Neuroscience fue Q3(2020)
  - Rol: Coautora de la publicación
  - La publicación se basa en la experimentación desarrollada en la sección 3.3.

- Publicación 2:
  - Lucía M. J., Revuelta P., García A., Ruiz-Mezcua B., Vergaz R., Cerdán V., Ortiz T. (2020). *Vibrotactile Captioning of Musical Effects in Audio-Visual Media as an Alternative for Deaf and Hard of Hearing People: An EEG Study*. IEEE Access, vol. 8, pp. 190873-190881, Oct.2020.
  - doi: 10.1109/ACCESS.2020.3032229.
  - El video publicado con este artículo recibió el gran premio del IEEE Access Best Multimedia Award 2020 (Parte 2).
  - El artículo fue destacado como IEEE *Access* Article of the Week en septiembre 2021
  - *IEEE Access* fue Q2(2020) y Q1(2019)
  - Rol: Coautora de la publicación
  - La publicación se basa en la experimentación desarrollada en la sección 3.4.

- Publicación 3:
  - Revuelta P., Lucía M. J., Ortiz T., Ruiz-Mezcua B., Sánchez-Pena J.M. (2020*). Mapping and Timing the (Healthy) Emotional Brain: A Review* in *Biomedical Signal and Image Processing*. Intechopen Book Series, Ene. 2021
  - doi: 10.5772/intechopen.95574
  - Rol: Coautora de la publicación

- Ponencia 1:
  Lucía M. J., López J.L, Ruiz-Mezcua B., González I. *State of Art of Speech-to-Text technologies and its application on accessible subtitle creation.* 2nd Global Conference on Applied Physics, Mathematics and Computing (APMC-18).

- Ponencia 2:
  Revuelta P., Lucía M. J., López J.L, Ruiz-Mezcua B., González I. *Tecnologías de Transcripción en la nube y su aplicación al subtitulado accesible.* IX Congreso de Accesibilidad a los Medios Audiovisuales para Personas con Discapacidad (AMADIS 2018).

- Ponencia 3:
  Revuelta P., Lucía M. J., Ruiz-Mezcua B., Sánchez-Pena J.M. *Blas-T: Sincronización automática del subtitulado en directo.* IX Congreso de Accesibilidad a los Medios Audiovisuales para Personas con Discapacidad (AMADIS 2018).

TABLA DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

LISTA DE ACRÓNIMOS

AAL	Automated Anatomical Labeling
CESYA	Centro Español del Subtitulado y la Audiodescripción
CNN	Convolutional Neural Network
CQT	Constant-Q transform
EEG	Electroencephalography
ERP	Event-related potential
fMRI	Resonancia magnética funcional
ISMIR	International Society for Music Information Retrieval
LLSS	Lengua de signos
LORETA	Low resolution electromagnetic mography
MEB	Musical Emotional Burts
MediaEval	Benchmarking Initiative for Multimedia Evaluation
MER	Music Emotion Recognition
MFCC	Mel Frequency Cepstral Coefficients
MIR	Music Emotion Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MNI	Atlas cerebral del Instituto Neurológico de Montreal
PPM	Pulsos por minuto
ResNet	Residual Network
SPM	Statistic Parametric Maps
STFT	Short-time Fourier transform

1 INTRODUCCIÓN

1.1 Justificación

Para las personas con discapacidad auditiva, el subtitulado es el medio principal para acceder a los medios audiovisuales como televisión, cine, teatros, conferencias, exposiciones. Los subtítulos son el "audio" las para personas sordas o con problemas severos de audición.

El subtitulado accesible tiene como objetivo ideal permitir que las personas con dificultades auditivas entiendan y disfruten del contenido audiovisual de la misma manera que las personas oyentes. En España el subtitulado accesible se rige por la norma UNE 153010 (AENOR, 2012), elaborada por la Asociación Española de Normalización y Certificación. Esta normativa, y en general todas las normativas europeas, incluyen requisitos sobre diferentes dimensiones de los subtítulos para garantizar su calidad y homogeneidad. Además, en la televisión, según la ley general audiovisual (BOE, 2010), el 90% de la programación de los canales públicos debe de estar subtitulada para sordos, así como el 75% de los canales comerciales.

El subtitulado convencional más extendido es el que se usa para subtitular películas en el idioma original o en otro idioma, al tiempo que permite escuchar la voz original de los actores, presentando en el borde inferior de la imagen el texto correspondiente a la narración o diálogo que se está desarrollando en el medio audiovisual. El subtitulado accesible para personas con discapacidad auditiva incluye, además de la transcripción del texto, información para poder identificar a los hablantes (mediante etiquetas o colores), información suprasegmental como la entonación, el ritmo o la fluidez del habla (“habla entrecortadamente”, “susurra”, “dice con tristeza”), información sobre los sonidos relevantes para la comprensión de la obra (“suena un teléfono”, “retumba la puerta, “suena un disparo”), y sobre la música que se está emitiendo (“se oye música triste”).

Según la norma UNE 153010 (AENOR, 2012), el subtítulo accesible debe aparecer centrado en la parte inferior de la pantalla, siempre que no se superponga a una información relevante. El texto debe de ser lo más literal y preciso posible, y en general no se deben presentar más de dos líneas simultáneamente. Desde el punto de vista temporal, debe de estar sincronizado para coincidir con el movimiento labial y favorecer la lectura labial de los usuarios sordos que acceden a esta información, y la velocidad de presentación no debe de superar los 15 caracteres por segundo. Cada cambio de hablante debe identificarse con una nueva línea, y con un código de color distinto o bien con una etiqueta delante del texto, por ejemplo: “(PEDRO) Voy a salir”. La información de contexto o suprasegmental debe aparecer delante del texto en el que se produce: (TARTAMUDEA) (LLORA). La norma también establece pautas sobre la tipografía, tamaño, y contraste de colores de los subtítulos.

Ilustración 1. Ejemplo de subtitulado accesible (Rodríguez, 2021)

Para los efectos sonoros, la norma accesible requiere que el subtítulo correspondiente aparezca en la parte superior derecha de la pantalla, entre paréntesis, con la primera letra en mayúscula, sincronizado con el relato audiovisual. En el caso de la música, la norma (AENOR, 2012) indica específicamente que:

"Se debe subtitular la música si es importante para ayudar a comprender la trama utilizando uno o más de los tres contenidos siguientes:

el tipo de música;

la sensación que transmite;

identificación de la pieza (título, autor…).[1]"

Ilustración 2. Ejemplo de subtitulado accesible de la música

[1] Para las personas sordas con amplia cultura musical previa a la sordera, la identificación de la pieza musical es importante, ya que les permite evocar la música con su memoria.

Con respecto a las canciones se transcribe la letra de estas incluyendo el símbolo "#" en el inicio del subtítulo.

Ilustración 3. Ejemplo de subtitulado accesible de canciones

Respecto a las tecnologías de subtitulado, existen dos tipos fundamentales: en directo y en diferido. También se pueden considerar modelos mixtos. Un subtítulo es en directo cuando la transcripción se produce en tiempo real, y un subtítulo se denomina en diferido cuando se dispone de la grabación, se produce el subtítulo y se emite con posterioridad. Un modelo mixto combina ambas aproximaciones en el mismo documento audiovisual. Las tecnologías subyacentes son específicas para cada caso. En el caso en diferido, el texto se inserta en los programas grabados con anterioridad a su emisión (películas, series, documentales, concursos, programas de humor), mientras que en el caso en directo los subtítulos se van insertando en tiempo real a la vez que se emite el programa. Ejemplos de programación que utilizan los subtítulos en tiempo real son eventos deportivos, noticieros y otros eventos en directo que no permiten preparar subtítulos fuera de línea.

El subtitulado en directo se realiza con el soporte de programas informáticos desarrollados a medida basados en la técnica del rehablado (Romero-Fresco & Eugeni, 2020), utilizada en la mayoría de las televisiones, o en la estenotipia informatizada (Fuentes, González, & Ruiz, 2007) y genera un subtitulado convencional basado en la transcripción literal del discurso. El rehablado consiste en una técnica de transcripción utilizando un sistema de reconocimiento automático del habla con dependencia del locutor. Un profesional especializado escucha desde una cabina insonorizada el programa en directo con el audio original y "rehabla" sobre la marcha lo que escucha a través de un micrófono. El micrófono está conectado a un software de reconocimiento de voz especialmente entrenado para la voz del profesional. La aplicación va mostrando en pantalla el resultado del reconocimiento, y permite mediante teclas rápidas, editar y corregir la puntuación o las palabras incorrectamente transcritas antes de la emisión. Normalmente se produce un retraso de unos pocos segundos respecto a la emisión. Cumplir los criterios de calidad es un reto, y, de hecho, la norma matiza que "los subtítulos en vivo deben alcanzar la máxima precisión según las posibilidades tecnológicas del momento" y admite un retraso máximo de 8 segundos (AENOR, 2012). Aun así, el rehablado profesional produce subtítulos de calidad.

Ilustración 4. Ejemplo de técnica de rehablado

La estenotipia informatizada es una tecnología que mediante un dispositivo de estenotipia permite teclear con pocas teclas el discurso de manera que se pueda hacer una transcripción en directo del discurso. Estos dispositivos de estenotipia se desarrollan mediante un teclado de reducido número de teclas con los que se generan pulsaciones de sílabas o palabras completas (Fuentes et al., 2007), aumentando notablemente la velocidad.

Ilustración 5. Dispositivo de estenotipia

En la actualidad, con la evolución de y mejora de las prestaciones de los sistemas automáticos de reconocimiento del habla, se empiezan a utilizar tecnologías de transcripción voz a texto para generar on-line el texto transcrito (Cesya, 2014).

La subtitulación en diferido o pregrabada es el modo estándar de subtitulación y produce subtítulos de mayor calidad. Es el caso de las películas de televisión, cine, documentales. En este caso, los subtítulos se preparan de antemano con códigos de tiempo precisos de entrada y salida sincronizados con el audio. Existen distintas herramientas software específicas que reducen el tiempo de preparación del subtitulado, facilitando tareas como la edición de los archivos de vídeo, localización de las tramas de audio, conversión automática del audio a texto, edición de los subtítulos, o previsualización. Las normas de subtitulado accesible sólo pueden cumplirse con las tecnologías en diferido.

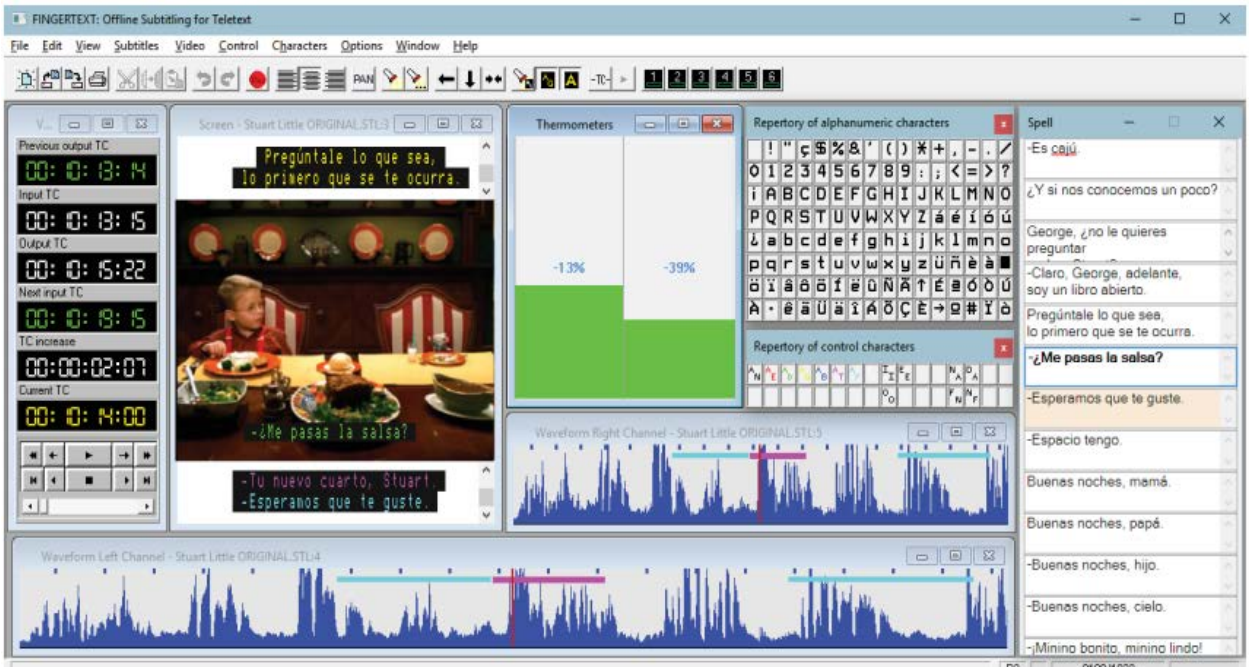


Ilustración 6. Editor de Subtitulación diferida [2]

El trabajo de subtitulación requiere una formación específica para desarrollar la capacidad de escribir de acuerdo con los requisitos del autor, y de acuerdo con las necesidades de las personas con discapacidad auditiva. Por ejemplo, en el caso del subtitulado de la música, es el subtitulador quien debe decidir cuándo es importante su subtitulado, y decidir la sensación que el autor ha querido transmitir con la música de la forma más objetiva posible. La definición de las competencias necesarias para definir el perfil profesional de subtitulador (Díaz, 2006) es un tema actual de debate, existiendo ya una normativa que regula los aspectos básicos de los cursos de especialización en subtitulación (BOE, 2019).

El subtitulado accesible, aun siendo la herramienta asistiva de referencia para la discapacidad auditiva, presenta algunas limitaciones. La presencia de subtítulos hace que la atención se desvíe de la imagen, por lo que se produce una reducción en el nivel de información de video asimilada, aunque se ha comprobado que, a cambio, los subtítulos proporcionan un mayor nivel de contexto (Gulliver & Ghinea, 2003). Otra limitación es la exigencia de una buena capacidad de comprensión lectora. Por ejemplo, se ha estudiado como una simplificación del lenguaje en los subtítulos mejoraba la comprensión en grupos de escolares con discapacidad auditiva, mientras que frustra a los usuarios adultos (Gulliver & Ghinea, 2003). Dentro de la discapacidad auditiva, se distinguen usuarios prelocutivos, en los que la sordera es anterior a la adquisición del lenguaje oral, y usuarios postlocutivos, en los que la sordera es posterior a la adquisición del lenguaje oral. Esta diferenciación categoriza a los usuarios del subtitulado como oralistas o signantes. Los oralistas son usuarios postlocutivos que han aprendido el lenguaje oral como lenguaje de comunicación, y los usuarios signantes, con sordera prelocutiva, han aprendido el lenguaje de signos (LLSS) para

[2] www.anglatecnic.com

comunicarse, y pueden preferir la comunicación por LLSS, aunque los subtítulos mejoran la comprensión cuando se añaden al subtitulado LLSS (Debevc, Milošević, & Kožuh, 2015).

Otra carencia importante se presenta en cuanto al subtitulado de la música de películas. La música es utilizada ampliamente en el cine como soporte a la narrativa por su capacidad de generar emoción (Donnelly, 2005), y se compone especialmente para una película por su poder de transmitir señales emocionales potentes (Eerola & Vuoskoski, 2011).

La música transmite de manera inmediata la tonalidad emocional de una escena: alegría, tristeza, miedo, informando así sobre el desarrollo de la acción dramática y completando la información verbal y suprasegmental. (Thompson, Russo, & Sinclair, 1994) comprobaron cómo modificando las notas finales de fragmentos musicales podía alterarse la sensación de conclusión de secuencias cinematográficas. (Pehrs et al., 2014) estudiaron la actividad cerebral de sujetos mientras veían escenas de besos de comedias románticas acompañadas de música triste, alegre, o sin música, observando que el tipo de música modulaba la emoción generada alterando la información procesada visualmente. En (Punset, 2011), se muestra como una misma escena puede interpretarse como una escena de seducción o como una escena de miedo en función de la música con la que se acompaña.

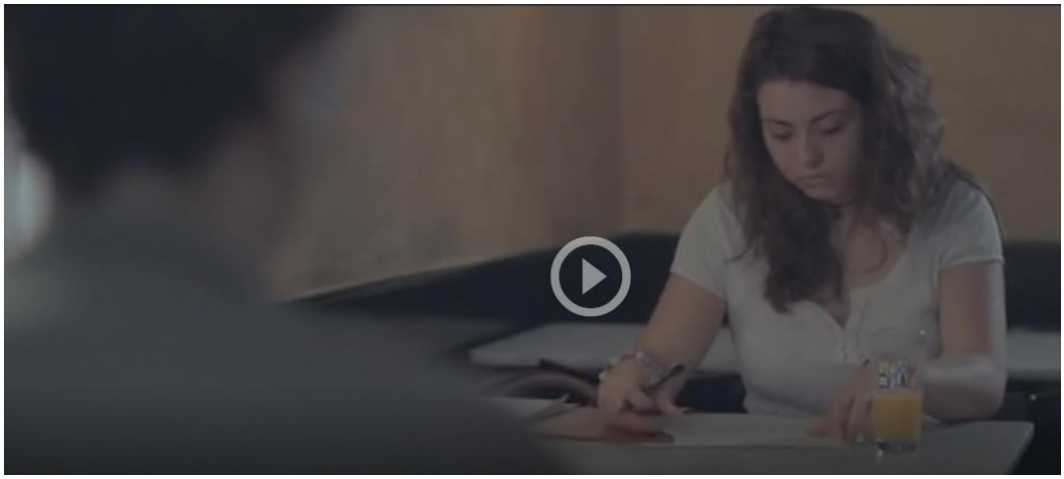


Ilustración 7. Ejemplo de escena utilizada en (Punset, 2011)

Pero las personas con discapacidad auditiva no pueden acceder a esta información de manera completa. El subtitulado textual que establece la normativa en el caso de la música no incluye esta información emocional. Es más, la representación visual de la música mediante un texto requiere un proceso cognitivo de atención consciente y selectiva para su lectura (Gulliver & Ghinea, 2003), lejos de la respuesta emocional inmediata que genera la música.

La predisposición emocional a la obra audiovisual que proporciona la música no es pues compartida por las personas sordas, y parcialmente por las personas con discapacidad auditiva que no se apoyan en dispositivos como audífonos o implantes cocleares en la escucha de las obras audiovisuales, incluso cuando la obra presenta un subtitulado accesible de calidad. Para mejorar la accesibilidad a los medios audiovisuales, se debería incluir esta

transmisión de la información contenida en la emoción musical por un canal alternativo al canal textual.

El canal vibro táctil ya se está explorando en otros ámbitos para facilitar la accesibilidad a la experiencia musical a las personas con discapacidad auditiva. La base es el diseño de dispositivos vibro táctiles que reproducen de forma simplificada los cambios de energía de las vibraciones sonoras. Así, para mejorar la calidad de experiencias como la asistencia a conciertos se han diseñado chalecos y sillones vibro táctiles (Jack, McPherson, & Stockman, 2015), o para facilitar la práctica en conjuntos musicales, se ha planteado el uso de pedales que transmiten a través de los pies las vibraciones del resto de instrumentos, permitiendo seguir el ritmo del conjunto, al tiempo que las manos quedan libres para tocar el instrumento musical (Hopkins, Maté-Cid, Fulford, Seiffert, & Ginsborg, 2016). Este canal vibro táctil podría ser la base para el subtitulado alternativo de la música en los medios audiovisuales.

1.2 Objetivos

El objetivo general de esta investigación es facilitar a las personas con discapacidad auditiva el acceso a la información emocional transmitida por la música de películas, explorando alternativas de subtitulado basadas en la transmisión vibro táctil, que ayuden a "sentir" esta emoción de forma directa e inmediata. A partir de este objetivo general, se plantean los siguientes objetivos específicos.

En primer lugar, al ser un campo de investigación nuevo, se necesita establecer un marco científico en el que apoyar la investigación, unos puntos básicos de los que partir. ¿Qué es la emoción? ¿Por qué la música emociona? Estas son algunas de las preguntas que se pretende abordar desde el estudio del estado del arte. Asimismo, sería importante conocer los tipos de emociones básicas que son perceptibles por los usuarios, y las posibilidades de clasificar automáticamente la emoción que una música cualquiera puede transmitir de manera mayoritaria. Se quiere abordar también el conocimiento científico sobre la percepción vibro-táctil, base de los dispositivos que ya se han empezado a utilizar. ¿Cuáles son las características de la percepción vibro táctil? ¿Qué tipo de vibraciones son perceptibles táctilmente? ¿Podemos transmitir parámetros musicales de forma vibro táctil?

En segundo lugar, se propone desarrollar experimentaciones que permitan aportar ideas base para el desarrollo de un framework de subtitulado accesible de la música de películas, que pueda incorporarse en un futuro a las tecnologías de subtitulado. Este framework estaría compuesto por dos funcionalidades: una funcionalidad de extracción de la emoción musical mediante la clasificación automática de los distintos fragmentos audio de la música, y una funcionalidad de transmisión de estas emociones a través del canal vibro táctil, estableciendo los correspondientes parámetros vibro táctiles.

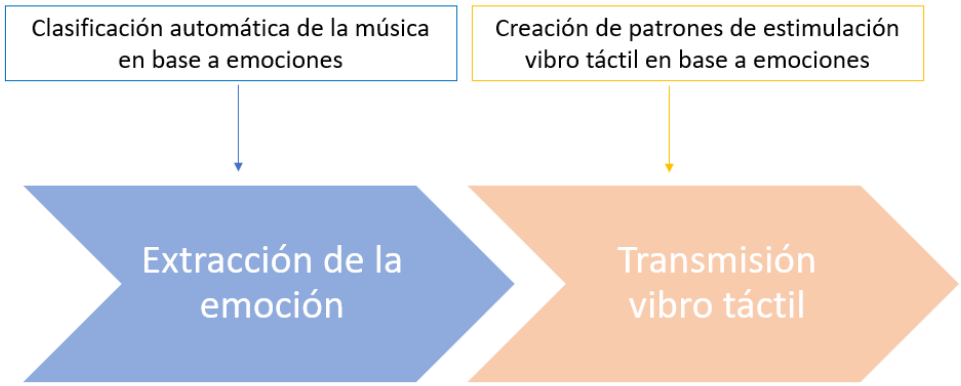


Ilustración 8. Framework básico para subtitulado accesible

Con este fin la propuesta de investigación incluye: la evaluación del canal vibro táctil como posible canal alternativo de transmisión de la emoción musical, y la evaluación de distintas estrategias de reconocimiento automático de emoción en la música.

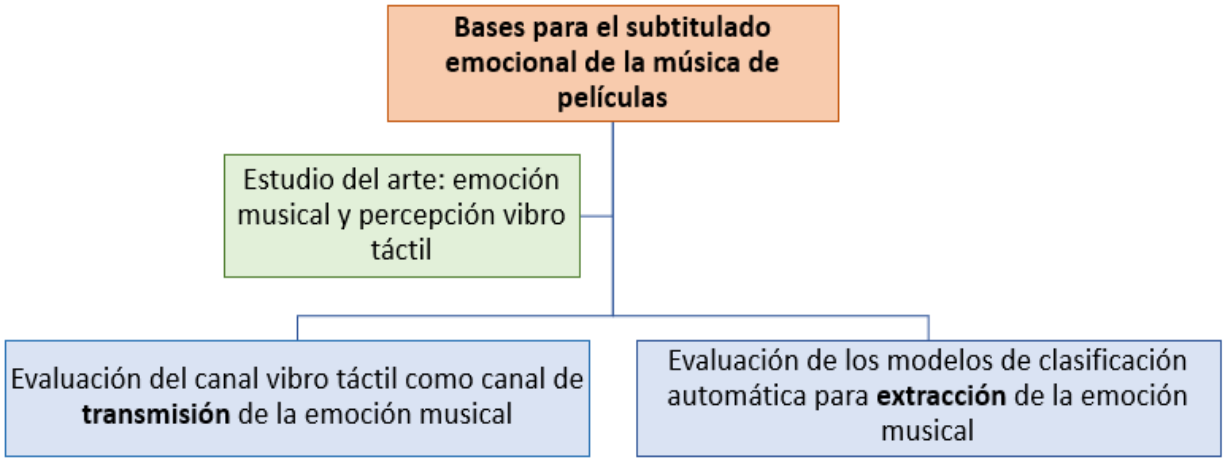


Ilustración 9. Propuesta de investigación

Como objetivo adicional, se considera el desarrollo de esta investigación desde un enfoque multidisciplinar integrando Ciencia Informática, Neurociencia, Teoría Musical y Electrónica.

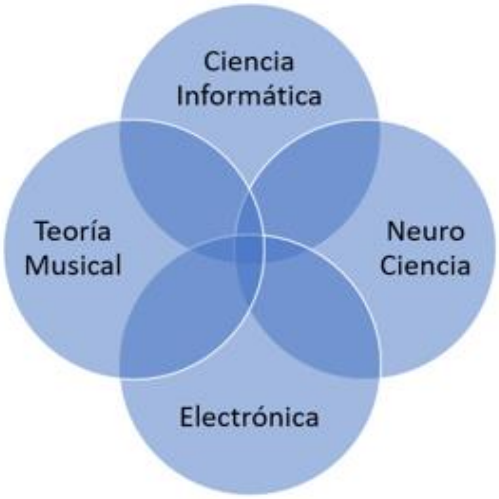


Ilustración 10. Enfoque multidisciplinar

Esta tesis se ha desarrollado en el marco del Centro Español del Subtitulado y la Audiodescripción (Cesya), centro asesor del Real Patronato sobre Discapacidad, en colaboración con los grupos de investigación de la Universidad Carlos III: Human Language and Accessibility Technologies (Hulat) y Grupo de Displays y Aplicaciones Fotónicas (Gdaf), y con el departamento de Psiquiatría de la Universidad Complutense de Madrid.

Ilustración 11. Marco de desarrollo de esta tesis

1.3 Hipótesis

Las hipótesis que se establecen para esta investigación son las siguientes.

- **Hipótesis 1:**

  El subtitulado accesible textual no transmite la información que aporta la música de forma inmediata a través de la emoción.

- **Hipótesis 2:**

  El tacto puede ser un canal de transmisión alternativo de emociones musicales básicas.

- **Hipótesis 3:**

  Los parámetros musicales pueden transmitirse de forma similar, aunque más limitada, mediante estimulación vibro-táctil

- **Hipótesis 4:**

  Los modelos de aprendizaje CNN permiten clasificar emociones básicas (alegría, tristeza, miedo) en fragmentos breves de música de película.

1.4 Estructura del documento

Este documento se estructura de acuerdo con los siguientes capítulos.

En el ***Capítulo 1****. Introducción* se explica la necesidad y motivación origen de esta investigación y se detallan los objetivos e hipótesis de partida.

En el ***Capítulo 2****. Estado del Arte* se lleva a cabo una revisión de la literatura científica en los dos ámbitos principales de interés de esta investigación. Por un lado, respecto a la emoción musical, se estudian los modelos científicos que desde la neurociencia explican el origen de la emoción y la relación de la música con la emoción, y los modelos que desde la ciencia informática se han desarrollado con el fin de extraer características musicales de las muestras de audio y determinar las emociones asociadas a estas características. Por otro lado, respecto a la percepción vibro táctil, se estudian sus bases fisiológicas, las analogías e interacciones entre los sentidos del oído y del tacto, la capacidad de percepción táctil de parámetros musicales, y distintos dispositivos que se han desarrollado recientemente para la transmisión de estos parámetros.

En el ***Capítulo 3.*** *El canal táctil como canal de transmisión* se detallan las experimentaciones realizadas, basadas en el estudio de la actividad cerebral mediante técnicas de electroencefalograma en participantes con y sin discapacidad auditiva, durante la presentación de materiales audiovisuales en distintas condiciones de audio, subtitulación, y estimulación

vibro táctil, con el fin de comparar las distintas condiciones y determinar la capacidad del canal vibro táctil como posible canal alternativo de transmisión de la emoción musical.

En el ***Capítulo 4****. Modelos CNN para extracción de la emoción musical* se detallan las experimentaciones realizadas con distintas estrategias de clasificación automática de la música basadas en redes CNN (Convolutional Neural Network), con el fin de definir un modelo de clasificación de fragmentos breves de audio en base a emociones básicas como alegría, tristeza, miedo, adecuado para la subtitulación automática de música de películas.

En el ***Capítulo 5.*** *Conclusiones y trabajos futuros* se establecen las conclusiones a partir de los objetivos e hipótesis establecidos, y de los resultados de las experimentaciones realizadas, y se proponen nuevas líneas de investigación para el desarrollo de un subtitulado accesible de la música de películas.

En el ***Capítulo 6.*** se relaciona la bibliografía consultada y de referencia para elaborar la investigación.

2 ESTADO DEL ARTE

2.1 Introducción

En este capítulo se propone abordar el primer objetivo de esta investigación, que es profundizar en el conocimiento de los campos principales de interés que van a permitir crear un marco científico de partida donde sustentar esta investigación. Por un lado, entender la emoción musical: ¿qué es la emoción? ¿por qué la música emociona? desde la perspectiva de la neurociencia y desde la ciencia informática. Por otro lado, estudiar un canal alternativo de transmisión de la información sonora y entender la percepción vibro-táctil, base de los dispositivos que ya se han empezado a utilizar con las personas con discapacidad auditiva para potenciar la experiencia musical, desde la perspectiva de la fisiología y la neurociencia, y también desde una perspectiva más técnica de desarrollo de dispositivos vibro táctiles.

La sección 2.2 se centra en la emoción musical, la sección 2.3 en la percepción vibro táctil. En la última sección, sección 2.4, se analizan los distintos modelos que desde la ciencia informática se han desarrollado en los ámbitos MIR (Music Information Retrieval) y MER (Music Emotion Recognition).

Con la información y limitaciones recogidas del análisis del estado del arte, se pretende construir las bases de un sistema completo de clasificación, transmisión y percepción de la emoción de la música para dar respuesta a las necesitades de los usuarios sordos y potenciar la emoción en el resto de los usuarios. De esta forma se cubrirán los objetivos de la presente investigación descritos en el capítulo 1 de la memoria.

2.2 Música y emoción: modelos neurocientíficos

La música es uno de los desencadenantes más efectivos de experiencias emocionales fuertes, siendo esta una de sus características definitorias (Gabrielsson, A., 2001) (Lonsdale & North, 2011). En este apartado se revisarán los diferentes modelos que se han desarrollado desde el punto de vista neurocientífico para explicar la emoción y en particular la emoción musical.

2.2.1 Modelos científicos a la emoción

La investigación científica de la emoción presenta muchas dificultades por la subjetividad inherente a la percepción de la emoción. Es un campo de investigación que se ha iniciado en décadas relativamente recientes y en el que queda mucho por conocer. Desde finales del siglo pasado, las investigaciones se han desarrollado en base a dos paradigmas básicos: el modelo categórico y el modelo dimensional de la emoción (Eerola & Vuoskoski, 2011). Ambos paradigmas consideran que las emociones son las representaciones subjetivas de circuitos y funciones neuronales que han evolucionado a partir de circuitos primarios fundamentales para la supervivencia, que se activarían priorizando eventos en un entorno complejo (Oatley & Johnson-laird, 1987), aunque difieren en la identificación de estos circuitos.

2.2.1.1 Modelo categórico de la emoción

El modelo categórico de la emoción presupone la existencia de un número limitado de emociones básicas, innatas y universales. Estas emociones básicas serían independientes unas de las otras y estarían asociadas a la activación de un circuito neuronal propio dentro del sistema nervioso central (Posner, Russell, & Peterson, 2005).

El modelo de Ekman (Ekman, 1992) es el más conocido. Considera como emociones básicas la alegría, la tristeza, el miedo, la ira, la sorpresa, y el asco. Ekman realizó estudios con muchas fotografías determinando que el rostro de las emociones básicas es universal, siendo análogo en cualquier cultura y raza. El resto de las emociones serían secundarias y podrían derivarse de estas, y su expresión facial ya no sería universal. A menor escala se han realizado estudios transculturales sobre el reconocimiento de las emociones básicas de alegría, tristeza, miedo e ira en expresiones vocales (Scherer, Banse, & Wallbott, 2001), o en extractos musicales (Fritz et al., 2009) (Balkwill & Thompson, 1999) (Balkwill, Thompson, & Matsunaga, 2004), comprobándose que los sujetos experimentales eran sensibles a la emoción transmitida en expresiones vocales y en extractos musicales de diferentes culturas. Los hallazgos sugieren que los oyentes son sensibles a las emociones en la música tanto familiar como desconocida, y esta sensibilidad está asociada con la percepción de señales acústicas que trascienden las fronteras culturales.

Ilustración 12. Fotos de expresiones faciales utilizadas por Ekman (Petisco, 2015)

En apoyo de este paradigma existen estudios de neuroimagen cerebral funcional que muestran por ejemplo cómo el reconocimiento de expresiones faciales de miedo parece ligado a sustratos neuronales específicos como la amígdala (Sprengelmeyer, Rausch, Eysel, & Przuntek, 1998) (Adolphs, Tranel, Damasio, & Damasio, 1994), mientras que el reconocimiento de expresiones faciales de asco estaría ligado a los ganglios basales y la ínsula anterior (Gray, Young, Barker, Curtis, & Gibson, 1997) (Sprengelmeyer et al., 1996).

Pero en general, los estudios fisiológicos y de neuroimagen funcional no han logrado establecer una evidencia clara para apoyar esta teoría (Lindquist, K. A., Wager, Kober, Bliss-Moreau, & Barrett, 2012).

2.2.1.2 Modelo dimensional de la emoción

El modelo dimensional de la emoción permite representar las emociones en un espacio continuo, generalmente de 2 o 3 dimensiones. En los últimos años, los modelos bidimensionales están prevaleciendo sobre el modelo categórico en las investigaciones científicas de la emoción (Eerola & Vuoskoski, 2011).

El modelo híbrido de "Circumplex model of affect" (Posner et al., 2005) propone que todos los estados afectivos surgen de interpretaciones cognitivas de las sensaciones neuronales centrales que son el producto de dos sistemas neurofisiológicos independientes: uno relacionado con la valencia y otro con la activación. La dimensión Valencia hace referencia a la negatividad/positividad de la emoción, y la dimensión Activación al nivel de activación/relajación producida por la emoción.

Desde la perspectiva de este modelo, las emociones discretas serían "etiquetas" psicológicas subjetivas que se pueden identificar con puntos de ese espacio continuo de dos dimensiones Valencia-Activación.

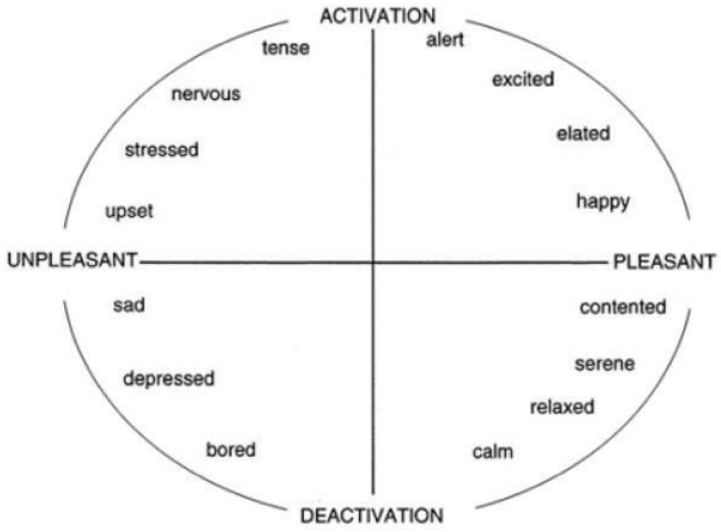


Ilustración 13. Circumplex Model of affect de Russell (Posner et al., 2005)

En apoyo de este modelo, algunos autores consideran que existen circuitos neuronales que se activan en función de la intensidad de los estímulos emocionales, independientemente de su valencia, mientras que otros circuitos de tipo hedónico se encargan de discriminar la valencia (Lang, Peter J. & Bradley, 2010)

(Thayer, 1989) también propone un modelo bidimensional, pero en base a las dimensiones de Energía (baja-alta) y Tensión (baja-alta), mientras que la valencia se podría explicar como una combinación de estas dos dimensiones.

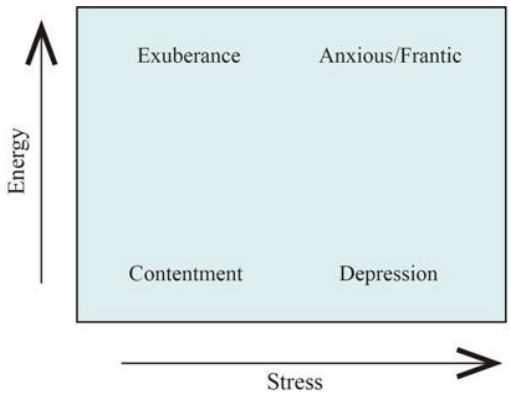


Ilustración 14. Modelo de Thayer (Seo & Huh, 2019)

En el modelo tridimensional (Eerola, Lartillot, & Toiviainen, 2009) (Lartillot, 2019) (Schimmack & Grob, 2000), se engloban las tres dimensiones: Valencia (positiva-negativa), Activación (baja-alta) y Tensión (baja-alta).

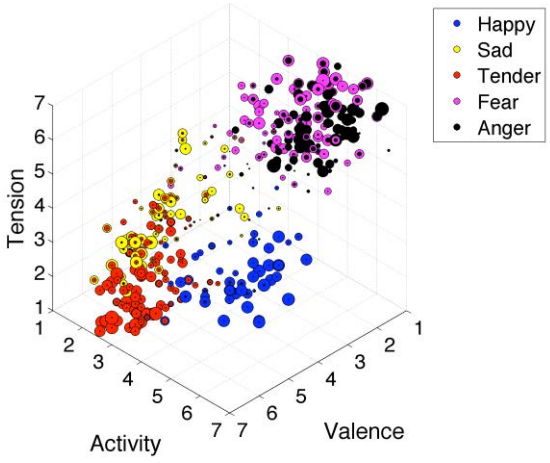


Ilustración 15. Modelo Valencia-Activación-Tensión (Eerola et al., 2009)

Aunque sea el modelo con mayor aceptación en los últimos años, el paradigma dimensional plantea distintos problemas. Como hemos visto, no hay acuerdo claro sobre las dimensiones a considerar. Tampoco está claro la linealidad de este espacio, ya que por ejemplo los eventos negativos producen respuestas más altas que los eventos neutrales o positivos (Taylor, 1991).

2.2.2 Medir la emoción

Los estudios científicos de la emoción se encuentran con la dificultad de la medida de la emoción, debido a la subjetividad de la experiencia emocional inherentes a cada una de las personas. Generalmente la experimentación científica típica en este ámbito se basa en la presentación a los sujetos experimentales de una serie de estímulos emocionales controlados, y en la medida de las reacciones emocionales que esos estímulos producen.

2.2.2.1 Estímulos emocionales

Los estímulos más utilizados son de tipo visual (imágenes), audiovisuales, y auditivos (Al-Nafjan, Hosny, Al-Ohali, & Al-Wabil, 2017). Las imágenes producen fuertes efectos emocionales (Ito, Cacioppo, & Lang, 1998), generando mayor respuesta emocional las imágenes simples que las imágenes con escenas complejas (Lang & Bradley, 2010) (Bradley, Margaret M., Hamby, Löw, & Lang, 2007). La música se utiliza también ampliamente, por su capacidad para provocar emociones, que es una de sus características definitorias (Gabrielsson, Alf & Juslin, 2003), y por su transculturalidad (Balkwill & Thompson, 1999) (Molnar-Szakacs & Overy, 2006). También son muy utilizados estímulos audiovisuales, como pequeños videoclips. Aquí es importante tener en cuenta el predominio de la imagen sobre el sonido (Brown & Cavanagh, 2017), en caso de ambigüedad o conflicto emocional entre imagen y sonido. Las palabras y los estímulos lingüísticos también se utilizan con frecuencia como desencadenantes emocionales (Thomas & LaBar, 2005) (Zhang, M., Ge, Kang, Guo, & Peng, 2018).

Para facilitar la reproducción, comparación y contraste de teorías y resultados, algunos autores y centros de investigación han desarrollado bases de datos con estímulos emocionales normalizados, validados con estudios experimentales y disponibles públicamente. Estas bases de datos incluyen palabras, imágenes, sonidos y combinaciones de estos, y las correspondientes etiquetas de acuerdo con uno o diferentes paradigmas. Algunos ejemplos son la base de datos de rostros de Ekman, Pictures of Facial Affect[3], compuesta por 110 imágenes de expresiones faciales que expresan emociones básicas, Surrey Audio-Visual Expressed Emotion[4] compuesta por clips audiovisuales con actores masculinos expresando distintas emociones, International Affective Picture System, conjunto de 12 series de 60 fotografías en color que varían en términos de valencia y activación, y representan una amplia gama de categorías semánticas (Lang, P., Bradley, & Cuthbert, 1997) , o las voces afectivas de Montreal (MAV), un conjunto de interjecciones vocales breves que expresan ira, disgusto, miedo, dolor, tristeza, sorpresa, felicidad, placer sensual y neutralidad (Belin, Fillion-Bilodeau, & Gosselin, 2008). También existen bases de datos musicales como las que se detallan en la sección 2.2.5.2.

[3] https://www.paulekman.com/product/pictures-of-facial-affect-pofa/

[4] http://kahlan.eps.surrey.ac.uk/savee/

2.2.2.2 Cuestionarios

Para medir las reacciones a estos estímulos, un instrumento de medida muy utilizado son los cuestionarios que se aplican tras la presentación del estímulo y que los sujetos deben rellenar para describir su reacción emocional. El formato del cuestionario queda determinado por el tipo de modelo teórico (Yang, X., Dong, & Li, 2018) (Yang, Y. & Chen, 2012).

En la experimentación basada en el modelo categórico, los cuestionarios de elección forzada suelen mostrar una selección de opciones con etiquetas correspondientes a unas pocas emociones, o dejan un formato libre en el que el sujeto debe escribir una o varias etiquetas que describan su reacción emocional. Este formato presenta varios problemas, como la granularidad de las etiquetas elegidas para describir las emociones percibidas por los sujetos. ¿Cuántas etiquetas se deben incluir en los cuestionarios? Si se eligen pocas etiquetas, pueden ser insuficientes para describir las reacciones emocionales producidas por los estímulos, o, por el contrario, si se definen demasiadas etiquetas pueden generar resultados confusos, ya que distintas y variadas etiquetas pueden no significar lo mismo para todos los sujetos.

Ilustración 16. Ejemplo de cuestionario de elección forzada (modelo categórico)

De acuerdo con el modelo dimensional, los cuestionarios suelen presentar escalas de tipo ordinal (Market, 2020) para cada una de las dimensiones del modelo, y el sujeto experimental debe determinar un valor en cada escala para describir su reacción emocional. Este formato tampoco está exento de problemas. Por ejemplo, en una escala de 1 a 5: ¿significa un 4 lo mismo para todos los sujetos? ¿Está bien definida la escala, es la distancia entre 0 y 1 equivalente a la distancia entre 4 y 5?

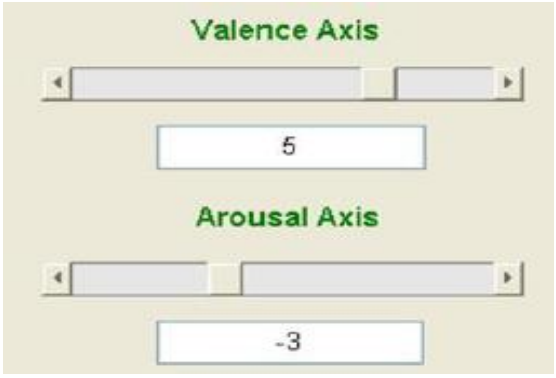


Ilustración 17. Ejemplos de cuestionario basado en escalas (modelo dimensional)

Tampoco está exento de problemas el entorno experimental, ya que se suelen utilizar situaciones controladas en las que se presenta un determinado tipo de estímulos. Estas situaciones distan mucho de las interacciones reales en las que múltiples estímulos externos e internos modulan nuestras reacciones emocionales.

2.2.2.3 Medidas EEG y fMRI

Con el desarrollo de las técnicas de electroencefalografía EEG (Farnsworth, 2019) y de resonancia magnética funcional fMRI (Glover, 2011), se han podido establecer medidas objetivas de la emoción a partir de la medida de la activación cerebral. La técnica EEG se desarrolló 1929, por el psiquiatra alemán Hans Berger, pero hasta mediados de los años 70 no empezaron a extenderse los estudios sobre la emoción basados en medidas EEG. La resonancia magnética funcional se desarrolló en la década de 1980. Desde entonces, la misma configuración experimental básica, como en (Petersen, Fox, Posner, Mintun, & Raichle, 1988), se ha replicado en la investigación hasta hoy: el sujeto experimental se conecta a un equipo EEG o fMRI, y se somete a la presentación de distintos estímulos emocionales mientras se registra su actividad cerebral.

El EEG se basa en la evidencia de que grupos masivos de neuronas se activan al mismo tiempo cuando trabajan sincronizados, produciendo pequeños cambios de voltaje a su alrededor (en el rango de milivoltios a microvoltios). El EEG mide de forma continua estos cambios directamente en la superficie del cerebro a través del cuero cabelludo. Normalmente se define una configuración estándar de la posición de los electrodos para facilitar la replicabilidad y comparativa de los estudios. Entre las ventajas más importantes de esta técnica se encuentran su precio, la posibilidad de usar equipos portátiles, y que se trata de una técnica no invasiva y segura cuando se aplica en el cuero cabelludo (Farnsworth, 2019).

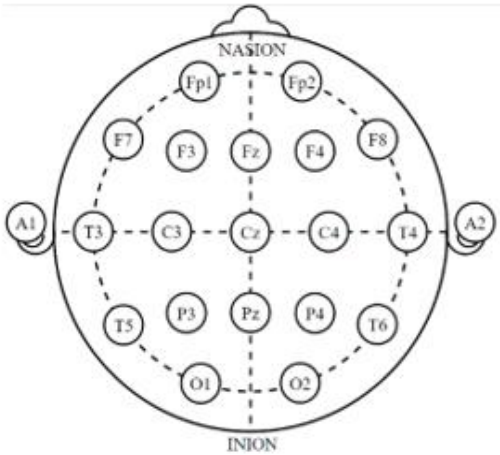


Ilustración 18. Configuración de electrodos 10-20 International

El EEG presenta una alta resolución temporal, de milisegundos, detectando cambios muy rápidos en los potenciales del cuero cabelludo. Permite el análisis de frecuencia, separar las distintas bandas de ondas cerebrales (delta, theta, alfa, beta, gamma), la medida del potencial espectral (la cantidad de energía en cada banda), y obtener información de fase y frecuencias instantáneas.

La primera información representativa que se encontró con el EEG fueron los llamados potenciales relacionados con eventos (ERP). Son señales que se producen como reacción a un estímulo, y que típicamente aparecen en un intervalo entre unos pocos cientos de milisegundos a varios segundos, como el denominado P300, pico positivo que aparece unos 300 ms después del inicio del estímulo.

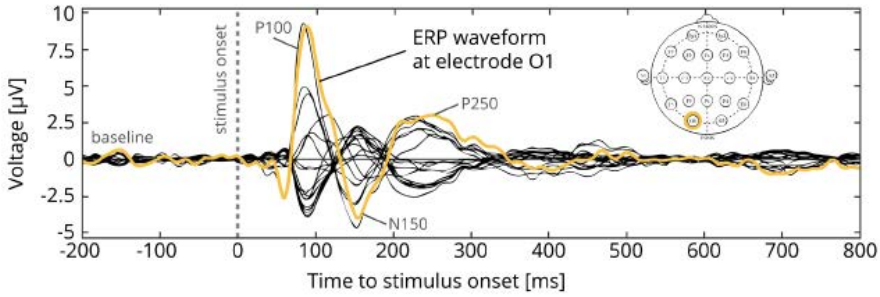


Ilustración 19. Señal típica EEG tras presentación de un estímulo (Farnsworth, 2019)

La contrapartida principal de esta técnica es que el registro de las señales en la superficie del cerebro (si no se colocan electrodos dentro del cerebro) limita la medición de fuentes internas de activación. Para solventar en parte esta limitación, se han desarrollado distintas técnicas basadas en algoritmos complejos que permiten recrear las fuentes internas a partir de su huella en el voltaje del cuero cabelludo, como la tomografía electromagnética de baja resolución denominada LORETA (Pascual-Marqui, Michel, & Lehmann, 1994), ampliamente utilizada. LORETA computa una distribución tridimensional, basada en una rejilla de 2394 vóxeles (unidades volumétricas de información) de 7x7x7 mm, de las fuentes de actividad neuronal compatible con la actividad registrada a nivel superficial. A partir de estos algoritmos se pueden generar mapas paramétricos estadísticos (SPM), como los de la Ilustración **20**, que muestran las zonas de activación diferencial detectadas en el cerebro.

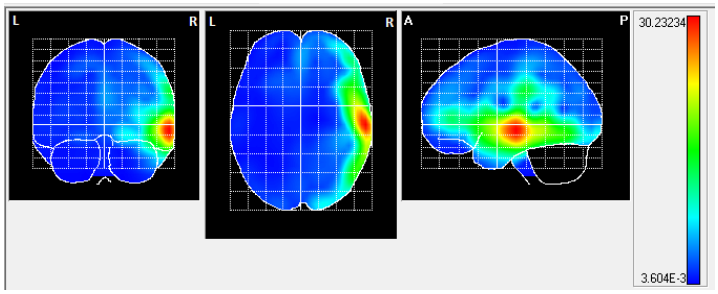


Ilustración 20. Mapas de activación cerebral (SPM) obtenidos con EEG

Otra contrapartida de los registros de EEG es que presentan los denominados artefactos (interferencias en la señal producidas por redes eléctricas, iluminación, movimientos musculares del sujeto experimental...) que deben eliminarse o filtrarse antes de poder interpretar los registros. Aunque se han propuesto algunos enfoques automáticos, los investigadores normalmente deben inspeccionar visualmente toda la señal registrada, y eliminar manualmente los artefactos detectados, lo que alarga el tiempo de procesado de las señales.

La resonancia magnética funcional (fMRI) detecta las activaciones diferenciales de las distintas zonas cerebrales (Cohen et al., 2008) midiendo las diferencias en la oxigenación de la sangre que fluye a través del cerebro (la oxihemoglobina y la desoxihemoglobina tienen diferente susceptibilidad magnética), diferencias que están correlacionadas con la activación neuronal. Estos datos se procesan estadísticamente para generar una representación significativa de estas diferencias, en los mapas paramétricos estadísticos (SPM), dando lugar a imágenes como las que se muestran en la Ilustración 21.

Las técnicas fMRI tienen las limitaciones temporales de los procesos fisiológicos en los que se basan (Lindquist, M. A., 2008), por lo que su resolución temporal es más pobre, en el rango de segundos. Sin embargo, la resolución espacial y la tridimensionalidad real obtenida permite generar un mapa de vóxeles de muy pocos mm^3 si no se requiere una alta resolución temporal, ya que hay un compromiso en esta técnica entre estos dos parámetros: por ejemplo, para un tamaño de vóxel de 3x3x5 mm3, la frecuencia de muestreo decae a aproximadamente 2s (Lindquist, 2008).

La fMRI requiere un imán masivo (típicamente alrededor de unos pocos teslas), lo que hace que la configuración experimental requiera mucho espacio y sea costosa.

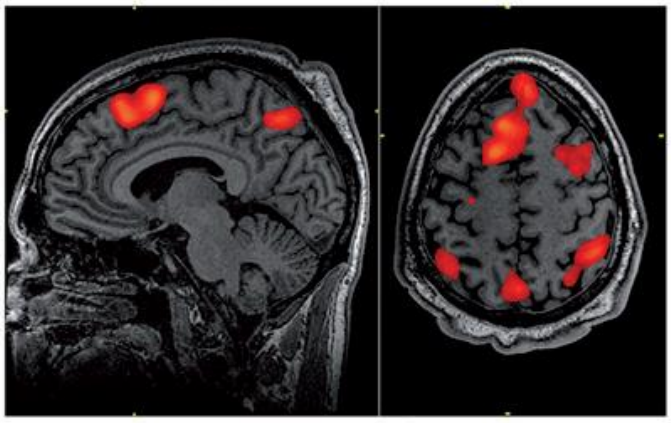


Ilustración 21. Mapas de activación cerebral (SPM) obtenidos con fMRI (G.Konstantina, CC BY-SA 4.0, via Wikimedia commons)

Dado que tanto el EEG como el fRMI se basan en procesos fisiológicos relacionados, es fácil encontrar correlaciones entre ambas medidas. Estas dos técnicas no invasivas de exploración del interior del cerebro funcional no son mutuamente excluyentes y ambas se utilizan ampliamente en la investigación neurocientífica en la actualidad.

2.2.3 Bases de datos científicas de música

Uno de los problemas que se encuentran los investigadores en los estudios sobre la emoción musical es decidir con qué estímulos musicales trabajar. Según (Eerola & Vuoskoski, 2011), la mayoría de los estudios han utilizado fragmentos de piezas musicales clásicas occidentales relativamente conocidas que han sido elegidas arbitrariamente por los investigadores, aunque también se han empleado extractos de música popular, música de películas o música sintética compuesta especialmente para la tarea de investigación en cuestión. Emplear músicas conocidas puede tener efectos no deseados si los participantes están familiarizados con ellas, mientras que los estímulos sintéticos están libres de estos problemas, pero suelen sonar de forma artificial y carecen de algunas de las características como el timbre importante para producir emoción. Otro problema es el número de ejemplos musicales relativamente bajo en comparación con los conjuntos de estímulos utilizados por los investigadores de la emoción en otros campos, como el dominio visual. Por ejemplo, el ya mencionado International Affective Picture System utiliza 12 series de 60 imágenes cada una (Lang et al., 1997). De aquí que algunos investigadores se hayan planteado la creación de unas bases de datos con elementos musicales específicamente desarrolladas para el estudio de la percepción de las emociones musicales desde distintos ámbitos científicos (neurociencia, psicopatología, etc.).

Se comentan a continuación las bases de datos desarrolladas por (Vieillard et al., 2008), (Eerola & Vuoskoski, 2011) y (Paquette, Peretz, & Belin, 2013), que son las más mencionadas como bases de datos validadas en el ámbito de la investigación de la emoción musical, y en particular son de nuestro interés por basarse las dos primeras en música de películas.

2.2.3.1 Musical excerpts for research on emotions (Vieillard et al., 2008)

(Vieillard et al., 2008) consideraron que no existía un set de estímulos musicales validado rigurosamente para la investigación de la emoción y crearon un set de estímulos con este objetivo.

Para ello seleccionaron las cuatro categorías de emociones: felicidad, tristeza, miedo y tranquilidad, que se reconocen con más precisión (Juslin, Patrik N. & Laukka, 2003) e inmediatez (Vieillard et al., 2008) (Peretz, Gagnon, & Bouchard, 1998) en la música, y que además pueden clasificarse en un espacio a lo largo de las dimensiones de valencia y activación.

La tranquilidad no se considera una emoción básica, pero se incluyó para contrastar con la amenaza, de manera similar a como la felicidad contrasta con la tristeza. La etiqueta de miedo de Ekman, que los autores consideraron ambigua en el caso de la música, se sustituyó, en los cuestionarios a los sujetos experimentales, por 'aterrador' para englobar en tanto los casos en los que la música expresa miedo como los casos en los que genera miedo.

El material inicial consistía en 56 fragmentos musicales (14 fragmentos para cada una de las cuatro emociones consideradas), de entre 9 y 16 segundos de duración, basados en una melodía con acompañamiento. Estos fragmentos fueron compuestos, específicamente para el estudio, en el género de la música de cine y siguiendo las reglas del sistema tonal occidental.

Los fragmentos alegres se escribieron en modo mayor, en un rango de metrónomo de 92-196 pulsos por minuto (PPM), y con línea melódica en el rango de tono medio-alto. Los fragmentos tristes se escribieron en modo menor a un tempo medio lento de 40-60 PPM. Los fragmentos tranquilos se compusieron en modo mayor, un tempo intermedio 54-100 PPM, con acompañamiento arpegiado. Los fragmentos de miedo fueron compuestos con acordes menores en el tercer y sexto grado, y ritmos variados de 44 a 172 PPM.

Para la evaluación de estos fragmentos realizaron tres experimentos. En el primer experimento se pidió a un grupo de sujetos experimentales que valoraran en una escala de "0 a 9" en qué grado reconocían las emociones de alegría/tristeza/miedo/tranquilidad. Otro grupo debía evaluar en una escala de 0 a 9 si los fragmentos eran placenteros-displacenteros, y en otra escala de 0 a 9 si eran relajantes-estimulantes. En el segundo experimento, se determinaba el número de eventos musicales necesarios para el correcto reconocimiento de la emoción. Finalmente, en el tercer experimento, se pedía a los oyentes que decidieran, entre pares de fragmentos musicales, en qué medida eran emocionalmente diferentes, para llegar a una clasificación de los fragmentos sin etiquetas verbales que pudieran influenciar la clasificación.

Como resultado de estos experimentos redujeron la muestra a 40 fragmentos musicales, 10 para cada tipo de emoción, alegría, tristeza, amenaza y tranquilidad, con tasas de reconocimientos de respectivamente 99%, 84%, 72% y 94%. Como conclusión, estos autores consideran que este set de 40 fragmentos musicales puede ser adecuado para la investigación sobre la emoción musical.

2.2.3.2 110 Film Music excerpts (Eerola & Vuoskoski, 2011)

En sus estudios, (Eerola & Vuoskoski, 2011) partieron del mismo objetivo de elaborar una base de datos musical, validada experimentalmente. Con el fin de incluir un material más realista, además de ejemplos muy característicos de las categorías emocionales básicas, decidieron incluir también ejemplos intermedios no atribuibles a una única categoría.

Para ello, un panel formado por 12 de expertos musicólogos seleccionaron 360 extractos musicales de 60 bandas sonoras de películas de temática variada (románticas, de ciencia ficción, de terror, de acción, de comedia y de drama), y de las décadas (1976-2006) para mantener una calidad sonora homogénea. Los autores decidieron utilizar música de películas por ser una música compuesta con el fin de transmitir estímulos emocionales potentes, y por ser un material relativamente "neutro" en términos de preferencias musicales y familiaridad.

La mitad de los expertos identificaron los extractos en base a seis emociones discretas (felicidad, tristeza, miedo, ira, sorpresa y ternura), elegidas por ser las más utilizadas en estudios anteriores sobre música y emoción (Juslin, P., 2000) (Kallinen, 2005) (Krumhansl, 1997). Para el modelo dimensional, los autores se basaron en el modelo tridimensional Activación-Energía-Tensión de (Schimmack & Grob, 2000), por lo que la otra mitad de expertos debía encontrar ejemplos de valencia positiva-negativa, tensión alta-baja y energía alta-baja. Para la dimensión de valencia los adjetivos utilizados eran: placentero-displacentero, bueno-malo y positivo-negativo. Para la dimensión de tensión eran: tenso-relajado, estresado-calmado y nervioso-descansado. Para la dimensión energía, los adjetivos eran: despierto-dormido, despierto-cansado y alerta-somnoliento.

Cada fragmento debía durar entre 10 y 30 segundos, y no contener diálogos ni efectos sonoros no musicales. Cada experto identificó una media de 30 fragmentos. El resultado fueron 360 clips de audio. A continuación, todo el panel de expertos evaluó los 360 extractos. La mitad de los expertos puntuaron de 1 a 7 la presencia de las 6 emociones discretas en los extractos. La otra mitad evaluó los extractos de 1 a 7 en las 3 dimensiones consideradas. Todos evaluaron de 1 a 3 si el extracto les resultaba familiar. Los resultados mostraron que las emociones objetivo se identificaban claramente salvo en el caso de la sorpresa, por lo que los extractos relativos a la emoción de sorpresa fueron descartados, así como los extractos que tuvieron una puntuación de 2 o 3 en familiaridad. Lo extractos se clasificaron a continuación en base a su tipicidad respecto a la emoción objetivo (teniendo en cuenta su puntuación en la emoción objetivo, y en las otras categorías emocionales).

Así se consideraron cinco mejores ejemplos de cada emoción discreta (felicidad, tristeza, ternura, ira y miedo) y cinco ejemplos "moderados", intermedios en la clasificación de tipicidad. En total se consideraron 50 ejemplos para las 5 emociones discretas (5 altas + 5 moderadas $\times$ 5 categorías). Así mismo se seleccionaron 60 fragmentos (20 por dimensión) representativos de la varianza sobre el espacio dimensional y no únicamente ejemplos de los extremos de cada dimensión. El resultado es un set de 110 fragmentos musicales clasificados de acuerdo con ambos modelos categórico y dimensional.

Hay que señalar que las mejores muestras de ira tienen una puntuación alta en la emoción de miedo también, lo que no ocurre a la inversa. Esto puede sugerir que la ira se confunde fácilmente con el miedo desde el punto de vista musical, ya que, como indica (Vieillard et al., 2008), en este caso el oyente puede confundir el miedo percibido con el miedo inducido a la hora de evaluar una música.

2.2.3.3 Musical Emotional Bursts (Paquette et al., 2013)

(Paquette et al., 2013) desarrollaron una base de datos de elementos musicales muy breves análoga a la base de datos Montreal Affective Voices (MAV). Esta última consiste en un conjunto validado experimentalmente de 90 vocalizaciones no verbales, muy breves, que representan distintas emociones básicas, para investigación en el procesamiento auditivo emocional.

En este caso los autores solicitaron a 10 violinistas y 10 clarinetistas profesionales que realizaran una serie de improvisaciones de alrededor de un segundo de duración representativas de las emociones de alegría, tristeza, miedo y una cuarta neutra. Los autores eligieron estas emociones por ser las más fáciles de reconocer en la música (Gabrielsson & Juslin, 2003) (Juslin & Laukka, 2003).

Preseleccionaron 120 extractos de cada instrumento y realizaron una prueba de validación con sesenta participantes. Veinte participantes evaluaron cada elemento en base a un cuestionario de selección forzosa entre alegría/tristeza/miedo/neutralidad ("Por favor, elija la emoción que cree que representa este estímulo entre las etiquetas de miedo, felicidad, tristeza y neutralidad"). Veinte participantes evaluaron para cada elemento la valencia percibida de la emoción expresada en una escala de "1 extremadamente negativa a 9 extremadamente positiva", y 20 participantes evaluaron para cada elemento la activación percibida de la emoción expresada en una escala de "1 nada activada a 9 extremadamente activada".

La mayor precisión se obtuvo en el violín para los estímulos que expresaban miedo y tristeza (88%) y en el clarinete para los que expresaban felicidad (92%).

Tras esta validación, redujeron las muestras a los 40 elementos de violín y los 40 elementos de clarinete más representativos de cada emoción (es decir categorizados en la emoción objetivo por el mayor número de participantes). Las muestras resultantes se denominaron Musical Emotional Burts (MEB).

Los autores consideran que los resultados con fragmentos musicales de máximo 2 segundos son bastante similares a los obtenidos por (Vieillard et al., 2008) que utilizaron estímulos musicales más largos y convencionales (inspirados en la música de películas), lo que sugiere que los MEB involucran procesos emocionales similares a los evocados por extractos de películas más elaborados.

DATASET	Elementos	Emociones etiquetadas	Resultados
MusicalExcerpts (Vieillard 2008)	56 fragmentos de 12-15 segundos Inspirados en música de película	happiness sadness threat peacefulness	- Modelo categórico: los sujetos identificaban con mucha precisión las cuatro emociones (alegría, tristeza, miedo, tranquilidad). - Modelo dimensional Valencia-Activación: los resultados eran más consistentes con un modelo con dimensiones Energía y Tensión.
Film Music excerpts (Eerola 2011)	110 fragmentos de 12-15 segundos Músicas de película	happiness sadness fear anger tenderness	- Puntuaciones muy consistentes en ambos modelos categórico y dimensional Activación-Energía-Tensión - Modelo dimensional más fiable en el caso de extractos musicales más ambiguos - El modelo tridimensional puede reducirse a uno bidimensional Valencia-Activación
Musical Emotional Bursts (Paquette, 2013)	80 fragmentos de 2-3 segundos Interpretación libre por músicos	happiness sadness fear	- Resultados bastante similares a los obtenidos por (Vieillard et al., 2008) - Consistentes con el modelo Valencia-Activación

Tabla 1. Tabla resumen de las características de las bases de datos

2.2.4 Música y los paradigmas de la emoción

Precisamente por su capacidad para emocionar, muchos estudios sobre las emociones han utilizado la música como estímulo para validar los dos modelos de la emoción, categórico y dimensional.

Desde el punto de vista categórico, varios estudios han mostrado que las emociones básicas de Ekman de alegría, tristeza y miedo se reconocen rápidamente en la música (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005) (Peretz et al., 1998) siendo las respuestas

muy consistentes y precisas entre distintos oyentes (Juslin & Laukka, 2003) (Bigand et al., 2005). Un reciente metaanálisis de 41 estudios sobre la interpretación de música muestra que la felicidad, la tristeza, la ira, la amenaza y la ternura son descodificadas con precisión por los oyentes (Juslin & Laukka, 2003), y que la alegría es la emoción que más fácilmente se reconoce en la música, mientras que la amenaza y la ira suelen confundirse.

En los últimos años, varios estudios han comparado los dos modelos categórico y emocional.

(Vieillard et al., 2008) utilizaron su base de datos de extractos musicales para analizar ambos modelos. Respecto al modelo categórico encontraron que los sujetos identificaban con mucha precisión las cuatro emociones (alegría, tristeza, miedo, tranquilidad). Respecto al modelo dimensional, aunque partieron de un modelo Valencia-Activación, consideran que sus resultados eran más consistentes con un mapa bidimensional con dos dimensiones salientes en términos de Energía y Tensión. La dimensión Energía, en la que contrastan tristeza y alegría, podría corresponder a una pulsión apetitiva, mientras que la dimensión de Tensión podría asimilarse al sistema defensivo.

(Eerola & Vuoskoski, 2011) utilizaron así mismo su base de datos de fragmentos de bandas sonoras de películas para evaluar el modelo categórico versus el modelo dimensional tridimensional (Valencia-Energía-Tensión), encontrando que ambos presentan una gran correspondencia. Las puntuaciones de los extractos mostraron ser muy consistentes en ambos modelos, aunque el modelo dimensional mostró ser más fiable en el caso de extractos musicales más ambiguos sin una única emoción claramente destacada. Sus resultados sugieren que el modelo tridimensional puede reducirse a uno bidimensional Valencia-Activación al menos en el ámbito de la emoción musical. Considera que sus resultados son compatibles con el modelo híbrido de Russell (Posner et al., 2005).

(Paquette et al., 2013) consideran que los resultados obtenidos en su estudio sobre los Musical Emotional Bursts encajan bien con la representación dimensional Valencia-Activación de las emociones: los estímulos alegres transmiten emociones positivas y excitantes, los estímulos de miedo transmiten emociones negativas y excitantes, los estímulos tristes transmiten emociones negativas moderadamente excitantes, y los estímulos neutros transmiten una valencia emocional neutra con poca excitación.

2.2.5 Cerebro y emoción

Las técnicas EEG y fMRI han permitido dilucidar algunas de las características de la actividad cerebral relativa a la emoción, pero aún queda mucho por esclarecer.

Desde los primeros modelos categórico y dimensional, se asume ampliamente que las emociones son las representaciones subjetivas de circuitos neuronales primarios, básicos para la supervivencia, que han evolucionado desde los primeros animales complejos (Lang & Bradley, 2010). Estos circuitos se ubican en el cerebro antiguo (el sistema límbico y otras regiones internas), y responderían ante estímulos críticos para la supervivencia moviendo a la acción (acercamiento ante estímulos positivos, o alejamiento ante estímulos negativos). Estos circuitos están también conectados con áreas más desarrolladas, como la corteza o el cerebelo, por lo que los estímulos externos producen respuestas más flexibles y adaptativas, moduladas por procesos cognitivos que actúan “frenando" estas reacciones primarias.

Los estudios basados en fMRI han permitido investigar estos núcleos de procesamiento internos del cerebro. En términos de activación, se ha encontrado que el área de mayor respuesta a estímulos emocionales en el cerebro es la amígdala, una estructura subcortical perteneciente al sistema límbico, situada en la parte interna del lóbulo temporal medial (Lang et al., 1997) (Koelsch et al., 2013).

En cuanto a la valencia, se ha correlacionado sobre todo con el circuito de premio-recompensa del cerebro, en particular con el Núcleo Accumbens (NAc), particularmente relevante en el procesamiento de la recompensa y el placer (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001) (Knutson, Adams, Fong, & Hommer, 2001). Otros estudios incluyen otras estructuras del sistema de recompensa como el hipotálamo y el área ventral tegmental (VTA) (Menon & Levitin, 2005), y la corteza cingulada que realiza un papel de conexión entre el sistema límbico y la corteza superior (Blood & Zatorre, 2001) (Bush, Luu, & Posner, 2000).

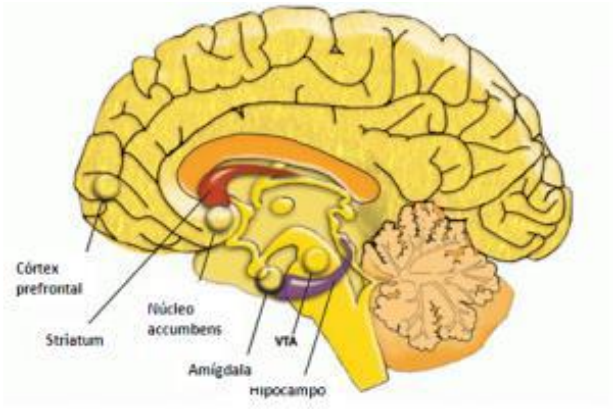


Ilustración 22. Estructuras del sistema de recompensa mesolímbico

La relación entre estos circuitos internos y algunas regiones frontales (como la corteza orbitofrontal y el giro frontal inferior) más encargadas del procesamiento cognitivo y la toma de decisiones, ha llevado a algunos investigadores a establecer una estrecha relación entre el procesamiento afectivo y el procesamiento cognitivo de distintos estímulos emocionales (Koelsch, Fritz, Cramon, Müller, & Friederici, 2006) (Cinzia & Vittorio, 2009) (Kawabata & Zeki, 2004).

Otros núcleos de la corteza cerebral importantes para el procesamiento emocional son las áreas parietal y temporal (Lang & Bradley, 2010) (Menon & Levitin, 2005). Cabe mencionar la ínsula, en la confluencia de los lóbulos temporal, parietal y frontal, que ha sido minuciosamente estudiada y definida como un relé entre el sistema límbico y el sistema motor (en la corteza), y puede ser el soporte fisiológico de estados autónomos subjetivos, como el dolor, el hambre, la percepción de la frecuencia cardíaca, o la conciencia emocional (Craig, 2002) (Craig, 2004).

Además de las interacciones del sistema límbico con las áreas cognitivas del cerebro, en el procesamiento de las emociones, muchas áreas diferentes y especializadas necesitan interactuar para dar cuenta de una experiencia subjetiva tan compleja. Por ejemplo, se han encontrado relaciones entre núcleos de procesamiento emocional y regiones visoespaciales y visomotoras, por ejemplo, durante la observación de esculturas clásicas y renacentistas, en lo

que parece una resonancia motora mental de los movimientos representados en las esculturas (Cinzia & Vittorio, 2009).

Por otra parte, las técnicas EEG han permitido analizar el patrón temporal de las reacciones emocionales. En un primer rango de 200-300 ms, aparece el ya mencionado P300 en las regiones occipital y temporal, con una amplitud relacionada con la intensidad del estímulo e independiente de su valencia (Hajcak, MacNamara, & Olvet, 2010) (Oatley & Johnson-laird, 1987). Otra de las firmas neuronales del procesamiento emocional es el llamado Potencial Positivo Tardío (LPP), considerado como la "significancia motivacional" de un estímulo (Bradley, M., 2009). Este potencial aparece alrededor de un segundo después de la presentación del estímulo, y su amplitud depende de la intensidad del estímulo (Hajcak et al., 2010) (Schupp et al., 2004). Los hallazgos de las señales EEG correlacionadas con el procesamiento emocional, muestran que los potenciales P300 y LPP rastrean los procesos emocionales, y que la información sobre la valencia se procesa antes que la información sobre la intensidad (Gianotti et al., 2008) (Olofsson, Nordin, Sequeira, & Polich, 2008).

2.2.5.1 Cerebro y música

Numerosos estudios EEG y fMRI se han centrado específicamente en medir las activaciones cerebrales que produce la música. (Frühholz, Trost, & Kotz, 2016) llevaron a cabo una revisión sobre el procesamiento de sonidos afectivos, y en particular de sonidos musicales, en la que compilaron las regiones del cerebro en las que de forma consistente se ha detectado activación en los estudios examinados, en función de los distintos tipos de estímulos sonoros. En la siguiente figura se muestran las regiones activadas por la música con los acrónimos indicados entre paréntesis.

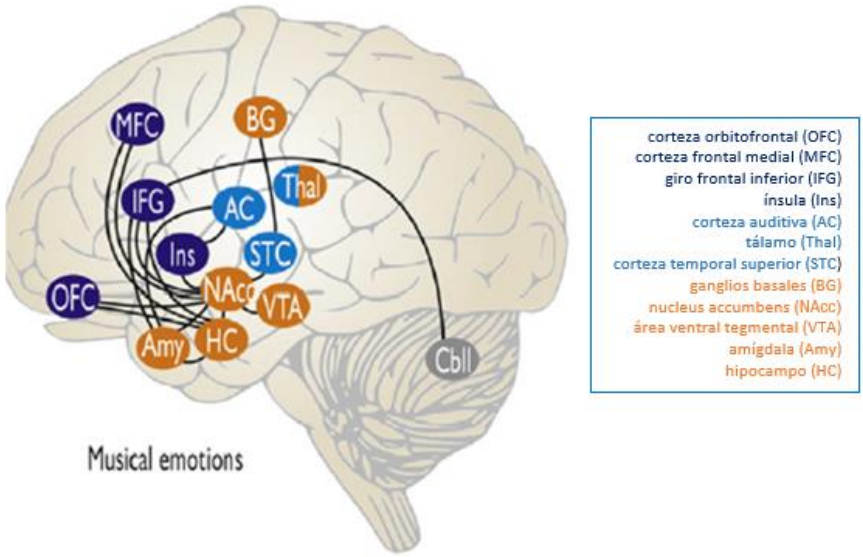


Ilustración 23. Regiones implicadas en el procesamiento de la música, de acuerdo con (Frühholz et al., 2016). Las líneas negras representan conexiones funcionales en la descodificación del contenido emocional de los sonidos.

(Blood & Zatorre, 2001) vieron que cuando la música producía emociones placenteras muy intensas, se activaban las regiones cerebrales de los circuitos de recompensa. (Frühholz et al., 2016) mostraron curiosamente en su revisión que el nucleus accumbens (NAcc) y el área ventral tegmental (VTA), pertenecientes al circuito cerebral de recompensa, se activan principalmente con los sonidos musicales, de entre todos los sonidos afectivos.

(Koelsch et al., 2006) encontraron mediante fMRI que una música agradable, consonante, activaba además el opérculo rolándico. Koelsh explica la activación del opérculo rolándico como un mecanismo de función espejo durante la percepción de lo melodías agradables, ayudando a la formación de representaciones (premotoras) para la producción de sonidos vocales durante la percepción de información auditiva agradable. También observó que las activaciones producidas en el cerebro aumentaban con el tiempo durante la presentación de los estímulos musicales, lo que indica que los efectos del procesamiento de las emociones tienen una dinámica temporal.

Tanto en la música como en el habla hay elementos emocionales comunes: por ejemplo, frases habladas o musicales felices tienden a ser más rápidas y en tonos más altos que las tristes. (Frühholz et al., 2016) proponen la existencia de una red neuronal común para procesar sonidos afectivos, y en particular sonidos musicales. Según este modelo, la corteza auditiva integraría información emocional descodificando las características complejas de la música que evolucionan con el tiempo, mientras que la amígdala descodificaría características musicales breves, como los cambios repentinos, que evocan emociones fuertes. La amígdala parece tener una función clave en el reconocimiento de la emoción (Koelsch et al., 2013), habiéndose comprobado que daños en la amígdala afectan al reconocimiento de la tonalidad emocional de una melodía (Vieillard et al., 2008). Incluso, aunque siempre se ha supuesto que el procesamiento de la estructura musical se daba a nivel cognitivo, el sistema límbico también desempeña un papel importante en el procesamiento de elementos como el tono (Gorzelańczyk, Podlipniak, Walecki, Karpiński, & Tarnowska, 2017).

La ínsula estaría involucrada en la traducción de la percepción auditiva de los sonidos musicales en una auto experiencia de la emoción (Kotz, Kalberlah, Bahlmann, Friederici, & Haynes, 2012) (Bamiou, Musiek, & Luxon, 2003). Las regiones frontales llevarían a cabo la evaluación y categorización de la música (Blood & Zatorre, 2001) (Panksepp & Bernatzky, 2002), y ponderarían la correspondiente respuesta adaptativa (Frühholz et al., 2016). Los ganglios basales, conectados a la corteza auditiva son sensibles a patrones temporales y podrían hacer predicciones temporales (Scherer & Zentner, 2001) (Salimpoor, Benovoy, Larcher, Dagher, & Zatorre, 2011), participando en la anticipación de eventos placenteros durante la escucha de la música, provocando así un aumento de la actividad en el sistema de recompensa nucleus accumbens - área ventral tegmental. El cerebelo participaría en respuestas motoras involuntarias a sonidos afectivos (Zald & Pardo, 2002).

Además (Frühholz et al., 2016) analizaron el índice de lateralidad en la activación cerebral de los distintos tipos de estímulos sonoros. Los resultados se resumen en la siguiente figura, mostrando que existiría una lateralización en el hemisferio derecho para el procesamiento de la prosodia afectiva.

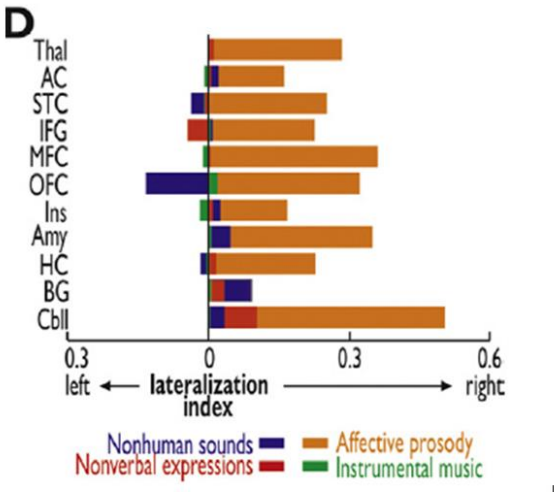


Ilustración 24. Índice de lateralidad para varias regiones cerebrales y distintos tipos de sonidos (Frühholz et al., 2016)

Parte de las regiones mencionadas, más superficiales, pueden ser medidas con EEG: corteza auditiva, corteza temporal superior, corteza orbitofrontal, corteza frontal medial, giro frontal inferior, ínsula y opérculo rolándico.

2.2.5.2 Oído relativo

Es interesante la forma en la que se procesa la música desde su entrada en el oído. El oído funciona como un analizador de sonidos. Cada armónico de un sonido musical excita un punto de la membrana basilar situada en el interior de la cóclea.

Ilustración 25. Rangos de frecuencia en la membrana basilar

El cerebro no percibe las frecuencias en una escala lineal. Se detectan mejor las diferencias en frecuencias más bajas que en frecuencias más altas. Por ejemplo, se percibe fácilmente la diferencia entre 500 y 1000 Hz, pero difícilmente se puede distinguir una diferencia entre 10,000 y 10,500 Hz, aunque la distancia entre los dos pares de frecuencias sea la misma. En 1937, (Stevens, Volkmann, & Newman, 1937)propusieron una unidad de tono tal que las distancias iguales en el tono sonaran igualmente distantes para el oyente, dando origen a la

escala Mel. El nombre Mel viene de la palabra melodía para indicar que se basa en la percepción humana de los tonos.

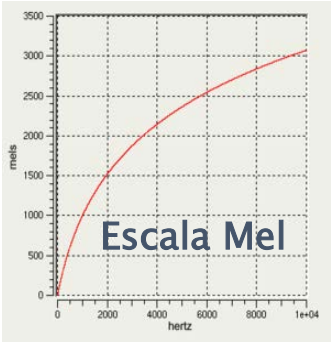


Ilustración 26. Escala Mel

Además, se perciben como una misma melodía intervalos musicales que tienen la misma relación de distancia. Por ejemplo, si entre dos notas hay la misma relación Nota1/Nota2 se percibe la misma melodía al escucharse las dos notas de forma secuencial, independientemente de la frecuencia. Lo que se percibe como esencial de una melodía son estas relaciones de frecuencia (Altozano, 2021).

Señalemos además como se detalla en la siguiente sección, que los resultados más agradables al oído se producen cuando la relación Nota1/Nota2 se expresa en relaciones matemáticas sencillas.

El origen de esta forma relativa de percepción puede residir en la importancia que para la supervivencia tiene distinguir las diferentes entonaciones (alegría, enfado, tristeza, ternura) de las voces masculinas (rango medio de 77 a 482 Hz) y femeninas (rango medio de 137 a 634 Hz). Las mujeres producen un rango de frecuencias más amplio, medido en hercios, para alcanzar el mismo resultado perceptivo en términos de tonos relativos que los hombres, porque su frecuencia natural es más alta que la de los hombres. Es decir, expresado en tonos relativos el rango tonal es similar entre ambos sexos, pero expresado en hercios no (Polo, 2019).

2.2.6 El origen de nuestra escala musical

Los primeros estudios sobre afinación conocidos fueron los realizados por Pitágoras (569-475 a.C.) y sus discípulos, en el siglo V a.C. (Senabre, 2018). El método que utilizaron los pitagóricos consistió en analizar el modo en que la altura de tono producido al pulsar una cuerda variaba en función de la longitud de esta, su grosor o la tensión aplicada. De este modo, los pitagóricos observaron que, en la construcción de intervalos, los resultados más agradables al oído se producían cuando dividían la longitud de las cuerdas utilizando relaciones matemáticas sencillas. Al dividir una cuerda por la mitad, el sonido obtenido era del todo consonante con el de la cuerda entera. Al dividirla nuevamente por la mitad, el resultado también lo era.

La regla general que obtuvieron fue que los sonidos más consonánticos se obtenían al combinar dos cuerdas con una relación entre sus longitudes que, expresada como una fracción, ofrecía un numerador y un denominador con números enteros y pequeños. Así las divisiones que mejores resultados proporcionaban siguiendo esta lógica, fueron:

- La mitad de la cuerda, una proporción que nos da el intervalo de octava (a esta proporción los pitagóricos la denominaron diapasón).
- Dos tercios de la cuerda, una proporción que nos da el intervalo de quinta justa, (denominada diapente).
- Tres cuartos de la cuerda, una proporción que nos da el intervalo de cuarta justa (denominada diatesarón).

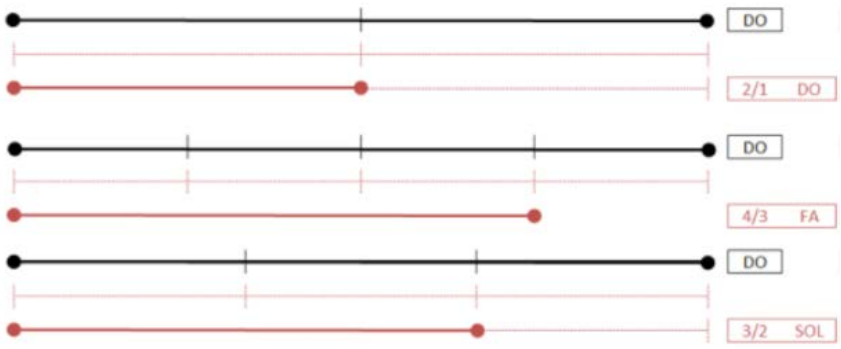


Ilustración 27. Intervalos perfectamente consonantes

A partir de estas observaciones establecieron un sistema de afinación basado en el intervalo de quinta justa (la relación de ese intervalo es de 3:2). El encadenamiento de quintas (Do, Sol, Re, La, Mi, etc.) genera un ciclo que abarca los doce tonos de nuestra escala cromática. Esto puede observarse en la siguiente figura:

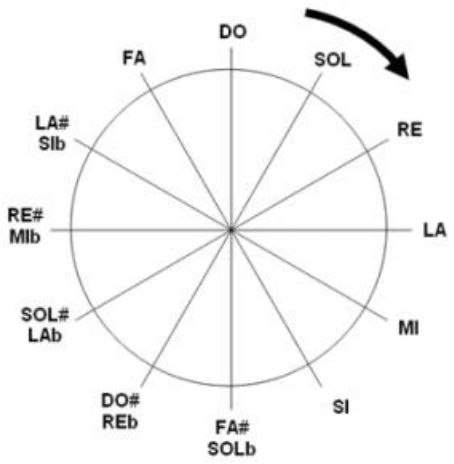


Ilustración 28. El círculo de quintas (Senabre, 2018)

Un instrumento construido con el método pitagórico sería capaz por tanto de emitir todos los sonidos usados para hacer música en cualquier tonalidad actual dentro de la música occidental.

El método, sin embargo, no es del todo satisfactorio ya que existe un problema en el fundamento matemático de la escala cromática así formada. El encadenamiento sucesivo quintas, con una relación 3:2, nunca da una frecuencia que se corresponda a la octava, de relación 2:1 (ningún número es al mismo tiempo potencia de 3 y de 2, salvo la unidad, que representa el unísono). La manera que utilizaron los pitagóricos para corregir la diferencia con la octava fue completar el ciclo utilizando una quinta ligeramente menor que el resto. La última quinta del ciclo se desafinaba para que éste pudiera cerrar en la octava.

La afinación pitagórica se deriva, por lo tanto, de la superposición de 11 quintas naturales (generalmente, se hacía desde Mib hasta Sol#), más una 12ª quinta que recibe toda la coma pitagórica. En este sistema, los intervalos de quinta son perfectamente consonantes, salvo en el caso de la que recibe la coma pitagórica (la 12ª, también conocida como «quinta del lobo»).

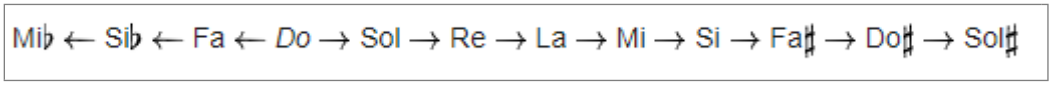


Ilustración 29. Secuencia de intervalos de quinta

La escala, en consecuencia, sonaba bien en el contexto de la música monódica, o para la superposición de quintas y cuartas justas (siempre que se evitara la denominada quinta del lobo) y por eso fue utilizada hasta la Edad Media (por ejemplo, en el canto gregoriano). El problema matemático persiste en la actualidad, aunque se resuelve en general con el sistema de afinación denominado "temperamento igual", construido mediante la división de la octava en doce partes iguales llamadas semitonos temperados, es decir se desafinan ligeramente los semitonos para que el intervalo de octava se mantenga siempre en la relación 2F. El temperamento igual hoy en día es considerado como la afinación estándar de la escala cromática occidental de 12 notas.

En la teoría musical actual, los intervalos de octava, quinta y cuarta, establecidos desde Pitágoras, se consideran perfectamente consonantes. Si se tocan por ejemplo un DO y un SOL a la vez, o dos DO separados por una octava, o un DO y un FA, suenan bien juntos, no producen tensión. Mientras que por ejemplo los intervalos FA-SI o SI-FA (cuarta aumentada o quinta disminuida), producen tensión, y se consideran disonantes.

2.2.7 Parámetros musicales y emoción

En los últimos años, el estudio de la relación entre parámetros musicales y emoción está cobrando así mismo mucho interés y se ha determinado con bastante precisión la relación entre distintos parámetros musicales y el tipo de emoción que generan. Una de las primeras recopilaciones es la de (Juslin, P. N. & Sloboda, 2001) "Music and emotion: theory and research" que incluye el capítulo de Gabrielsson, "The role of structure in the musical

expression of emotions". En general estos estudios se han centrado en las emociones básicas: alegría, tristeza, y miedo, (Gabrielsson, A., Lindström, E., 2010), que son las más fáciles de reconocer en la música, añadiendo una cuarta emoción que dependiendo de los autores suele ser paz o ternura. Respecto a los parámetros musicales más estudiados se describen en la siguiente tabla.

MODO	Tonalidad mayor o menor de la escala utilizada en la composición. El modo de una melodía viene determinado por la distancia entre los grados I y III de una escala. Por ejemplo, en la escala de DO, sería la distancia entre el primer grado DO y el tercer grado MI. En el modo mayor son 2 tonos, y en el modo menor son 1 tono y un semitono.
TEMPO	Velocidad de ejecución
DINÁMICA	Graduaciones en la intensidad del sonido
ARTICULACIÓN	Cómo se produce el ataque y finalización de cada nota, y la transición entre notas. Es LEGATO si el sonido y el volumen entre nota y nota son continuos. Es STACCATO cuando se produce una interrupción del sonido entre nota y nota.
ATAQUE	Tiempo que el sonido tarda en pasar de cero a su punto máximo de intensidad en una nota.
CAÍDA	Tiempo que el sonido tarda en pasar de su punto máximo de intensidad a cero en una nota.
TIMBRE	Cualidad del sonido, depende de la cantidad de armónicos que tenga un sonido y de la intensidad de cada uno de ellos.

Tabla 2. Parámetros musicales principales

El modo de una melodía viene determinado por la distancia entre los grados I y III de una escala, difiriendo el modo mayor y menor por medio semitono. Un acorde de 3 notas en modo mayor se siente alegre, y un acorde de 3 notas en modo menor se percibe triste. A nivel de frecuencias, la diferencia es muy pequeña, como se puede observar en la Ilustración 30, con los acordes de Do mayor y Do menor en la escala central del piano. Solo difieren en la nota MI/MIbemol en unos 18 Hz, y esta pequeña diferencia en hercios marca una diferencia emocional alegría/tristeza.

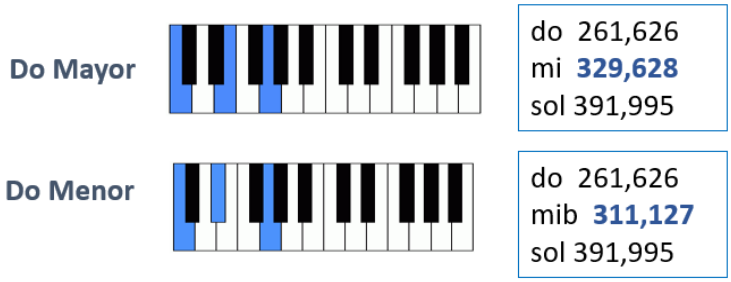


Ilustración 30. Frecuencias de las notas en los acordes de Do Mayor y Menor

(Gabrielsson, A., Lindström, E., 2010) recopilaron la asociación más frecuente en los estudios entre emociones básicas y parámetros musicales, y confirmaron en su estudio estas asociaciones. También comprobaron que la importancia relativa de estos parámetros variaba para cada emoción. Por ejemplo, el modo es muy importante para la tristeza, y la articulación para el miedo. (Eerola, Friberg, & Bresin, 2013) encontraron resultados similares, y consideraron que el parámetro más importante era el modo seguido de tempo, registro, dinámica, articulación y timbre, aunque comprobaron igualmente que la clasificación variaba con cada emoción. Por ejemplo, el modo es extremadamente importante para emociones felices y tristes mientras que tiene un impacto relativamente bajo en la emoción de miedo. También estableció que estos parámetros operaban de manera aditiva.

En la tabla siguiente se resumen los resultados de ambos estudios respecto a las relaciones entre emociones básicas y parámetros musicales.

	ALEGRÍA	TRISTEZA	MIEDO	PAZ/CALMA
MODO	mayor	menor	menor	mayor
TEMPO	rápido	lento	rápido	lento
REGISTRO	alto	bajo	bajo-alto	medio
DINÁMICA	alta	baja	alta	baja
ARTICULACIÓN	staccato	legato	staccato	legato
TIMBRE	brillante	suave	brillante	suave

Tabla 3. Parámetros musicales y emociones básicas

Por su parte (Bresin & Friberg, 2011) utilizaron la base de datos de (Vieillard et al., 2008), seleccionando las muestras musicales con más alta puntuación en las emociones de tristeza, alegría, miedo y neutro. Pidieron a músicos que ajustaran los parámetros de tempo, volumen, articulación, y registro de estas muestras según su criterio para transmitir mejor la emoción correspondiente. A partir de estos datos calcularon los valores medios de estos parámetros para cada una de las emociones: alegría, tristeza, miedo y paz. Los parámetros en los que los músicos estuvieron más de acuerdo fueron el tempo y el registro. La siguiente tabla resume el resultado de estos estudios. Los parámetros se muestran en orden de importancia de acuerdo con (Eerola et al., 2013), y con los valores medios evaluados por (Bresin & Friberg, 2011).

	ALEGRE	TRISTE	MIEDO	PAZ
TEMPO (notas/seg)	4,9	1,3	4,4	2,2
REGISTRO (Hz)	1660	932	659	124
ARTICULACIÓN	staccato	ligato	staccato	ligato
ATAQUE (mseg)	60,5	370,4	317,3	294,9

Tabla 4. Valores medios de parámetros musicales y emociones básicas

Los estudios de (Peretz et al., 1998) muestran que los juicios emocionales son consistentes entre sujetos, y determinados por la estructura musical (modo y tempo). Respecto al instrumento con el que se toca la música, tiene un impacto importante en el reconocimiento de emociones, ya que determina el timbre. Por ejemplo, (Hailstone et al., 2009) han comprobado que las melodías suenan menos alegres cuando se tocan con el violín que con otros instrumentos. (Paquette et al., 2013) observaron comparando violín y clarinete, que el miedo no se reconocía bien cuando se expresaba con el clarinete, y que los estímulos tristes y de miedo se identificaban con más precisión cuando se interpretaban con el violín, mientras que los extractos alegres se identificaban con más precisión cuando se interpretaban con clarinete.

Los intérpretes también utilizan los parámetros musicales para expresar la emoción. (Juslin, 2000) encontró dos dimensiones básicas, tempo y articulación, además del nivel de sonido, para explicar la transferencia emocional entre intérpretes y audiencia. Parece que incluso los niños (entre 4 y 12 años) son capaces de usar estas variables para expresar emociones en una canción (Adachi & Trehub, 1998).

Otro aspecto que se ha estudiado es la diferencia entre la emoción identificada por el oyente, es decir la emoción representada por la música, que el compositor o el intérprete quieren transmitir, y su propia respuesta emocional. La frontera entre ambas es muy difusa (Gabrielsson & Juslin, 2003) y los estudios muestran que suele haber mucho acuerdo en la identificación de la emoción independientemente del entrenamiento musical, inteligencia, cultura, o circunstancias personales de los oyentes (Fritz et al., 2009).

2.2.8 ¿Por qué la música emociona?

Cómo se ha visto la música activa los circuitos neuronales primarios del cerebro más interno, básicos para la supervivencia (amígdala, circuito premio-recompensa), por lo que la emoción musical tendría su origen en funciones de supervivencia (Koelsch, 2014), contrariamente al pensamiento tradicional que consideraba que el procesamiento de la música se realizaba fundamentalmente a nivel cognitivo. El origen de la emoción musical estaría en el mecanismo original de la emoción, un mecanismo primitivo adaptativo que se dispara ante estímulos críticos para la supervivencia, moviendo a la acción.

Así algunos autores consideran que la música podría activar circuitos emocionales biológicamente importantes para el procesamiento de vocalizaciones como risas, o gritos, prosodias, entonaciones (Peretz, 2010). Las emociones serían las encargadas de convertir los

sonidos que percibimos en algo comprensible, lo que hace que entendamos las situaciones en las que estamos (Koelsch, 2014).

De hecho, el reconocimiento de las emociones básicas en la música no sólo es consistente entre oyentes de la misma cultura (Vieillard & Gilet, 2013), también se ha comprobado que la emoción producida por la música sigue patrones universales, pudiendo distintas culturas distinguir los mismos sentimientos de alegría, tristeza o ira (Balkwill & Thompson, 1999) (Balkwill et al., 2004), incluso culturas sin exposición previa a la música occidental, pueden distinguir en ella emociones como alegría, tristeza y miedo (Fritz et al., 2009).

Además, este reconocimiento es inmediato, se produce en menos de dos segundos, con un simple acorde o unas pocas notas, por lo que la percepción de la música se puede incluir entre los procesos cognitivos y emocionales de acción rápida como los de la percepción de la cara y la voz (Filipic, Tillmann, & Bigand, 2010) (Belin et al., 2008). (Paquette et al., 2013), con su conjunto MEB (Musical Emotional Burts) de clips musicales de muy corta duración, en promedio 1.6 segundos, comprobaron cómo los sujetos experimentales categorizaban correctamente y con gran precisión las emociones asociadas a estos clips, y que incluso 250 milisegundos desde el inicio de la música eran suficientes en algunos casos para distinguir una música triste de una música alegre. (Vieillard et al., 2008) obtuvieron tiempos medios de reconocimiento de 483 milisegundos para fragmentos musicales representativos de alegría, 1445 milisegundos para fragmentos representativos de tristeza, 1737 milisegundos para fragmentos representativos de miedo, y 1261 segundos para fragmentos neutros.

El reconocimiento de la emoción en la música se produce además desde muy temprana edad. Desde los 2 a los 4 meses los bebés prefieren la música placentera, consonante a la música disonante (Trainor, Tsang, & Cheung, 2002). Desde los tres años, los niños pueden discriminar la tonalidad alegre o triste de breves pasajes musicales (Kastner & Crowder, 1990). Con cinco años discriminan entre músicas tristes o alegres en base a las diferencias de tempo, y a los seis años discriminan las tonalidades mayores versus menores (Dalla, Peretz, Rousseau, & Gosselin, 2001). Desde los seis años pueden distinguir tristeza, amenaza o enfado en la música (Dolgin & Adelson, 1990).

Ilustración 31. Bebé escuchando música [5]

Algunos autores ven el origen de las estructuras musicales en el desarrollo neurológico del feto, en el que los sonidos intrauterinos se almacenarían como patrones musicales en el

[5] https://www.planetacurioso.com/bebe-emocionado-con-la-musica/

cerebro, y consideran que es posible hacer coincidir los sonidos del útero con los elementos que se pueden encontrar en la música de todas las culturas (Teie, 2016). Así el ritmo se originaría a partir de la combinación de los sonidos de la respiración de la madre y los latidos de su corazón, mientras que las notas musicales y la melodía tendrían su origen en las modulaciones de la voz materna que el feto escucha dentro del útero. Este origen explicaría la asociación entre parámetros musicales y emoción. Los sonidos intrauterinos podrían explicar parámetros como tempo, registro, dinámica. Las prosodias, los cambios de entonación en las voces con las emociones de tristeza o alegría podrían explicar las diferencias entre los modos mayor y menor.

Esta característica de los modos mayor o menor en los que un semitono marca la diferencia entre una sensación triste o alegre ocurre también en músicas de otras culturas. En China por ejemplo secuencias de semitonos se utilizan para expresar quejidos, o el intervalo de tercera menor para transmitir llanto.

Desde el punto de vista evolutivo, (Montagu, 2017) considera que el ritmo, junto con la tendencia innata que se da también en animales a moverse al mismo ritmo, tendría una función adaptativa cohesiva, favoreciendo la vinculación entre grupos al moverse en grupo al mismo ritmo.

2.2.9 Resumen y Limitaciones

Existe consenso en que la emoción musical, al igual que la emoción en general, está entroncada con los circuitos neuronales del cerebro más internos, los más antiguos y básicos para la supervivencia: dos circuitos neuronales internos, que se activan en función de la intensidad (amígdala), o de la valencia del estímulo (circuito premio-recompensa). En su origen la emoción sería un mecanismo adaptativo que daría la alarma ante estímulos críticos para la supervivencia, positivos y negativos, moviendo a la acción, de acercamiento o de huida. Estos mismos mecanismos adaptativos ante estímulos como sonidos de la naturaleza, vocalizaciones, prosodias, percibidos incluso desde el útero materno, estarían en el origen de las estructuras musicales básicas. Este origen explicaría la inmediatez y universalidad de la emoción musical. La activación del circuito premio-recompensa (más destacada con los sonidos musicales que con el resto de los sonidos afectivos) explicaría el placer y "adicción" que produce la escucha de la música.

El cerebro procesa relaciones entre frecuencias y le *gustan* las relaciones matemáticas sencillas: los resultados consonantes más agradables al oído se producen cuando la relación por ejemplo entre dos notas Nota1/Nota2 es una relación matemática sencilla. Sin embargo, no existe un modelo teórico-científico consensuado que explique los mecanismos perceptivos subyacentes que dan lugar a las diferentes emociones musicales. En la mayoría de los estudios se siguen considerando conjuntamente los modelos categóricos, con unas pocas emociones básicas, y dimensional, sin acuerdo sobre sus dimensiones, aunque predomina el modelo Valencia-Activación.

Las emociones básicas e intensas de alegría, tristeza y miedo son las más claramente identificables en extractos musicales. Estas emociones se pueden identificar también como puntos del espacio dimensional continuo, mientras que es más confusa está identificación en

el caso de emociones intermedias. También hay gran consenso, tanto desde la musicología como desde la neurociencia, sobre los valores de los parámetros musicales que se relacionan con la alegría, la tristeza y el miedo, sobre todo en lo que se refiere a los parámetros modo, tempo, registro, dinámica, articulación y timbre. La emoción básica de ira, que se distinguen claramente desde el punto de vista emocional, es más difícil de identificar desde el punto de vista musical, ya que se produce una confusión en el oyente entre miedo percibido y miedo inducido. Igualmente, las bases de datos de fragmentos musicales científicamente validadas sobre las emociones musicales se basan en las emociones de alegría, tristeza, miedo, y una cuarta emoción, paz/tranquilidad/neutralidad, que no se considera emoción básica pero que se identifica fácilmente como un estado emocional musical percibido.

Por tanto, el consenso general desde los puntos de vista modelo teórico, relación con parámetros musicales, y bases de datos musicales, se da únicamente en las emociones: alegría, tristeza, miedo, cuando se expresan con intensidad.

2.3 Música y percepción vibro-táctil

Las investigaciones recientes muestran cada vez más analogías entre los mecanismos perceptivos del tacto y la audición, ambos sentidos respondiendo a variaciones de energía mecánica. Originalmente el estudio de la percepción vibro-táctil se inició desde la fisiología, con el propósito de investigar las respuestas de los mecanorreceptores de la piel, y para evaluación diagnóstica de disfunción nerviosa. Actualmente esta investigación está en auge en otros ámbitos, como el de las tecnologías hápticas que se basan en el tacto, en la percepción de texturas, formas o vibraciones, para facilitar las interacciones con distintos objetos o instrumentos. También en el ámbito de la investigación musical, en el que se está considerando con el objetivo de facilitar la experiencia musical a las personas con discapacidad auditiva, y en general, para mejorar la experiencia musical con un enfoque multimodal.

En esta sección se revisa la fisiología del sentido del tacto relacionada con la percepción de vibraciones, y se recapitulan estudios recientes sobre la percepción táctil de parámetros musicales como el tono y el ritmo.

2.3.1 Fisiología de la percepción vibro-táctil

El sentido del tacto responde a la energía mecánica. Los receptores sensoriales del tacto se activan por estímulos mecánicos y activan a su vez a la neurona sensorial aferente primaria. La neurona aferente primaria es la primera de un conjunto de neuronas dispuesta en serie que actúan secuencialmente a lo largo de una vía sensorial ascendente que recorre la médula espinal, el tallo cerebral, el tálamo, hasta la corteza cerebral.

Dentro de la piel glabra existen cuatro tipos de receptores táctiles especializado: Meissner, Pacini, Ruffini, y Merkel. La capacidad de sentir objetos a través del tacto es el resultado combinado de la activación de estos receptores (Jones & Lederman, 2007).

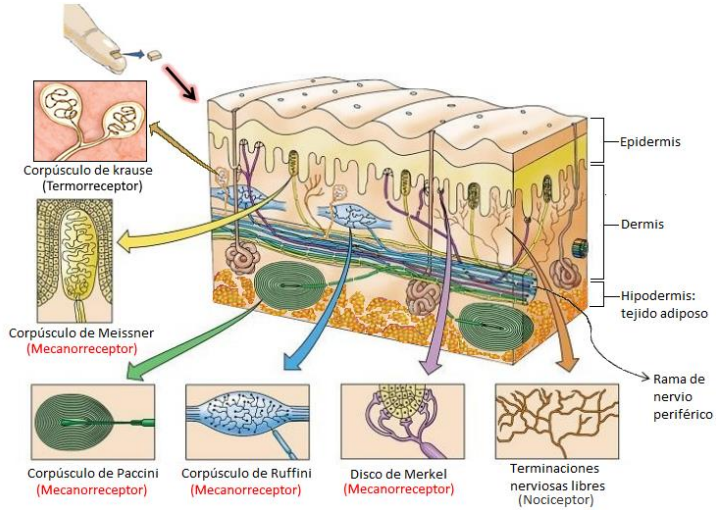


Ilustración 32. Receptores táctiles en la piel[6]

[6] https://eltamiz.com/elcedazo/2017/03/18/los-sistemas-receptores-04-el-tacto/

Los receptores de Merkel y Ruffini son de adaptación lenta, se mantienen activados mientras dura el estímulo, e informan sobre su intensidad. Por el contrario, los receptores de Meissner y Pacini, de adaptación rápida, sólo se activan cuando cambia la intensidad del estímulo, informando sobre sus variaciones. Son los que captan las sensaciones vibratorias. Las fibras nerviosas procedentes de estos receptores son gruesas y rápidas, con velocidades de 30 a 110 m/s, y transmiten con fidelidad temporal y espacial las sensaciones vibratorias.

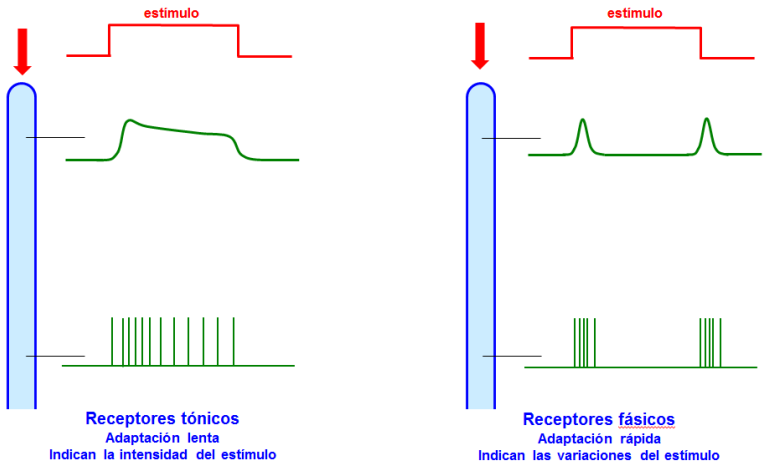


Ilustración 33. Adaptación lenta y rápida

Los corpúsculos de Pacini se encuentran en la parte más profunda de la capa dérmica. De adaptación muy rápida, son extremadamente sensibles a la vibración táctil, activándose con frecuencias entre 50 y 1000 Hz, con una sensibilidad máxima alrededor de los 200-300Hz. Generan sensación de vibración como la que se produce al deslizar una herramienta sobre una superficie texturizada (Jones & Lederman, 2007).Los corpúsculos de Meissner son superficiales, de adaptación rápida, y su sensibilidad máxima a las vibraciones táctiles está en el rango de frecuencias entre 5 y 50Hz. Generan sensación de aleteo, como la que se produce al deslizar un dedo sobre una superficie rugosa. La sensibilidad a las vibraciones depende de estos dos tipos de receptores, Meissner y Pacini, que conjuntamente cubren el de 5-1000Hz.

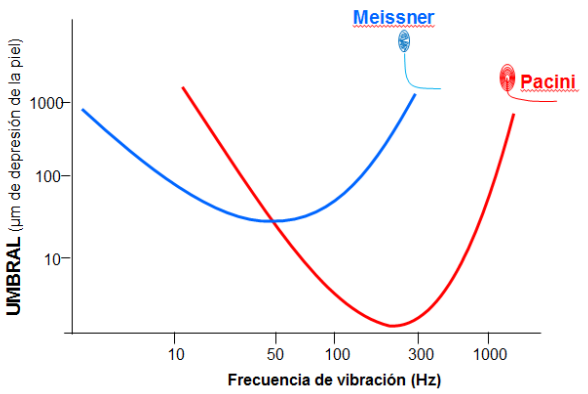


Ilustración 34. Umbrales en función de la frecuencia

Los receptores de Meissner tienen campos receptores superficiales pequeños (3-5mm), con una mayor densidad en la yema de los dedos.

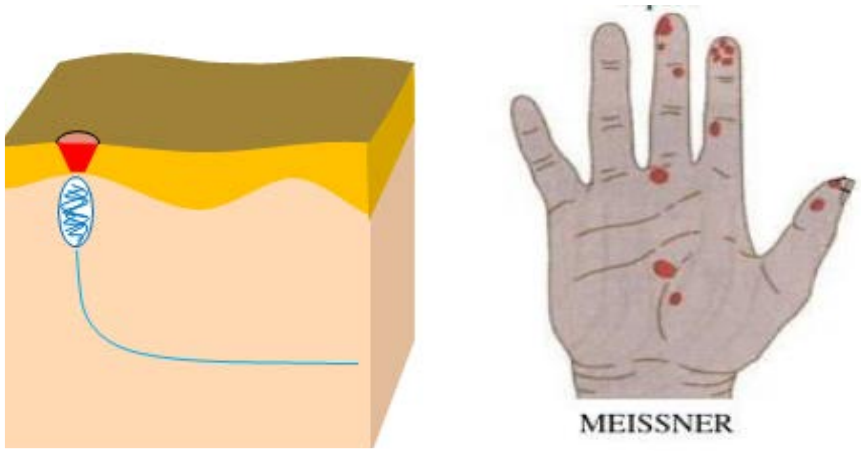


Ilustración 35. Localización de los corpúsculos de Meissner

Los receptores de Pacini tienen campos receptores profundos y más amplios.

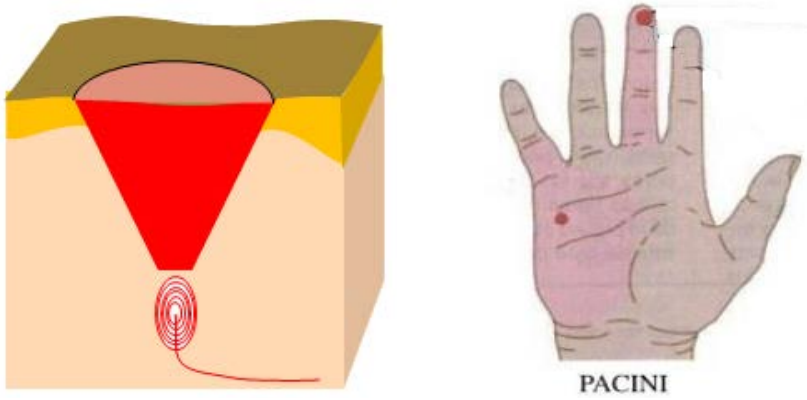


Ilustración 36. Localización de los corpúsculos de Pacini

Estos dos tipos de receptores tienen un umbral de activación bajo, pudiendo activarse con deformaciones mecánicas de la piel del orden de 1 Amstrong, aunque depende de la duración del estímulo y de la superficie estimulada. (Papetti, Stefano, Järveläinen, Giordano, Schiesser, & Fröhlich, 2017) estudiaron cómo los umbrales bajaban a mayor superficie de contacto y mayor duración del estímulo. Así, trabajando con ondas sinusoidales, los corpúsculos de

Pacini detectan mejor estímulos de 1 segundo de duración, y estímulos aplicados sobre superficies del orden de 2 cm de diámetro. Para una vibración sinusoidal de 200-300 Hz el umbral necesario es una deformación del orden de 1 μm.

Sin embargo, los mecanismos neurológicos que codifican la información sobre la amplitud y la frecuencia de una vibración táctil no están claramente identificados. La amplitud de la vibración podría estar relacionada con el número de picos de intensidad por intervalo de tiempo. Respecto a la frecuencia, la codificación podría estar relacionada con el patrón temporal de activación del impulso nervioso (Morley & Rowe, 1990).

2.3.2 Interacción oído y tacto

Cada vez hay más investigaciones sobre las correspondencias entre el sentido del tacto y el sentido del oído. Ambos sentidos cuentan con receptores sensibles a los estímulos de presión y tienen la capacidad de procesar vibraciones, aunque con diferentes umbrales y rangos de percepción, siendo el rango táctil de percepción de frecuencias de 5-1000 Hz, mientras que el auditivo es de 20- 20.000 Hz.

Según el modelo tradicional de procesamiento sensorial, la información del estímulo llega inicialmente a la corteza sensorial primaria. La siguiente figura muestra las áreas sensoriales primarias de los distintos sentidos.

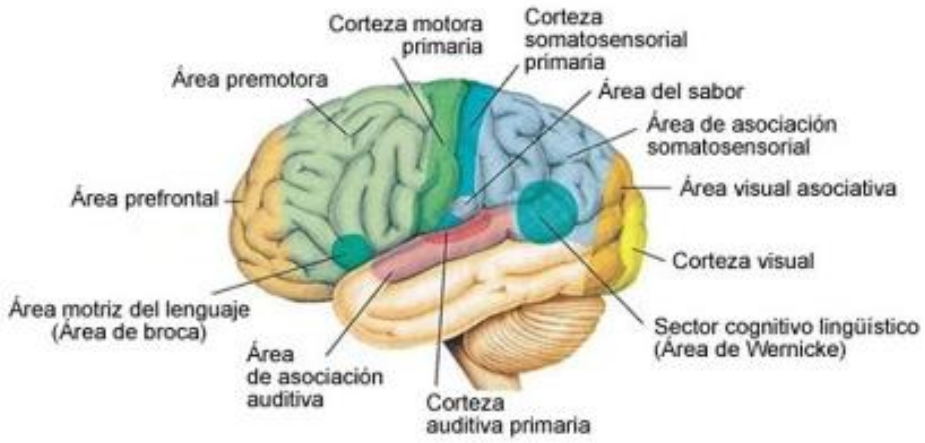


Ilustración 37. Áreas cerebrales funcionales

Desde la corteza sensorial primaria, la información iría pasando de una región a otra de la corteza cerebral en un proceso ascendente de análisis, regulado por el tálamo (Sherman & Guillery, 2001). La siguiente figura ejemplifica este modelo en el caso de una estimulación visual. La integración multisensorial ocurriría a nivel de la corteza asociativa multimodal.

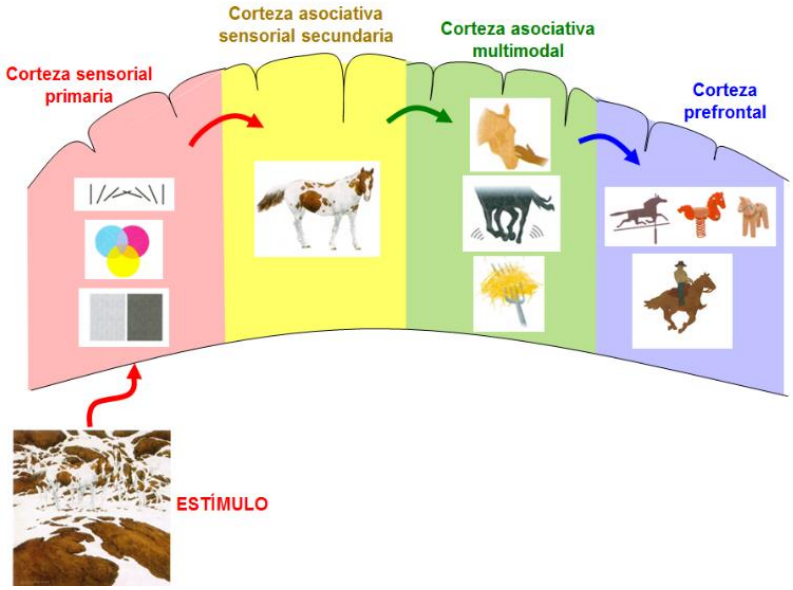


Ilustración 38. Modelo tradicional de procesamiento sensorial[7]

Sin embargo (Kayser, Petkov, Augath, & Logothetis, 2005) demostraron con técnicas de alta resolución fMRI en monos macacos que la integración multisensorial táctil-auditiva ocurre antes, muy cerca de las áreas sensoriales primarias, antes de la intervención de mecanismos pre-atencionales, y que era más intensa cuando los estímulos coincidían temporalmente. También se han registrado indicios de entrada vibro táctil en el córtex auditivo humano con vibraciones de baja frecuencia detectadas por los sistemas táctil y auditivo (Caetano & Jousmäki, 2006). La interacción entre oído y tacto también se ha mostrado en otros estudios sobre la percepción de frecuencia, que muestran como las señales de audio modifican la percepción de las frecuencias táctiles y viceversa (Convento, Wegner-Clemens, & Yau, 2019) (Crommett, Pérez-Bellido, & Yau, 2017) (Huang, Gamble, Sarnlertsophon, Wang, & Hsiao, 2012).

Aunque los mecanismos neurológicos que codifican la información sobre la vibración táctil aún no están claramente identificados, Kayser et al. sugieren que existe un área de coactivación vibro táctil-auditiva (Kayser et al., 2005) y que la integración auditivo-táctil se produce de forma temprana y cerca de las áreas sensoriales primarias. Según (Huang et al., 2012) la corteza auditiva estaría involucrada tanto en el procesamiento temporal táctil como en el ritmo auditivo.

(Huang et al., 2012) relacionan la percepción del ritmo vibro táctil con la activación de los receptores de Pacini, que codificarían los patrones temporales de las vibraciones de forma similar a la codificación de ondas sonoras en la cóclea del sistema auditivo, y consideran que el ritmo se procesa posteriormente por un mismo mecanismo perceptivo común a ambos

[7] https://slideplayer.es/slide/14538901/

sentidos, y no por vías sensoriales distintas. (Rahman, Barnes, Crommett, Tommerdahl, & Yau, 2020) demostraron, mediante fMRI, que la información de frecuencia temporal contenida en los estímulos audio-táctiles era procesada por sistemas somatosensoriales y auditivos específicos para cada característica, y comunes para ambos sentidos.

Respecto al timbre, (Russo, Ammirante, & Fels, 2012) proponen que, análogamente a la discriminación auditiva del timbre, la discriminación vibro táctil de distintos timbres podría ser el resultado de la integración a nivel cortical de salidas de receptores táctiles sintonizados que funcionarían como filtros paso banda en las distintas frecuencias. Así un tono complejo se descompondría en sus frecuencias componentes, contribuyendo a una percepción táctil del timbre en base a las intensidades de cada banda. (Young, Murphy, & Weeter, 2015), a partir de su estudio sobre la respuesta táctil a vibraciones puras y vibraciones complejas, apoyan también la existencia de una operación combinada de filtros paso banda a nivel de los mecanorreceptores de la mano, que darían respuestas individuales al desplazamiento mecánico de la piel, funcionando como filtros sintonizados para los tonos complejos.

2.3.3 Percepción vibro-táctil de parámetros musicales

La sensación de tono musical es producida por vibraciones mecánicas periódicas. En base a las correspondencias entre el sentido del oído y del tacto, siendo ambos sensibles a estímulos mecánicos de presión y capaces de procesar vibraciones, se han incrementado en los últimos años los estudios sobre la percepción táctil de parámetros musicales como son el tono, el ritmo, o el timbre.

En general, estos estudios se basan en la presentación de vibraciones en las yemas de los dedos de los sujetos, palmas de las manos o plantas de los pies, aunque también hay estudios que se basan en sillas o respaldos vibratorios. Se han centrado sobre todo en la discriminación de tonos puros, generalmente sinusoidales, y en la discriminación de distintas formas de ondas. Como ejemplos de dispositivos las siguientes figuras muestran un actuador de dedo y un banco. En ambos casos, el disco en contacto con el dedo, o el banco, están conectados a vibradores electrodinámicos. En general se utilizan auriculares aislantes para evitar que los ruidos de las vibraciones lleguen a los participantes y contaminen los experimentos.

Ilustración 39. Ejemplos de dispositivos experimentales (Hopkins et al., 2013) (Merchel & Altinsoy, 2013)

Respecto a la discriminación del tono, los estudios psicofísicos se han centrado sobre todo en la discriminación vibro-táctil de tonos puros. El rango de percepción táctil de tonos puros es variable en función de las condiciones de presentación del estímulo (Rovan & Hayward, 2000). En estos estudios se determinó que un rango de 70 a 800 Hz era perceptible para los humanos, y que el umbral de la "audición" táctil está vinculado a la duración del evento. La intensidad del estímulo, es decir la presión sobre piel, puede aumentarse hasta 55dB por encima del umbral de detección, punto a partir del cual la vibración se torna molesta. (Merchel & Altinsoy, 2013) estudiaron la capacidad de discriminación entre varias frecuencias, mediante vibraciones aplicadas en la espalda de los sujetos. Los resultados obtenidos muestran que el intervalo mínimo detectable aumenta con la frecuencia, desde aproximadamente 7 Hz alrededor de los 20 Hz, hasta 66 Hz alrededor de los 90 Hz. (Wyse, L. Nanayakkara, S. Seekings, P. Ong, S. H. Taylor, E. A., 2012) aplicaron las vibraciones en toda la palma de la mano encontrando que los sujetos con discapacidad auditiva eran sensibles a las vibraciones a frecuencias de 2000 Hz y 4000 Hz, aunque las amplitudes requeridas para la detección eran 30-40 dB más altas que a una frecuencia de 250 Hz.

(Kuroki, Watanabe, & Nishida, 2017) mostraron presentando distintas frecuencias simultáneamente en distintos dedos de una misma mano, que la percepción de la frecuencia surge no sólo de un canal mecanorreceptor específico en una localización cutánea concreta, sino también de la integración de los estímulos sincrónicos a través de diferentes canales y diferentes ubicaciones de la piel.

Son particularmente interesantes los estudios de (Hopkins, Mate-Cid, Seiffert, Fulford, & Ginsborg, 2013) sobre la discriminación de los tonos de la escala musical diatónica, aplicando vibraciones táctiles con la frecuencia pura de cada tono en la yema de los dedos y en la planta del pie de los sujetos durante 1 segundo. Considerando que la deformación mecánica de la piel necesaria para detectar la vibración aumenta con la frecuencia, y disminuye con tiempo de presentación del tono, y para evitar efectos negativos relacionados con la exposición a la vibración, se estableció que, el rango de notas musicales que pueden considerarse para la presentación vibro-táctil de la música es de C1 a G5. En este rango las notas pueden presentarse con una amplitud pequeña, de unos 10dB por encima del umbral de detección de la vibración, pero con duraciones no inferiores a 0.125s para ser detectadas con comodidad. Esto representa un rango musical de trabajo muy amplio, que cubre las frecuencias fundamentales de la voz humana, y seis de las ocho octavas de un piano, como se muestra en la figura siguiente. En cuanto a la diferenciación entre dos notas, se encontraron muchas limitaciones para intervalos menores de tres tonos (tercera menor). Además, observaron que estos resultados eran similares en personas con y sin discapacidad auditiva.

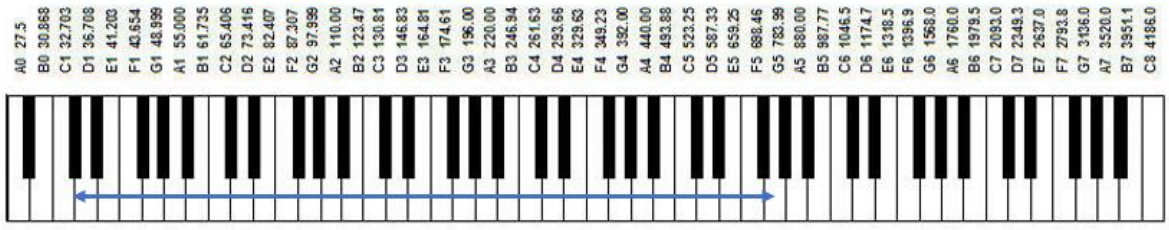


Ilustración 40. Teclado del piano

(Kuroki et al., 2017) presentando dos vibraciones distintas a diferentes dedos de las manos, mostraron como los sujetos percibían una frecuencia integrada correspondiente al promedio ponderado por la intensidad de las dos vibraciones.

Respecto al ritmo, los estudios muestran que la percepción auditiva del ritmo es mejor que la percepción táctil, y que la percepción táctil es mejor que la percepción visual del ritmo. A nivel experiencia, los sujetos sin discapacidad auditiva, consideran la vibración táctil muy satisfactoria para seguir el ritmo, con una calidad de experiencia muy similar a la auditiva (Kosonen & Raisamo, 2006). (Tranchant et al., 2017) compararon a personas sordas y oyentes bailando al ritmo de estímulos vibro-táctiles (sin sonido) sobre una plataforma vibratoria, encontrando resultados similares en ambos grupos: la sincronización táctil-motora precisa en un contexto de baile ocurre sin la experiencia auditiva, aunque la sincronización auditiva-motora es de calidad superior.

Respecto al timbre, (Russo et al., 2012) investigaron la capacidad de discriminar entre timbres musicales presentados de forma vibro-táctil a través del respaldo de una silla. Encontraron que la tasa de éxito en la discriminación entre los tonos de violonchelo, piano y trombón para una misa frecuencia estaba por encima del azar. (Young et al., 2015) trabajaron con distintas formas de ondas puras y complejas a una frecuencia fundamental fija que se presentaban aisladamente a la mano. Los resultados indicaron que los sujetos podían distinguir entre diferentes formas de onda a través de la estimulación táctil.

2.3.4 Tecnología vibro-táctil y música

En base a este potencial de percepción vibro-táctil de los parámetros musicales, se está desarrollando en los últimos años un nuevo campo de investigación y creatividad basado en la experiencia multimodal de la música. El objetivo ya no es sólo escuchar la música, sino "sentir" la música, tanto para facilitar el acceso a la música a las personas con discapacidad auditiva, como para ofrecer al público en general una experiencia más rica de la música.

Para las personas con sordera profunda es muy gratificante el poder sentir físicamente el sonido (a través de las manos, torso o piernas). Teniendo en cuenta que el rango de percepción de las frecuencias táctiles es de 5-1000Hz frente a los 20-2000Hz del oído, es más fácil sentir el sonido en conciertos con gran componente de bajos, como es el caso en la música heavy, o con muchas claves visuales para poder seguir el ritmo (Jack et al., 2015). De hecho, las personas sordas participan en actividades sociales de baile, y son capaces de bailar sincronizadamente con la música a través de las vibraciones que sienten en su cuerpo (Tranchant et al., 2017). Para el público en general, en la música moderna, con un alto nivel de decibelios, se busca no sólo que la música se pueda oír, sino que también se pueda sentir. (Teie, 2016) considera que el deseo de un sentido táctil de la música se remonta a las sensaciones del útero, donde el líquido que rodea al feto permite que las ondas de presión que emanan del latido del corazón materno se sientan en el cuerpo del feto.

Dentro de los diseños que se están realizando para potenciar la experiencia musical en las personas con discapacidad auditiva se encuentran distintos tipos de dispositivos que tienen como objetivo mapear la señal musical acústica a una señal vibro-táctil. (Jack et al., 2015)

diseñaron unos sillones en cuyos brazos o respaldos se introducen unas bobinas que transmiten vibraciones al cuerpo de la persona que está sentada y que reproducen parte de los parámetros de la música que se está escuchando (frecuencia, ritmo, dinámica, timbre). Las respuestas de los usuarios eran muy positivas cuando la música traducida a vibración táctil era muy rítmica, mientras que músicas más armónicas, con ritmos poco marcados producían sensaciones de zumbido o vibración difusa. (Hopkins et al., 2016) han diseñado unos dispositivos vibradores que se colocan en los pies de los músicos, dejando libres las manos para que puedan tocar su instrumento, mientras en los pies sienten reproducidas las vibraciones de los otros instrumentos del grupo, y pueden seguir el ritmo del conjunto. De esta forma se facilita que las personas con discapacidad auditiva puedan tocar en grupos musicales, y sincronizarse correctamente con el resto de los instrumentos del grupo.

Ilustración 41. Dispositivo vibrador para músicos (Hopkins et al., 2016)

El reto en este tipo de dispositivos es estudiar como los parámetros musicales (tono, ritmo, timbre…) pueden mapearse eficazmente a un patrón vibro-táctil. De momento se ha conseguido transmitir la percepción rítmica, mientras que los esfuerzos por transmitir el resto de los parámetros no consiguen buenos resultados y en general generan unas sensaciones de vibración indeterminada. Por ejemplo, los armónicos de un tono pueden generar una sensación de hormigueo al ser traducidos a vibración táctil (Papetti, S. & Saitis, 2018). Se están pues estudiando distintas aproximaciones de tratamiento de la señal acústica para su transformación en señal táctil, en general considerando que podría no ser necesario codificar toda la información auditiva disponible: por ejemplo, utilizando filtros paso bajo, o bien eliminando armónicos de la frecuencia fundamental, o bajando una octava la música original, o comprimiendo el rango de frecuencias, aunque sin resultados claros todavía que eliminen las sensaciones de hormigueo o vibración difusa. Otra aproximación del problema es trabajar la articulación de las notas para eliminar las vibraciones continuas y así evitar las sensaciones de hormigueo constantes, por ejemplo, iniciando la nota con un ataque prolongado, y luego finalizando la nota de forma más breve dejando un breve "silencio" antes de la siguiente nota. Otro enfoque para transmitir el tono, ya que la discriminación del tono táctil no es tan fina como en el oído, es codificar la información de tono en una dimensión táctil espacial, por ejemplo, aplicando bandas de frecuencias distintas en zonas distintas de la piel con múltiples actuadores (Baijal et al., 2012) (Branje & Fels, 2014) (Papetti & Saitis, 2018).

Desde el ámbito de la experiencia multimodal de la música para el público en general, un enfoque pionero es componer directamente música vibro-táctil, sin intentar traducir una música preexistente. (Baijal et al., 2012) desarrollaron un experimento de creatividad musical

considerando únicamente el sentido del tacto. A partir de una silla con distintos actuadores que vibraban en distintas bandas de frecuencia, propusieron a los participantes, seleccionados entre músicos oyentes y cineastas con discapacidad auditiva, que compusieran música táctil mediante un interfaz que permitía lanzar vibraciones con distintas características a la silla. El objetivo era componer pequeñas secuencias vibro-táctiles para subtitular un clip de 2 minutos con extractos de una película de Harry Potter. Los participantes consideraron esta tecnología muy prometedora, y valoraron mucho la experiencia de poder sentir las vibraciones en varias partes del cuerpo, sugiriendo incluso aumentar el número de zonas. (Branje & Fels, 2014) también han desarrollado un sistema de teclados ("Vibrochord") que permite enviar patrones de vibración a través de una silla.

Ilustración 42. Vibrochord (Branje & Fels, 2014)

(Gunther & O'Modhrain, 2003) consideran la composición táctil como un fin en sí mismo, convirtiendo la sensación táctil en el fin estético perseguido. Trabajan con un sistema que facilita la composición y la percepción de patrones espaciotemporales estructurados de vibración en la superficie del cuerpo. El objetivo final que se persigue es poder desarrollar un lenguaje de composición para el sentido del tacto. Son muy interesantes sus consideraciones sobre cómo debería ser este lenguaje. Por ejemplo, las secuencias de tonos, que tienen un papel de primer orden en la música, traducidas a vibraciones táctiles continuas producen sensación de zumbido en el caso de tonos bajos, y sensación de vibración difusa en frecuencias más altas, en vez de sensación musical. Para evitar esto, en el lenguaje táctil, el tono puede sustituirse por una dimensión tempo-espacial, en la que se distribuyen las bandas de frecuencia por distintas zonas de la piel pudiendo activarse de acuerdo con secuencias temporales. Estos autores proponen en el contexto de la composición táctil considerar la superficie completa del cuerpo. Por ejemplo, las frases musicales que se repiten en una composición musical tendrían su equivalente en secuencias de vibraciones a lo largo del cuerpo, que se repetirían con la misma secuencia espaciotemporal. Respecto a la intensidad, se puede trabajar con amplitudes de deformación de la piel que van desde los límites de detección de la vibración hasta el punto en que la vibración empiece a resultar molesta. Para la articulación, estímulos vibro-táctiles de menos de 0.1 segundos darían sensación de staccato, mientras que estímulos de más duración, con inicios y finales suaves de notas, darían sensación de frases musicales ligadas. También hay que tener en cuenta no estimular de forma

continua la misma parte del cuerpo, para evitar la adaptación que ocurre con relativa rapidez en la piel.

El tacto se está pues convirtiendo en un nuevo ámbito de creatividad artística, con un gran potencial para favorecer la accesibilidad de las personas con discapacidad auditiva a la experiencia musical, y como expresión artística para el público en general.

2.3.5 Resumen y Limitaciones

Los estudios muestran una estrecha relación en el procesamiento de vibraciones mecánicas entre los sentidos del oído y del tacto, y que parámetros vibro-táctiles como el ritmo o la frecuencia se procesarían por un mismo mecanismo perceptivo común a ambos sentidos. Por lo que podríamos considerar la estimulación vibro-táctil como una forma alternativa de transmitir parámetros musicales y por tanto emoción.

Como se ha visto en la sección 2.2.7, los parámetros musicales más relevantes respecto a la emoción son el modo y tempo, seguidos de dinámica, articulación y timbre. El reto es conseguir reproducir estos parámetros con la estimulación vibro táctil.

Diferentes tipos de dispositivos vibradores, como sillones, chalecos, pulseras, pedales vibradores, etc., ya han sido diseñados para mapear la señal musical acústica a una señal vibro táctil y se ha llegado a algunas conclusiones. En general hay que filtrar la señal auditiva ya que el rango de percepción de la frecuencia táctil es de 5-1000 Hz frente a 20-20000 Hz para la percepción auditiva, por ejemplo, mediante el uso de filtros pasa bajo, o eliminando armónicos de la frecuencia fundamenta. Cuando se utiliza la envolvente de una música para reproducirla mediante actuadores vibro táctiles, se ha visto que la música rítmica produce un feed-back positivo por parte de los usuarios, pero que músicas más armónicas producen sensaciones de zumbido o vibración difusa.

La reproducción tonos requiere otra aproximación, como puede ser codificar la información tonal mediante una dimensión táctil espacial, aplicando diferentes bandas de frecuencia en diferentes áreas de la piel. Sería interesante también comprobar si se puede conseguir una percepción de intervalos consonantes, o de modos mayor o menor investigando con la aplicación de distintas frecuencias vibro táctiles. El mapeo de la articulación también podría realizarse con una aproximación similar, produciendo vibraciones de forma escalonada en una dimensión táctil, con mayor o menor velocidad, emulando el ataque o decaimiento de una nota. La dinámica de una melodía podría conseguirse mediante cambios en la amplitud de la vibración o sumando vibraciones en distintas zonas espaciales.

Pero estos enfoques están en un estadio muy inicial y no existen estudios concluyentes sobre el adecuado mapeo de los parámetros musicales a la vibración táctil.

Por último, es interesante resaltar que la reacción a la estimulación vibro táctil es similar en sujetos con y sin discapacidad auditiva, lo que permite simplificar el reclutamiento de participantes en parte de las investigaciones determinando un canal alternativo a los usuarios sordos y de potenciación de la emoción a los usuarios oyentes.

2.4 Música y emoción: modelos computacionales

La capacidad de la música para inducir emociones ha dado lugar también en un ámbito totalmente diferente, el ámbito de la ciencia informática y la computación afectiva, a un campo de investigación dedicado a identificar las características de la música que generan los distintos estados emocionales. Este campo, denominado Music Emotion Recognition (MER), está despertando mucho interés en los últimos años, sobre todo por el gran auge de las plataformas de reproducción de música vía streaming y el uso de recomendadores musicales automáticos, que sugieren músicas en base a preferencias o incluso a los estados de ánimo del oyente (Yang et al., 2018). MER se basa en el análisis de características de nivel bajo o medio de la música. Estas características se obtienen de las muestras digitales de audio utilizando las técnicas de otro campo cercano de investigación en auge, el denominado Music Information Retrieval (MIR).

De acuerdo con la revisión de (Yang et al., 2018), el primer artículo en este campo fue publicado por (Feng, Zhuang, & Pan, 2003). En este trabajo los autores proponían un sistema de clasificación de canciones en 4 categorías emocionales: alegría, tristeza, ira y miedo, basado en dos características musicales, tempo (rápido o lento) y articulación (staccato o legato). Desde entonces se están produciendo muchos estudios sobre algoritmos de clasificación de emociones, sobre todo en el contexto de los Music Information Retrieval Evaluation eXchange[8] (MIREX), que se celebran como parte de las conferencias anuales que organiza la International Society for Music Information Retrieval (ISMIR), una organización sin ánimo de lucro. También son importantes los talleres organizados por la comunidad MediaEval[9] (Benchmarking Initiative for Multimedia Evaluation), dedicada a la evaluación de nuevos algoritmos de acceso y recuperación de contenidos multimedia.

Estas conferencias son el principal foro mundial de investigación sobre el tratamiento, la búsqueda, la organización y el acceso a los datos relacionados con la música. Pero MER sigue siendo un campo muy complejo, todavía en sus inicios, y con mucha diversidad de métodos de investigación.

2.4.1 Modelo MER (Music Emotion Recognition)

El esquema típico de desarrollo de un modelo MER consta de las siguientes fases (Yang et al., 2018):

- Selección y etiquetado de las muestras musicales digitales ("ground truth"). El etiquetado de los datos se basa en anotaciones realizadas a las muestras musicales por un grupo de sujetos que evalúan la emoción percibida en cada muestra.
- Selección y obtención de características ("features") de las muestras digitales a través de técnicas MIR (Music Information Retrieval).
- Aplicación de métodos de aprendizaje automático ("machine learning") para establecer un mapeo entre características y etiquetas emocionales.

[8] https://www.music-ir.org/mirex/wiki/MIREX_HOME

[9] https://multimediaeval.github.io/about/

Una vez definido el modelo se puede utilizar para predecir las emociones de nuevos fragmentos.

Cada una de estas fases presenta varios problemas que comentamos a continuación.

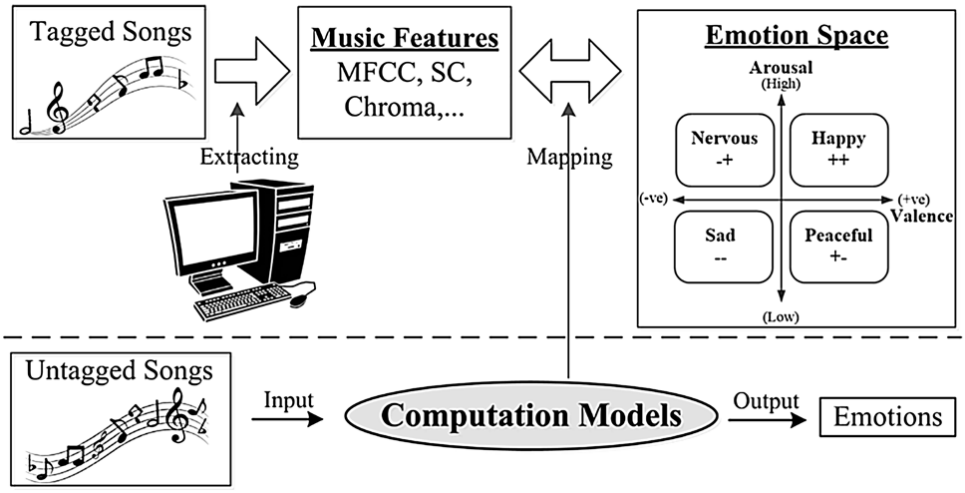


Ilustración 43. Esquema típico de desarrollo de un modelo MER (Yang et al., 2018)

2.4.2 "Ground Truth"

A diferencia de los estudios neurocientíficos que seleccionan unas pocas muestras musicales de corta duración, de tipo orquestal para eliminar variables como contenido verbal, y validadas en estudios con rigor científico (ver sección 2.2.3), en el ámbito computacional de MER se utilizan en general muchas muestras musicales extraídas de música popular de distintos géneros, frecuentemente canciones completas, lo que conlleva muchos problemas respecto a los derechos de propiedad intelectual, que pueden variar según el país, respecto a la calidad de las muestras, respecto a la calidad del etiquetado, y por tanto respecto a la replicabilidad de estudios (Yang et al., 2018).

Para intentar paliar estos problemas, han surgido varias iniciativas. En el ámbito de las ya mencionadas conferencias Music Information Retrieval Evaluation eXchange (MIREX), en 2007, se creó un conjunto de datos, denominado AMC ("audio mood classification"), compuesta por 600 clips de audio de 30 segundos en formato wav, creada con el fin de obtener una línea de base de comparación para investigadores (Hu, Downie, Laurier, Bay, & Ehmann, 2008). Los clips fueron evaluados por 21 voluntarios de 5 países diferentes de la comunidad MIR (8 voluntarios terminaron de evaluar los 250 clips asignados mientras que el resto realizaron entre 1 y 140 evaluaciones). AMC supuso un hito en este ámbito, pero presenta varios problemas: en el etiquetado se utilizaban 5 tipos de emociones discretas, algunas sin base neurocientífica, variaba el número de evaluaciones por clip, y sólo podía accederse a los datos dentro de la comunidad MIREX.

DEAM y MTurk son dos conjuntos de datos públicos, etiquetados segundo a segundo, ampliamente conocidos utilizados en el ámbito MIR/MER.

Ilustración 44. Ejemplo de base de datos de música popular con distintas etiquetas[10]

DEAM[11] (Database for Emotional Analysis of Music) se creó entre 2013 y 2015, en el contexto de los talleres organizados por la comunidad MediaEval, con el objetivo de facilitar los estudios de reconocimiento de la emoción musical. El conjunto de datos DEAM es público, consta de 1802 muestras (entre extractos de 45 segundos y canciones completas), libres de derechos de autor, y anotadas por varios evaluadores en las dimensiones de valencia y activación tanto de forma continua (segundo a segundo) como a lo largo de toda la canción. También incluye metadatos que describen la duración y género musical de los distintos clips. Las anotaciones se hicieron en distintas fases por distintos equipos en los encuentros Mediaeval 2013-2015, cambiando entre fases el set de etiquetas utilizado para evaluar la canción completa. Los autores consideran que las anotaciones en la dimensión de activación son de mejor calidad que las realizadas en la dimensión de valencia, y que las anotaciones de valencia más fiables son las de 2015.

MTurk (Speck, Schmidt, Morton, & Kim, 2011) contiene 240 canciones populares, libres de derechos, y también anotadas de forma continua segundo a segundo en las dimensiones de valencia y activación. En este caso, para facilitar el reclutamiento de evaluadores, se utilizó un crowdsourcing en línea a través de Mechanical Turk, el programa de subcontratación de Amazon. A cada sujeto se le pidió que seleccionara aleatoriamente 11 segmentos entre los propuestos, y que anotara los valores de activación y valencia, segundo a segundo, mediante una interfaz gráfica. Este método adolece de falta de rigor en la selección y entrenamiento de

[10] https://www.gracenote.com/

[11] http://cvml.unige.ch/databases/DEAM/

los evaluadores, por lo que se incrementa el ruido en las anotaciones (Panda, R., 2019), y se requieren pasos adicionales para filtrar los datos. En este caso se utilizó un procedimiento de verificación automática para eliminar las puntuaciones de baja calidad.

GTZAN[12] es otro dataset ampliamente utilizado. Comprende 1000 fragmentos de canciones de 30 segundos, pertenecientes a 10 géneros musicales, etiquetados con un único género, y adolece también de problemas en el proceso de etiquetado (Sturm, 2013).

Million Song Dataset también se utiliza en el ámbito MER. Consiste en una base de datos, accesible públicamente, de metadatos y características audio relativos a un millón de canciones de música popular contemporánea. Los audios clips no son accesibles públicamente, pero se pueden descargar vistas previas. Las anotaciones adolecen del mismo problema de ser el resultado de una anotación libre y abierta a cualquier usuario e incluyen más de 50 campos por canción.

MagnaTagATune (MTAT) es otro dataset utilizado para el benchmarking de modelos de clasificación automática de la música. Contiene anotaciones sobre género, estado anímico, instrumentación para 25.877 segmentos de 29 segundos de audio correspondientes a distintos géneros musicales (desde música clásica a música a música punk). Con el fin de obtener un etiquetado de calidad, las anotaciones se basan en una estrategia de juego en la que dos jugadores tienen que etiquetar un audio clip, el mismo o distinto audio clip (Law, West, Mandel, Bay, & Downie, 2009). Después cada jugador tiene que decidir, en base a las etiquetas del otro jugador, si estaban calificando el mismo audio clip. Sólo se asignan las etiquetas en las que más de dos jugadores coinciden.

Como se ha visto, el proceso de etiquetado de estos conjuntos de datos no es riguroso: los evaluadores y los criterios cambian entre fases de etiquetado, no hay una selección rigurosa de los sujetos evaluadores, tampoco un entrenamiento supervisado de los sujetos para la tarea ni una verificación del procedimiento seguido por los sujetos. Por lo que se pueden tener ciertas reservas sobre la precisión del conjunto de datos obtenido "ground truth" para su uso en los modelos de aprendizaje automático supervisado. Además, se debe añadir, al igual que en los modelos neurocientíficos, la dependencia del etiquetado del modelo de emoción categórico o dimensional elegido. Como ya se ha mencionado, en el modelo categórico la dificultad reside en elegir un número adecuado de categorías de emociones, mientras que, en el modelo dimensional, el problema reside en el significado variable que puede tener para los sujetos los distintos grados de la escala y los conceptos de valencia y activación. En su revisión (Yang et al., 2018) señala que generalmente en MER el método preferente de etiquetado se basa en el modelo dimensional Valencia/Activación.

Otra dificultad es elegir la longitud de los segmentos musicales a evaluar. En el caso de una canción, de unos pocos minutos de duración, el contenido emocional puede fluctuar temporalmente. Normalmente, para conseguir resultados más precisos, se divide la canción en pequeños segmentos, detectando la emoción para cada segmento. Por ejemplo, la longitud típica de segmentación suele ser:

- 25-30 segundos para música popular (Yang et al., 2018)

[12] https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

- Para música clásica, (Xiao, Dellandrea, Dou, & Chen, June 2008) encontraron que los resultados eran óptimos con longitudes de 8-16 segundos (se probaron con 4-8-16-32)

Finalmente, otro problema adicional en el etiquetado de canciones completas se deriva del hecho que una etiqueta puede referirse sólo a un pequeño segmento de la canción, pero queda asignada a la canción completa cuando no se detallan las posiciones exactas a las que se refiere el etiquetado.

(Panda, Renato, Malheiro, & Paiva, 2020) consideran que la mayoría de los trabajos MER no obtienen buenos resultados en parte por la falta de uniformidad en los datasets utilizados, algunos privados, lo que dificulta la replicabilidad y evaluación comparativa de diferentes estudios. Además considera que la taxonomía definida para anotar las emociones carece del apoyo teórico de la psicología musical.

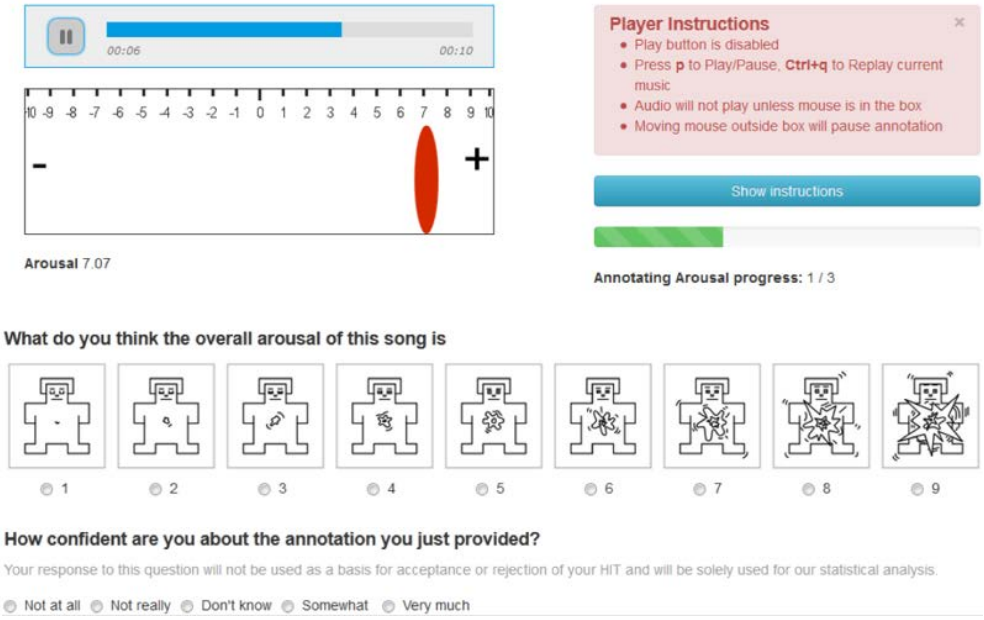


Ilustración 45. Ejemplo de cuestionario MER para etiquetar una canción (Anna Aljanaki, Yi-Hsuan Yang, & Mohammad Soleymani, 2017)

2.4.3 Extracción de características musicales

Las características de audio representan la información extraída de una señal de audio y se utilizan en distintos ámbitos como los efectos audio digitales, las transcripciones automáticas, o el propio MER. La extracción de características entra en el terreno MIR (Music Information Retrieval). En los últimos años se han desarrollado diferentes frameworks especializados capaces de extraer un gran número de características de audio, algunos más orientados para su uso en la industria y otros más orientados a la investigación, en general. Ejemplos de estos

frameworks son Essentia[13], MirToolbox[14] y Librosa[15], ampliamente utilizados en la investigación MER (Panda, 2019).

Essentia (Bogdanov et al., 2013) es una librería open-source desarrollada en C++ específicamente para la recuperación de la información musical en base al análisis audio. Contiene numerosos algoritmos incluyendo descriptores espectrales, temporales, tonales y descriptores musicales de alto nivel como género o instrumentación. Está desarrollada con el objetivo de ser una plataforma robusta, y eficiente en términos de velocidad y uso de memoria, útil tanto para aplicaciones industriales como para la investigación.

MirToolbox (Lartillot & Toiviainen, 2007) se desarrolló inicialmente con el objetivo principal de investigar la relación entre las características musicales y la emoción y actividad neuronal inducidas. Está construido sobre Matlab y ofrece un amplio conjunto integrado de funciones dedicadas a la extracción de características musicales de archivos de audio. El diseño es modular: los diferentes algoritmos se descomponen en una serie de funciones elementales. Estas funciones elementales engloban distintas estrategias y variantes de extracción de características, de forma que se pueden seleccionar y parametrizar distintas aproximaciones a un mismo problema. MirToolbox se ha utilizado en algunos de los mejores algoritmos clasificación de emociones desarrollados en los encuentros MIREX (Panda, R. & Paiva, 2012) (Wang, J., Lo, Jeng, & Wang, 2010).

Librosa es un paquete de Python de código abierto para análisis de audio y música. Ha sido diseñado con interfaces y nombres de variables estandarizadas, por compatibilidad con implementaciones de referencia existentes. Las características disponibles de Librosa son principalmente características espectrales, aunque también están disponibles descriptores de características rítmicas. Es un paquete que se integra muy bien con las librerías Python en el desarrollo de redes neuronales.

Tomando como ejemplo MirToolbox podemos ver cómo se desarrolla la extracción de características musicales de las muestras de audio (Lartillot, 2019). Se parte de la señal audio digital, en general en formato *wav* o mp3 y con una frecuencia de muestreo de 44.000 Hz. Sobre esta señal se pueden realizar distintas operaciones básicas como son:

- Descomposición en ventanas temporales: El análisis de la señal temporal permite obtener valores medios. Para tener en cuenta la evolución dinámica de la señal se estudia a partir de pequeñas ventanas temporales (frames) desplazadas cronológicamente a lo largo de toda la señal.

- Descomposición en bandas de frecuencias: La descomposición se realiza mediante una serie de filtros, que seleccionan un rango particular de valores de frecuencia. Esta transformación simula el proceso de la percepción humana, que procesa el sonido en base a bandas de frecuencia en la cóclea.

[13] https://essentia.upf.edu/

[14] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox

[15] https://librosa.org/doc/latest/index.html

- Envolvente: El cálculo de la envolvente muestra la forma exterior global de la señal. Es particularmente útil para mostrar la evolución en el tiempo de la señal.

- Análisis cepstral: Este análisis permite reducir una serie armónica de frecuencias a su frecuencia fundamental

- Función de autocorrelación: Se utiliza para evaluar periodicidades en señales, autocorrelando distintos segmentos de la señal de audio.

- Distancia entre ventanas consecutivas: Mide la distancia entre ventanas sucesivas de la señal de un descriptor determinado, por ejemplo, el flujo espectral mide las diferencias en el espectro. Los picos obtenidos indican la posición temporal de contrastes importantes en la señal. Son útiles en la detección de ritmos.

- Detección de picos: La detección de picos permite estimar la posición de eventos musicales y permite segmentar la señal de audio en base a esos eventos.

- Cálculo de Espectrogramas: La descomposición de la señal audio a lo largo de las frecuencias se puede realizar usando la transformada rápida de Fourier (STFT). La STFT se calcula en sucesivas ventanas temporales de la señal, parcialmente solapadas. El espectrograma resultante es una gráfica tridimensional que representa la energía del contenido frecuencial de la señal a lo largo del tiempo. También se utilizan los espectrogramas Mel, en los que la escala de frecuencias se transforma en la escala Mel auditiva, o los Mel Frequency Cepstral Coefficients (MFCC) que se obtienen a partir de la escala Mel, tomando el logaritmo de las energías de la escala Mel y aplicando la transformada de coseno discreta a esos logaritmos. En los espectrogramas CQT, las frecuencias se representan en escala logarítmica, correspondiendo a las diferentes notas y se muestran los niveles de energía estimados en los distintos tipos de tonos musicales. El contraste espectral basado en octavas (OSC) considera el pico y el valle espectrales en cada subbanda En general los picos corresponden a componentes armónicos y los valles a componentes no armónicos o ruido, por los que la diferencia entre picos refleja la distribución del contraste espectral.

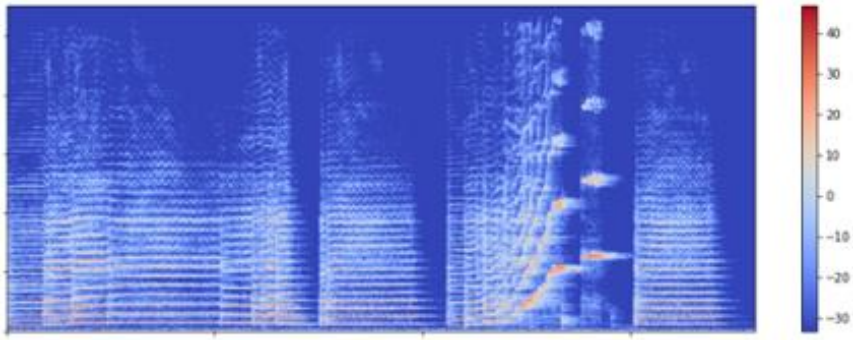


Ilustración 46. Espectrograma STFT

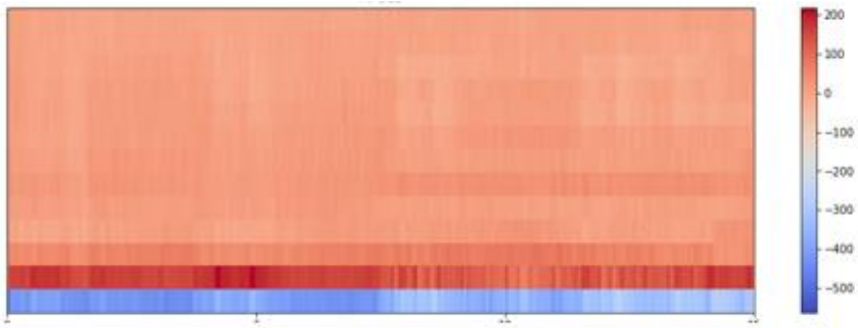


Ilustración 47. Espectrograma MFCC

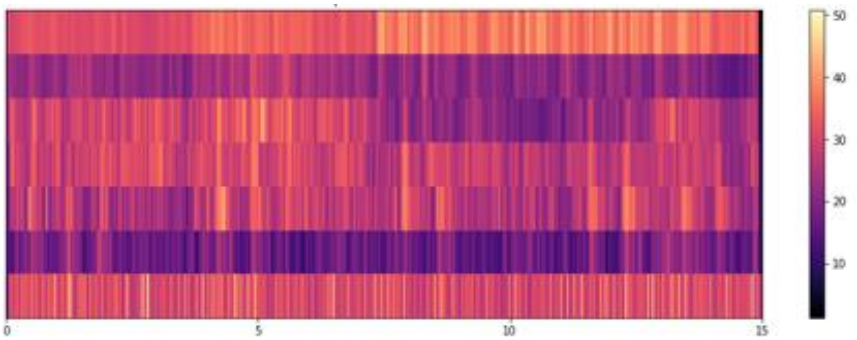


Ilustración 48. Espectrograma OSC

A su vez estas operaciones básicas son la base del desarrollo de algoritmos complejos para extraer distintas características de la señal audio. Estas características se pueden obtener sobre la señal completa, o sobre los distintos frames en los que se haya descompuesto la señal si se quiere obtener el detalle de su evolución temporal. En el cuadro siguiente se muestran parte de las características que se han desarrollado relacionadas con parámetros musicales.

Características audio relacionadas con la DINÁMICA MUSICAL (graduaciones de intensidad del sonido)	
Energía cuadrática media	Amplitud media de la señal.
Distribución temporal de la energía	Porcentaje de ventanas con energía cuadrática por debajo de la energía cuadrática media de toda la señal.
Segmentación en base a silencios	Segmentación de la señal en base a posiciones en las que la energía cuadrática cae por debajo de un umbral determinado.
Características audio relacionadas con el RITMO y TEMPO MUSICAL (el tempo hace referencia a la velocidad de ejecución)	
Espectro de pulsos	Medida de similitud acústica en función de intervalos. La música muy estructurada o repetitiva tendrá fuertes picos de espectro en los tiempos de repetición.
Ubicación temporal de eventos	Identificación de picos de energía.
Ataque/Decaimiento	Identifican las características de estos picos, duración y gradiente de las pendientes de inicio y fin de los picos identificados.
Densidad de eventos	Estimación del número de eventos por segundo.
Tempo	Estimación del tempo a partir de las periodicidades de la curva de detección de eventos.
Cambios de Tempo	Estimación de cambios en el tempo.
Claridad de pulso	Estimación del grado de claridad del ritmo, en base a la fuerza de los pulsos estimados.
Características audio relacionadas con el TIMBRE MUSICAL (calidad percibida del sonido, permite por ejemplo distinguir voces o instrumentos)	
Ataque/Decaimiento Duración/Pendiente/Salto	La duración y pendiente del tiempo de ataque y decaimiento de los eventos puede dar información sobre las características tímbricas.
Duración de los eventos	Duración del evento entre el fin de la fase de ataque y el inicio de la fase de decaimiento.
Tasa de cambio de signo	Número de veces que la señal cambia de signo, es un indicador del ruido de la señal.
Energía alta frecuencia/ Brillantez	Frecuencia a partir de la cual un % determinado de la energía total se encuentra por debajo de esa frecuencia. Medida de la cantidad de energía por debajo de esa frecuencia.
Descriptores estadísticos de la distribución espectral	Centro de masa, distancia entre las sucesivas ventanas, asimetrías, de la distribución espectral.

Características audio relacionadas con el TONO y TONALIDAD MUSICAL (frecuencia fundamental percibida de un sonido y tonalidad de la escala musical)	
Estimación de tonos	Algoritmos de detección de notas.
Inarmonicidad	Estimación de la cantidad de energía fuera del ideal armónico.
Chroma	Distribución de energía entre los tonos o clases de tonos estimados en la señal audio. Es un vector de 12 posiciones correspondientes a los 12 semitonos entre las notas C y G#.
Probabilidad de las tonalidades	Probabilidad de las tonalidades estimadas en base a las correlaciones entre los coeficientes del cronograma obtenidos y los cromogramas correspondientes a las distintas tonalidades.

Tabla 5. Características audio y parámetros musicales

Aunque se han establecido estas relaciones entre características audio y parámetros musicales todavía se requiere mucha investigación para entender realmente estas relaciones (Panda et al., 2020). Por otra parte, hay que tener en cuenta que parte de estos parámetros se han desarrollado para resolver otros problemas como por ejemplo la transcripción automática de música, o separación de pistas e instrumentos, y no está claro cuáles de estos parámetros tienen relevancia emocional.

2.4.4 Selección de características musicales

La selección de las características adecuadas para los modelos MER es una tarea difícil, porque no está claro que los algoritmos desarrollados hasta ahora sean los idóneos y porque todavía no hay una identificación clara de qué características audio son relevantes pare capturar el contenido emocional de una música (Panda et al., 2020).

Así, en los modelos de aprendizaje automático en el ámbito MER se han utilizado distintos conjuntos de estas características para establecer un modelo predictivo. De acuerdo con la revisión de (Yang et al., 2018) las características generalmente más utilizadas en MER son:

- Mel-frequency Cepstrum Coefficients (MFCCs)
- Descriptores estadísticos del espectro
- Contraste espectral basado en octavas (OSC)
- Chroma

Pero en general el conjunto de características empleado es muy complejo y agrupa numerosos elementos. Consideremos como ejemplo el modelo desarrollado para la detección de emociones por (Eerola et al., 2009) para el encuentro Mirex 2009, y desarrollado en base a las características ofrecidas por MirToolbox. Trabajaron considerando respecto al modelo categórico 5 emociones básicas (alegría, tristeza, ira, miedo y ternura), y respecto al modelo

dimensional las 3 dimensiones de (Schimmack & Grob, 2000): valencia (positiva-negativa), activación (baja-alta) y tensión (baja-alta).

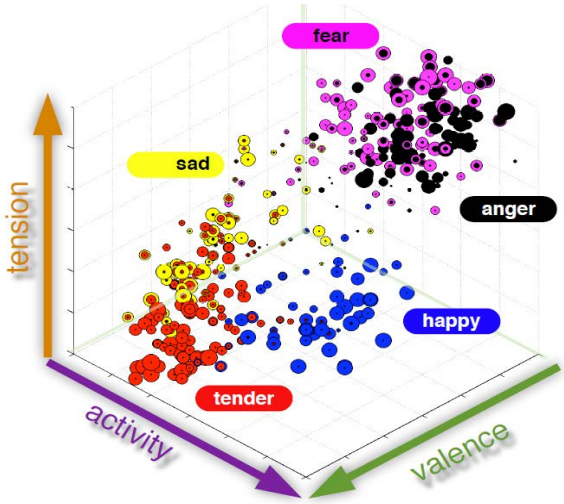


Ilustración 49. Modelo Mixto utilizado por (Eerola et al., 2009)

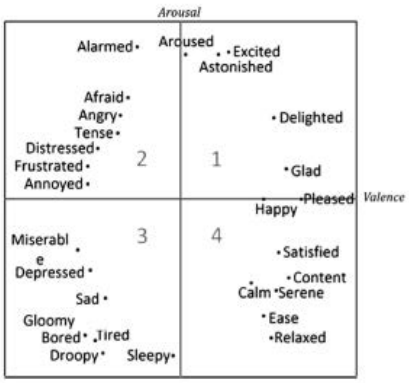
En el estudio estadístico de regresión lineal múltiple que realizaron, estudiaron la correlación entre emociones básicas y dimensiones emocionales y varios conjuntos de características musicales. Las máximas correlaciones obtenidas se resumen en la siguiente figura, donde el color rojo representa características espectrales, el color azul características tonales, y el verde características de energía.

	HAPPY	SAD	TENDER	ANGER	FEAR
Maximum value of summarized fluctuation	**β = 0.7438**				β = -0.2538
Spectral spread averaged along frames	β = -0.3965	β = 0.4324			
Standard deviation of the position of the maximum of the unwrapped chromagram	β =0.4047	β = -0.3137			
Key clarity (2nd output of mirkey) averaged along frames	**β = 0.7780**		**β = 0.5192**	**β = -0.5802**	**β = -0.9860**
Mode averaged along frames	**β = 0.6620**	**β = -0.5201**			
Averaged HCDF (β = -0.6017)		**β = -0.6017**	β = -0.3995		β = -0.3144
Averaged novelty from wrapped chromagram (β = 0.4493)		β = 0.4493			
Spectral centroid averaged along frames (β = -0.2709)			β = -0.2709		
Standard deviation of roughness (β = -0.4904)			β = -0.2709		
Averaged spectral novelty			β = 0.3391		
Roughness averaged along frames				**β = 0.5517**	
Entropy of the smoothed and collapsed spectrogram, averaged along frames				β = 0.2821	
Standard deviation of RMS along frames					β = 0.4069
Averaged attack time					**β = -0.6388**
Averaged novelty from unwrapped chromagram				β = -0.2971	

Ilustración 50. Características audio más relevantes para 5 emociones (Eerola et al., 2009)

Otro ejemplo que refleja la variedad y complejidad de las características musicales utilizadas en los modelos MER, es el conjunto de parámetros propuesto por (Panda et al., 2020) como mejores predictores para detección de la correspondencia de la emoción musical con los cuatro cuadrantes del modelo valencia-activación. En este estudio los investigadores partieron de 1702 características que redujeron a 989 descartando características que correlacionaban entre ellas. Esta gran cantidad de características se debía en parte a que se consideraban distintos segmentos temporales para cada canción dentro del dataset de 900 canciones utilizado en el estudio.

Según las conclusiones de (Panda et al., 2020) parece que el uso de características espectrales da mejores resultados que el uso de características dinámicas, de ritmo o tonales.



Top 5 Features for Each Quadrant Discrimination

Q	Feature	Type	Concept	Weight
Q1	FFT Spectrum - Spectral 2nd Moment (median)	base	Tone Color	0.1467
	Transitions ML1 -> ML0 (Per Sec)	novel	Texture	0.1423
	MFCC1 (mean)	base	Tone Color	0.1368
	Transitions ML0 -> ML1 (Per Sec)	novel (voice)	Texture	0.1344
	Fluctuation (std)	base	Rhythm	0.1320
Q2	FFT Spectrum - Spectral 2nd Moment (median)	base	Tone Color	0.2528
	Roughness (std)	base	Tone Color	0.2219
	Rolloff (mean)	base	Tone Color	0.2119
	MFCC1 (mean)	base	Tone Color	0.2115
	FFT Spectrum - Average Power Spectrum (median)	base	Tone Color	0.2059
Q3	Spectral Skewness (std)	base	Tone Color	0.1775
	FFT Spectrum - Skewness (median)	base	Tone Color	0.1573
	Tremolo Notes in Cents (Mean)	novel	Tremolo	0.1526
	Linear Spectral Pairs 5 (std)	base	Tone Color	0.1517
	MFCC1 (std)	base	Tone Color	0.1513
Q4	FFT Spectrum - Skewness (median)	base	Tone Color	0.1918
	Spectral Skewness (std)	base	Tone Color	0.1893
	Musical Layers (Mean)	novel	Texture	0.1697
	Spectral Entropy (std)	base	Tone Color	0.1645
	Spectral Skewness (max)	base	Tone Color	0.1637

Ilustración 51. Características audio más relevantes (resaltadas en amarillo) para 4 cuadrantes (Panda et al., 2020)

Pero en general la mayoría de los trabajos MER no obtienen buenos resultados, entre otros factores, porque no está claro que las características audio utilizadas sean lo suficientemente relevantes para el problema, por la falta de uniformidad en las características utilizadas, y por la falta de características que expresen conceptos de más alto nivel como las técnicas expresivas de la música (vibratos, trémolos, …), (Panda et al., 2020). Por otra parte, los experimentos han demostrado que el uso de más características no conduce a una mejora del rendimiento (Yang et al., 2018).

2.4.5 Modelos de aprendizaje

Los modelos de aprendizaje MER se basan en mapear las características audio de bajo nivel de las muestras de audio con la emoción etiquetada en cada muestra. Los modelos difieren en función de las hipótesis de partida.

De acuerdo con la revisión de (Yang et al., 2018) en los modelos que utilizan muestras etiquetadas de acuerdo con el modelo emocional categórico, se utilizan en general algoritmos de clasificación como Gaussian Mixture Models (GMM), K-Nearest Neighbour (KNN), Support Vector Machines (SVM), siendo SVM el clasificador que obtendría los mejores resultados (Han, Rho, Dannenberg, & Hwang, 2009) (Panda et al., 2020). También se han utilizado clasificadores multietiquetas como Multilabel SVM en los casos de canciones o fragmentos musicales con varias etiquetas en función del segmento temporal.

Mientras que en los modelos que utilizan muestras etiquetadas de acuerdo con el modelo emocional dimensional, se utilizan modelos regresión como Support Vector Regression (SVR), o Gaussian Process Regression.

A pesar de los esfuerzos en el ámbito MER, los resultados siguen siendo limitados. Las soluciones desarrolladas en MER todavía no son capaces de resolver los problemas más simples como la clasificación en 4 o 5 emociones (Panda et al., 2020). En la revisión de (Yang et al., 2018) se indica que la mayor precisión en clasificación de emociones se había obtenido en MIREX 2011 con un 69,5% considerando 5 categorías emocionales. Cuando se añaden más categorías emocionales disminuye la precisión. (Panda et al., 2020) comparan distintos resultados utilizando SVM para clasificar fragmentos musicales en los cuatro cuadrantes del modelo valencia-arousal, método que consideran da mejores resultados que otros métodos, obteniendo precisiones de hasta 76,4%.

En los últimos años, estos modelos basados en la selección de características se están sustituyendo por el uso de las redes neuronales convolucionales (CNN) con resultados prometedores, utilizando como parámetros de entrada espectrogramas como STFT o Mel, por lo que vamos a analizarlos con más detalle.

2.4.6 Modelos de Redes neuronales CNN

Las redes neuronales convolucionales (ConvNets o CNN) son un tipo de redes neuronales muy eficaces en el ámbito de reconocimiento de imágenes. El desarrollo de estas redes se inició en la década de los 90 con los trabajos de Yann LeCun (Lecun, Bottou, Bengio, & Haffner, 1998), que definió la denominada arquitectura LeNet. Desde entonces, se han propuesto continuamente mejoras a esta primera arquitectura y actualmente las CNN se consideran la arquitectura más importante en el reconocimiento visual.

Los elementos principales de una red CNN (Ujjwalkarn, 2016) son:

- **Input:** los datos (píxeles) sin procesar de la imagen.

- **Convolution layer**: extrae características ('features') de las imágenes de entrada.
- **Pooling layer**: reduce la dimensionalidad de las imágenes, pero retiene la información más importante.
- **Fully connected layer**: funciona como un clasificador.

Esencialmente, cada imagen puede representarse como una matriz de valores de píxeles. Una imagen de una cámara digital estándar se compone de 3 canales: rojo, verde y azul, y puede representarse como 3 matrices de 2 dimensiones, cada una con valores de píxeles en el rango de 0 a 255. Una imagen en escala de grises tiene un único canal, variando el valor de cada píxel en la matriz de 0 a 255: cero indica negro y 255 indica blanco.

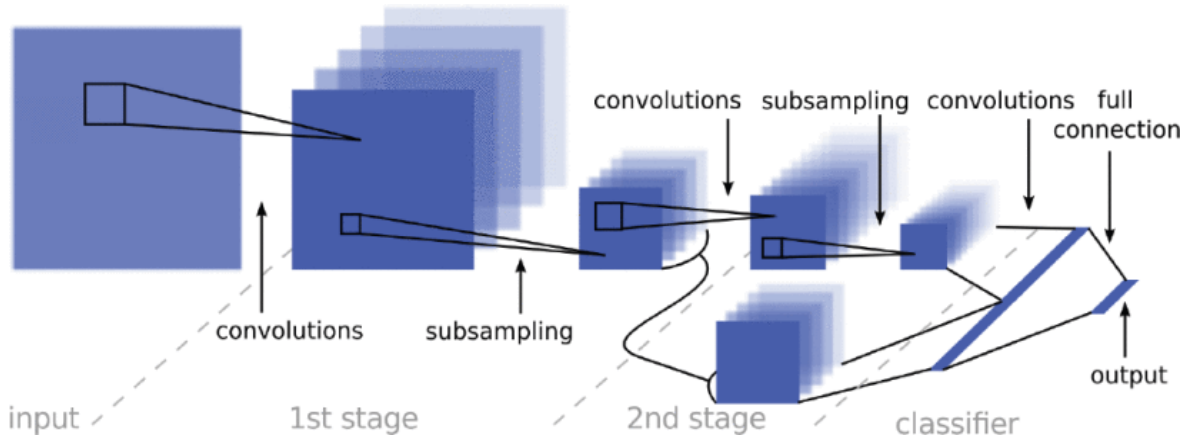


Ilustración 52. Arquitectura CNN (Sermanet & LeCun, 2011)

En la fase de convolución (Convolution Layer) aplicamos distintos filtros sobre la imagen original, desplazando estos filtros a lo largo de la imagen. Estos filtros, denominados 'filter', 'kernel' or 'feature detector', son matrices cuadradas que computan el producto escalar entre estas y los segmentos de imágenes que van recorriendo. La convolución conserva la relación espacial entre píxeles extrayendo las características de la imagen considerando pequeños cuadrados de datos de entrada. La matriz de salida de cada filtro se denomina 'Convolved Feature' o 'Activation Map' o 'Feature Map'.

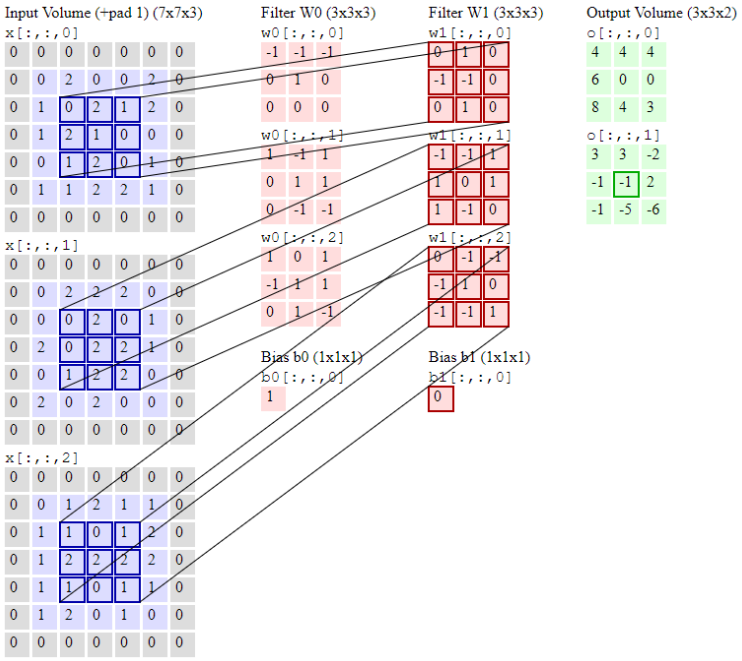


Ilustración 53. Fase de convolución (Stanford University, 2021)

En la práctica, una CNN aprende los valores de estos filtros por sí misma durante el proceso de entrenamiento, aunque se deben especificar parámetros como el número de filtros o el tamaño del filtro. Cuantos más filtros, más características de la imagen se extraen.

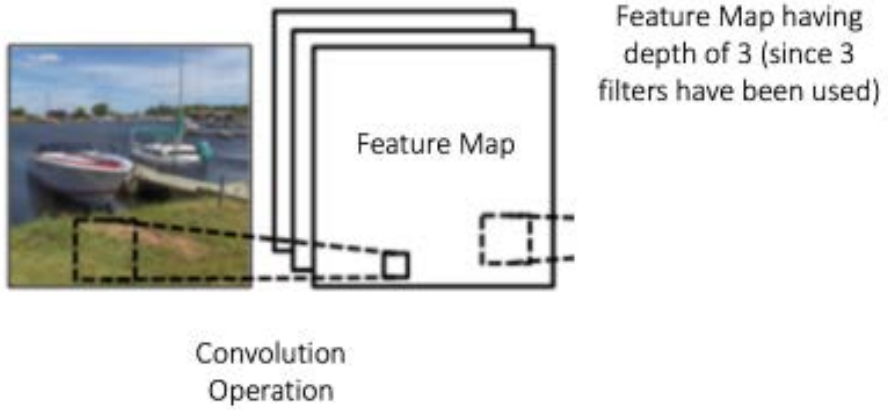


Ilustración 54. Operación de convolución (Ujjwalkarn, 2016)

El tamaño de los "feature maps" obtenidos depende de 3 parámetros que deben decidirse también antes del proceso de entrenamiento:

- Profundidad, que corresponde a la cantidad de filtros que usamos para la operación de convolución.

- "Stride": paso, número de píxeles que se usan para desplazar la matriz de filtro sobre la matriz de entrada.
- "Zero-padding": Decide si se completa la matriz de entrada con ceros alrededor del borde de la imagen, para poder aplicar los filtros a los elementos del borde de la imagen.

La fase "Pooling Layer" reduce la dimensionalidad de cada mapa de características, pero conserva la información más importante. La agrupación espacial puede basarse en máximos, media, suma, etc. Por ejemplo, se pueden definir ventanas de 2×2 y tomar el elemento mayor de esa ventana, o el promedio de esa ventana. La función de esta fase es reducir progresivamente el tamaño espacial de la imagen de entrada en agrupaciones más pequeñas y manejables reduciendo el número de parámetros y cálculos en la red, consiguiendo que la red sea invariante a pequeñas transformaciones, distorsiones y traslaciones en la imagen de entrada (una pequeña distorsión en la entrada no cambiará la salida de Pooling, ya que tomamos el valor máximo / promedio en un vecindario local).

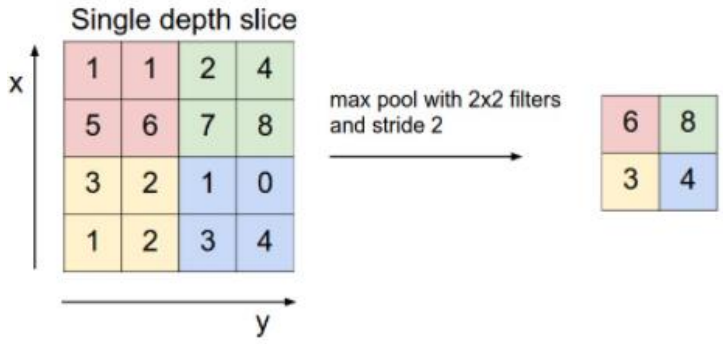


Ilustración 55. Ejemplo de Max Pooling (Stanford University, 2021)

La capa "Fully Connected Layer" es un perceptrón multicapa tradicional que usa una función de activación en la capa de salida, y en la que cada neurona de la capa anterior está conectada a cada neurona de la siguiente capa. La salida de las capas convolucional y "Pooling Layer" representan características de alto nivel de la imagen de entrada. El propósito de la capa "Fully Connected Layer" es utilizar estas características para clasificar la imagen de entrada en varias clases según el conjunto de datos de entrenamiento.

Desde las primeras redes convolucionales, ha habido una tendencia hacia arquitecturas cada vez más profundas para mejorar el rendimiento, con un número creciente de niveles convolucionales (Fleuret, 2021).

Network	Nb. layers
LeNet5 (leCun et al., 1998)	5
AlexNet (Krizhevsky et al., 2012)	8
VGG (Simonyan and Zisserman, 2014)	11–19
GoogleLeNet (Szegedy et al., 2015)	22
Inception v4 (Szegedy et al., 2016)	76
Resnet (He et al., 2015)	34–152
Resnet (He et al., 2016)	1001
Resnet (Huang et al., 2016)	1202

Ilustración 56. Evolución de las CNN (Fleuret, 2021)

Pero la acumulación de capas conlleva problemas como la degradación del proceso de aprendizaje, como muestran (He, Zhang, Ren, & Sun, 2016) con pruebas sobre la base de datos CIFAR-10[16] (compuesta por 60.000 imágenes en color categorizadas en 10 clases), comparando redes neuronales de 20 y 56 capas.

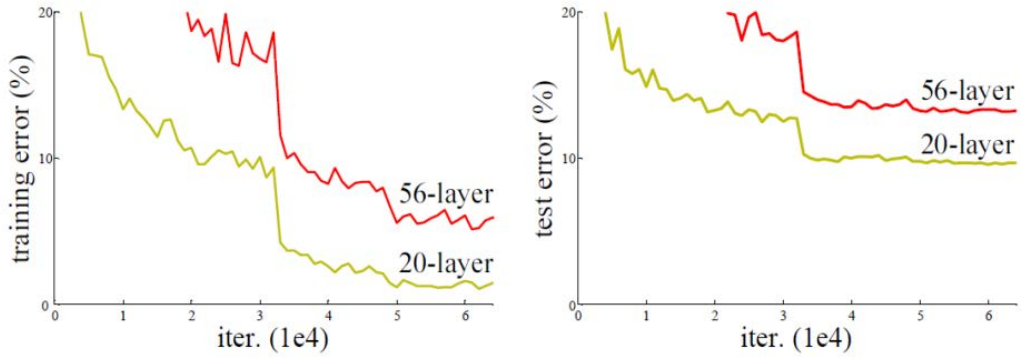


Ilustración 57. Training error en función del número de capas

Para reducir este fenómeno de degradación, (He et al., 2016) propusieron las denominadas **redes residuales (ResNets)**. ResNet se basa en la idea de alimentar la salida de dos capas convolucionales consecutivas y la entrada inicial, como entrada en la siguiente capa para mejorar el flujo de información entre capas.

[16] https://www.cs.toronto.edu/~kriz/cifar.html

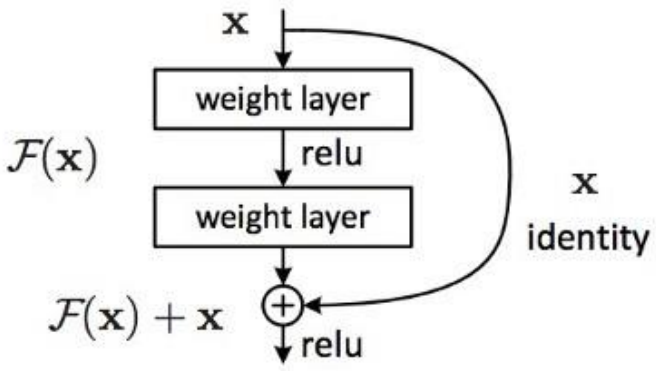


Ilustración 58. Bloque de construcción ResNet (Culurciello, 2017)

Estos bloques utilizan dos convoluciones 3x3. Una red ResNet profunda es una arquitectura modular que se construye a partir de múltiples unidades de construcción ResNet.

Squeeze-and-Excitation Networks (SENets) introducen otro tipo de bloque a las CNN para mejorar la dependencia entre canales. En una red CNN cada "feature map" F se genera a partir de los diferentes canales de entrada formando un único canal de salida de la convolución. Los SENets tienen como objetivo ponderar cada canal en función de su relevancia. La operación Squeeze-and-Excitation se ejecuta al final de la convolución para obtener un peso para cada canal de la salida. Una vez que se dispone de los pesos por cada canal, la salida final de la convolución se obtiene reevaluando el "feature map" F de acuerdo con esos pesos (Sermanet & LeCun, 2011).

Otro de los problemas que se ha tratado de resolver es la variación en tamaño de la información importante de una imagen. Por ejemplo, dependiendo de la imagen, un perro que intentamos identificar puede ocupar una parte muy pequeña de la imagen o toda la imagen. Esto dificulta la decisión sobre el tamaño de los filtros, siendo mejor utilizar filtros grandes si la información se distribuye globalmente en la imagen, o filtros pequeños si la información ocupa un espacio pequeño de la imagen.

La denominada arquitectura **Inception** intenta dar una solución a este problema incorporando filtros de distinto tamaño en un mismo nivel. Las salidas se concatenan y se envían al siguiente bloque Inception. La idea es que estructuras paralelas de convoluciones con diferentes filtros pueden capturar distintas características a distintos niveles 1×1, 3×3 and 5×5. La Ilustración **59** muestra una arquitectura Inception que combina paralelamente filtros convolucionales 1×1, 3×3 y 5×5. El uso de bloques convolucionales 1×1 permite reducir el número de características para evitar "cuellos de botella" computacionales.

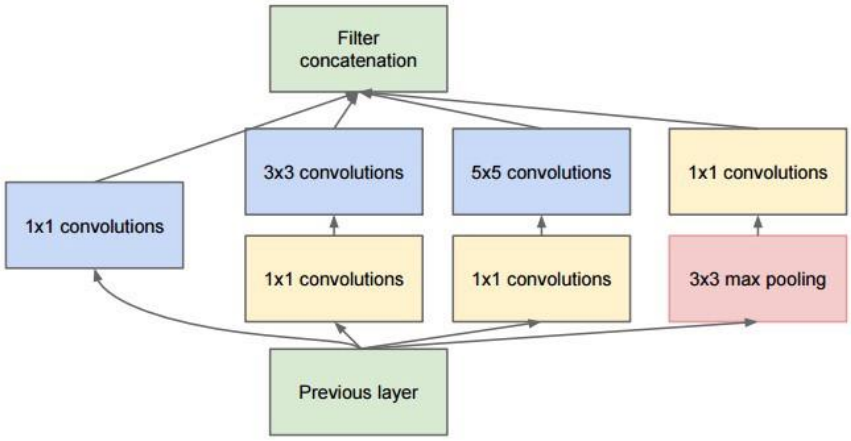


Ilustración 59. Módulo Inception (Culurciello, 2017)

2.4.7 Aplicación de las redes CNN en MER

El éxito de las redes neuronales en el reconocimiento de imágenes ha despertado el interés en la aplicación de estas redes en el ámbito MER utilizando como parámetros de entrada equivalentes a las imágenes los espectrogramas, como STFT o Mel, obtenidos a partir de las muestras de audio.

En 2010 (Li, Chan, & Chun, 2010) utilizaron como enfoque novedoso una red CNN para clasificación de género musical utilizando espectros MFCC (Mel Frequency Cepstral Coefficients) mostrando que las CNN tenían un gran potencial para extraer características de las muestras audio.

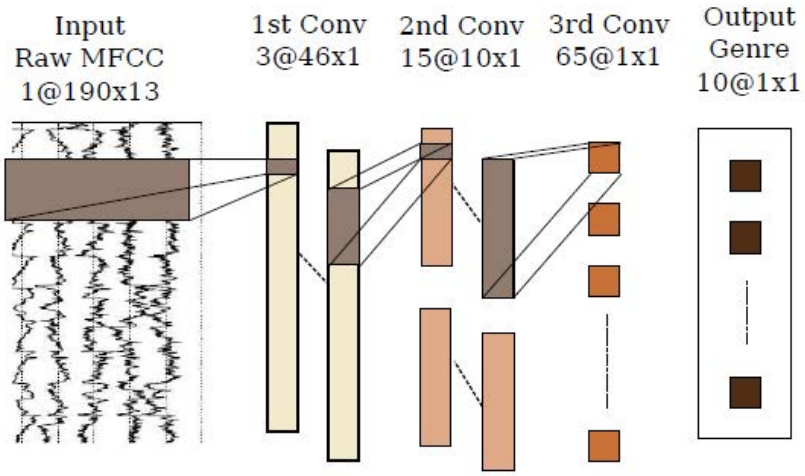


Ilustración 60. Red CNN utilizada por (Li et al., 2010)

Con el mismo objetivo de clasificación de género musical (Zhang et al., 2016) compararon el rendimiento de 2 redes convolucionales similares, incluyendo en una de ellas una conexión tipo Resnet, llegando a alcanzar precisiones del 87% con la base de datos musical GTZAN[17].

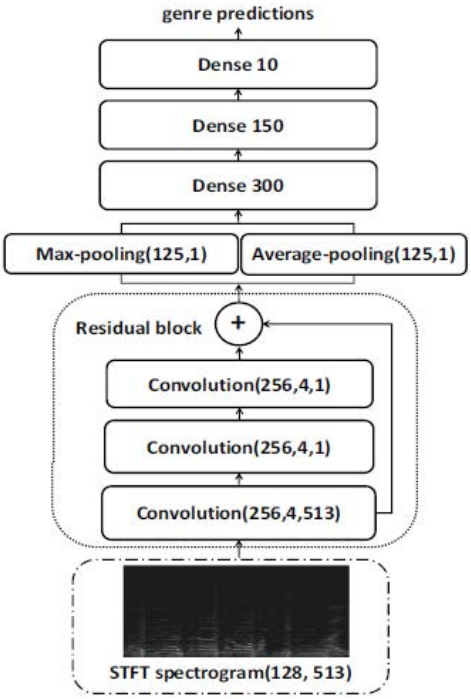


Ilustración 61. Arquitectura CNN-ResNet utilizada por (Zhang et al., 2016)

Recientemente (Won, Ferraro, Bogdanov, & Serra, 2020) han realizado un benchmarking de las últimas arquitecturas CNN propuestas en la clasificación de género musical. Para la clasificación utilizaron etiquetado múltiple, considerando hasta 50 etiquetas relativas al género, al estilo, o al estado de ánimo reflejado en las distintas muestras. Dentro de los modelos CNN evaluados, los modelos que utilizaban las muestras más pequeñas de audio (de unos 3 segundos de duración) tenían las siguientes características:

1. Short Chunk CNN: espectrogramas Mel de 3,69 segundos como datos de entrada, CNN con capas convolucionales 2D y filtros cuadrados (Ilustración 62a), y variante añadiendo conexiones ResNet.
2. Musicnn: espectrogramas Mel de 3 segundos como datos de entrada, filtros verticales y horizontales, concatenados, y seguidos de CNN con capas convolucionales 1D (Ilustración 62b).

[17] https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

3. Sample Level CNN: muestras directas audio de 3,69 segundos, CNN con capas convolucionales 1D (Ilustración 62c), y variante añadiendo bloques SE (Squeeze-and-Excitation).

La siguiente figura ilustra las distintas presentaciones de los datos y filtros utilizados en estos modelos. En el caso (a) se utilizan filtros cuadrados sobre un espectrograma Mel. En el caso (b) se utilizan filtros verticales y filtros horizontales para capturar tanto relaciones armónicas como temporales sobre espectrogramas Mel. En el caso (c) se utilizan como datos de entrada directamente las muestras de audio.

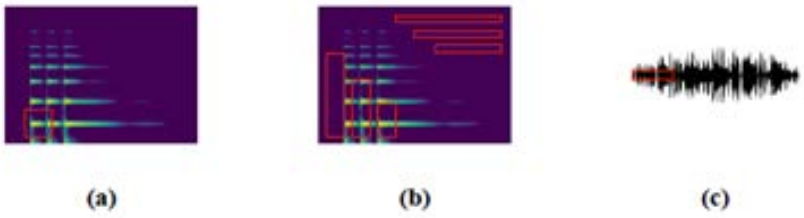


Ilustración 62. Distintos modelos evaluados por (MinzWon et al., 2020)

El primer modelo Short Chunk CNN fue el que obtuvo mejores resultados, una sencilla red CNN sobre espectrogramas Mel (sin la variante ResNet). Las redes CNN entrenadas directamente con las muestras de audio también obtuvieron resultados buenos. En general, los resultados mostraron que las arquitecturas más simples aplicadas sobre fragmentos musicales cortos (de unos 3 segundos) obtenían los mejores resultados.

Para mejorar los resultados de las redes CNN se están considerando actualmente enfoques que incorporan información adicional a la información extraída únicamente del audio. La información adicional puede ser la letra en el caso de las canciones, incluyendo redes utilizadas en el reconocimiento de voz (Pandrea, Gómez Cañón, & Herrera Boyer, 2020), o el uso de metadatos como las etiquetas o comentarios de los oyentes (Korzeniowski et al., 2020).

2.4.8 Resumen y Limitaciones

Como se ha visto, los sistemas MER adolecen de muchas limitaciones, principalmente porque no existe un framework común de experimentación, lo que hace muy difícil la replicabilidad de los estudios.

Primero, está la ausencia de *datasets* públicos, consensuados y adecuadamente validados, a lo que se añade la dificultad del proceso de anotación, generalmente realizado en base a

taxonomías variadas, no basadas en los estudios neurocientíficos, y en un entorno poco controlado. Tampoco hay uniformidad en la longitud de las muestras de audio estudiadas.

Segundo, está la dificultad de seleccionar y obtener las características de audio significativas para la captura de la emoción musical, ya que no está clara la bondad de los algoritmos, ni qué características son relevantes para la detección de la emoción asociada a una música.

Y, finalmente, está la dificultad de crear modelos de aprendizaje automático robustos para capturar las relaciones música-emoción. Los resultados apenas superan el 75% de precisión siempre que se consideren 4 o 5 emociones básicas, o 4 cuadrantes en el modelo dimensional.

Por otra parte, los modelos computacionales están muy lejos de reflejar los procesos cerebrales de percepción de la emoción. Los estudios neurocientíficos sobre la emoción musical muestran un cerebro capaz de procesar la emoción musical con mucha precisión, en muy poco tiempo, y en base a pocos parámetros, como puede ser por ejemplo la tonalidad de un acorde en modo menor o mayor, independientemente del instrumento y tímbrica empleados. Sin embargo, los modelos computacionales parecen muy alejados de estos estudios, enfrascados en la búsqueda de correlaciones entre múltiples características audio de una muestra digital y unas emociones etiquetadas con poco rigor científico.

Los modelos computacionales parecen perderse en el laberinto de los algoritmos MIR y los algoritmos de aprendizaje automático, distanciándose de los fundamentos neurocientíficos de la percepción de la emoción musical. En este sentido, las redes neuronales CNN parecen ofrecer un enfoque más sencillo al partir de los datos audio originales sin un filtrado previo de las características, salvo los que ocurren en la generación de los espectrogramas.

3 EL CANAL TÁCTIL COMO CANAL DE TRANSMISIÓN

3.1 Introducción

En este capítulo se detallan las experimentaciones desarrolladas con el objetivo de validar las tres primeras hipótesis planteadas en esta investigación:

1. El subtitulado accesible textual no transmite la información que aporta la música de forma inmediata a través de la emoción.
2. El tacto puede ser un canal de transmisión alternativo de emociones musicales básicas.
3. Los parámetros musicales pueden transmitirse de forma similar, aunque más limitada, mediante estimulación vibro-táctil

El objetivo general de las experimentaciones era presentar a los sujetos, con y sin discapacidad auditiva, un material audiovisual asociado a distintos estímulos sonoros, visuales (subtítulos, emoticonos), y vibro táctiles, al tiempo que se registraba su actividad cerebral mediante EEG, para comparar la actividad cerebral promedio generada por los distintos tipos de estímulos.

Se desarrollaron dos experimentaciones basadas en la misma metodología. La primera tenía como objetivo comparar la actividad cerebral generada por el audio en un material audiovisual versus la actividad cerebral generada por los subtítulos correspondientes. La segunda experimentación tenía como objetivo comparar la actividad cerebral generada por la música en un material audiovisual versus la actividad generada por una estimulación vibro táctil.

En primer lugar, se describen las características comunes de la metodología, y a continuación, se describe el detalle de cada una de las dos experimentaciones.

3.2 Metodología y materiales

La metodología experimental se basa en la configuración experimental más extendida en la experimentación científica en el ámbito de la emoción, consistente en la presentación a los sujetos experimentales de una serie de estímulos emocionales controlados, mientras se mide su actividad cerebral con EEG.

3.2.1 Estímulos

Los estímulos utilizados fueron distintos vídeos acompañados por estímulos adicionales: subtítulos, efectos sonoros, emoticonos, o estimulación vibro táctil.

Los vídeos fueron creados específicamente para las experimentaciones, componiendo secuencias de imágenes, extraídas de películas o documentales, no asociadas con ningún

diálogo o acción dramática, dado que las imágenes deben mantenerse lo más neutrales posible para permitir la medición de los efectos producidos por los estímulos adicionales a las imágenes (Gerdes et al., 2013) (Wang, Y. et al., 2020).

Para la elaboración de los subtítulos accesibles se contó con la participación de personal especializado del CESyA.

Para la estimulación vibro táctil se utilizó un guante, creado por el Grupo de Displays y Aplicaciones Fotónicas (GDA) de la Universidad Carlos III de Madrid, que permitía aplicar una suave vibración táctil en las yemas de los dedos y la palma de la mano de los participantes (ver Ilustración 63). El guante se implementó con un Inesis Golf Glove 100, de tela elástica y con un tamaño que permitía asegurar el ajuste del guante a las yemas de los dedos y la palma de la mano. Sobre este guante se colocaron motores de monedas Uxcell 1030 para proporcionar la estimulación cutánea adecuada en cada una de las tres ubicaciones específicas: dedo índice, dedo anular y palma. Estas ubicaciones fueron seleccionadas en base a la localización de los receptores de Pacini y Meissner descritos en la sección 2.3.1. Estos motores funcionan con 3 voltios CC y 70 miliamperios y fueron seleccionados debido a su pequeño tamaño. Los motores se pegan al exterior del guante y transmiten las vibraciones a la piel debido al ajuste de la tela elástica. Antes de cada prueba, se verificaba el ajuste del guante, si los participantes se sentían cómodos con él y podían sentir claramente las vibraciones generadas (Lucía et al., 2020).

Ilustración 63. Guante háptico utilizado para la estimulación vibro táctil

La señal de conducción de los motores consistía en una ráfaga de 102 ms de señal cuadrada de 1 kHz, generada por un Arduino UNO activado por un PC de control y sincronizado con la proyección de los vídeos. Se utilizó una frecuencia de 1 kHz para producir una activación rápida de la vibración del motor, ya que esa era su frecuencia de resonancia. El ritmo al que

se disparaban las activaciones era el principal estímulo. Se seleccionó el ritmo como parámetro vibro táctil ya que es el más fácil de reproducir y tiene una mejor respuesta del usuario cuando se traduce en vibración táctil (Jack et al., 2015).

3.2.2 Participantes

En ambos experimentos se reclutaron dos grupos de participantes: un grupo de control con participantes voluntarios sin discapacidad auditiva y un grupo experimental de voluntarios con discapacidad auditiva. Se gestionaron sus datos conforme a la Ley de protección de datos.

Para la experimentación se estableció un protocolo ético que salvaguarda los derechos de los participantes en la misma. Todos los participantes fueron informados del objetivo y el procedimiento general de los estudios. Los participantes firmaron un documento de consentimiento informado, aprobado por el Comité de Bioética de la Universidad Carlos III de Madrid, y completaron una encuesta sobre información demográfica, nivel de estudios y grado de hipoacusia.

El grado de pérdida auditiva, declarado por las participantes con discapacidad auditiva, fue clasificado de acuerdo con la Clasificación audiométrica de las deficiencias auditivas de la Oficina Internacional de Audiofonología (BIAP, 1996) :

- Pérdida auditiva leve (entre 20 y 40 dB)
- Pérdida auditiva moderada (entre 41 y 70 dB, se percibe el habla si la voz es fuerte, y el sujeto comprende mejor lo que se dice si puede ver a su interlocutor)
- Pérdida auditiva severa (entre 71 y 90 dB, se percibe el habla si la voz es fuerte y cerca del oído, y también se perciben ruidos fuertes)
- Pérdida auditiva muy severa (entre 91 y 119 dB, no se percibe el habla y solo se perciben ruidos muy fuertes)
- Pérdida total de audición (más de 120 dB).

3.2.3 Registro y análisis de la actividad cerebral

Para los registros de EEG, se utilizó un equipo de EEG multicanal de 64 canales, con un casco Neuroscan y el sistema de registro ATI Vertex (Advantek SRL).

El casco EEG se ajustaba en la cabeza de los participantes previamente a la realización de las pruebas. Los electrodos de referencia se colocaban en las mastoides y el de tierra en la frente de los participantes. Además de los 64 canales de registro, se utilizaron canales adicionales para monitorizar los movimientos oculares.

Ilustración 64. Equipamiento de registro EEG

Ilustración 65. Colocación del casco EEG

Se disponía además de dos PC de sobremesa, uno de control y otro de registro EEG. El PC de control se destina a la proyección del material audiovisual, y envío de señales de sincronización al EEG para localizar temporalmente el momento en que se inician los estímulos en los potenciales registrados. También se utiliza para lanzar las señales de control al guante para que inicie o detenga la correspondiente estimulación vibro táctil, y al mismo tiempo enviar las correspondientes marcas temporales al EEG.

El segundo equipo se destina al registro de las señales EEG. Los datos se registran con un filtro paso de banda de 0.05-30 Hz y una frecuencia de muestreo de 512Hz/1000 Hz respectivamente en cada experimentación. En estos registros se evalúa la actividad global, sin distinguir las bandas de frecuencia involucradas (delta, theta, alfa, beta, gamma). En el registro resultante, junto a las ondas de actividad cerebral registrada en los distintos electrodos en el tiempo, aparecen las distintas marcas temporales enviadas desde el PC de control para determinar los tiempos de emisión de los estímulos.

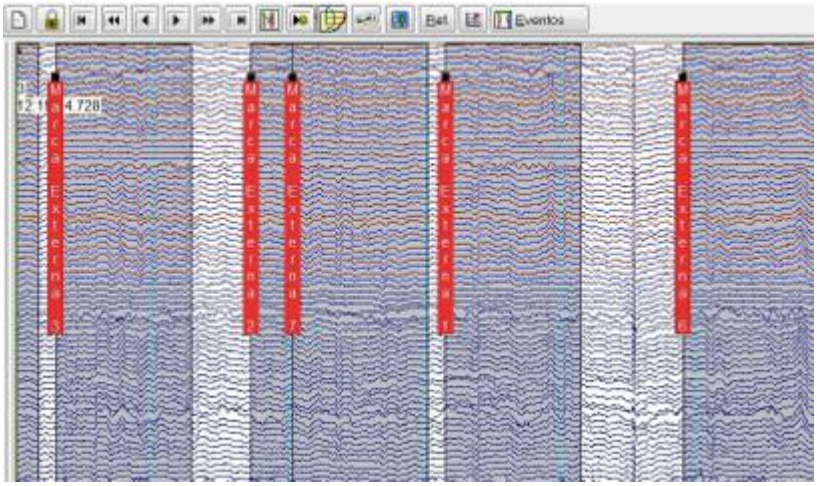


Ilustración 66. Ejemplo de registro con marcas temporales (en rojo) que señalan el tiempo en el que se produce el estímulo adicional (subtítulo, audio, etc.)

El análisis de los registros requiere varias fases de procesado. Primero se inspeccionan visualmente los registros de cada participante en cada condición para verificar que el registro se ha realizado limpiamente y para localizar los denominados artefactos, señales que provienen de fuentes ajenas a la actividad cerebral que se desea medir (en general movimientos de los ojos y parpadeos) y que se solapan con la señal principal. Estos artefactos se identifican mediante inspección visual y el segmento temporal en el que ocurren (un segundo en general) se descarta. En la Ilustración 67 se pueden observar este tipo de artefactos. La herramienta de registro ATI Vertex utilizada presenta distintas funcionalidades para facilitar la eliminación de estos segmentos.

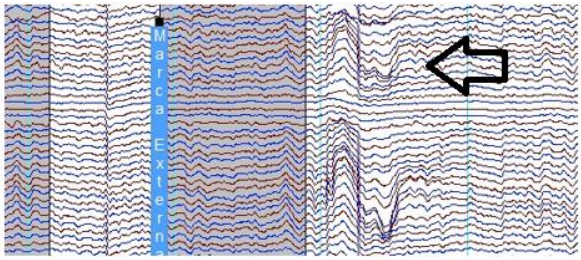


Ilustración 67. Artefacto en un registro

A partir de los registros verificados y limpios de artefactos, se calculan los promedios de activación cerebral para cada participante y condición. Los promedios pueden calcularse sobre el registro completo o sobre determinados segmentos seleccionados a partir de las marcas temporales. Por ejemplo, se puede realizar el promedio de activación en todos los segmentos de una duración configurable previos y/o posteriores a un determinado tipo de marca temporal.

A partir de estos promedios de activación a nivel de cuero cabelludo, se utilizan distintos algoritmos para generar los mapas de activación cerebral correspondientes. El algoritmo

LORETA (Low resolution Electromagnetic Tomography), ya mencionado en la sección 2.2.2.3, permite identificar las fuentes de corriente neuronal subyacentes a los potenciales registrados a nivel de cuero cabelludo. Utilizando estas fuentes, se generan los correspondientes mapas cerebrales, de acuerdo con el modelo de atlas cerebral promedio del Instituto Neurológico de Montreal (MNI) (Evans et al., 1993). Estos mapas muestran las áreas de máxima activación. En el caso del algoritmo LORETA los valores numéricos mostrados en los mapas cerebrales corresponden a unidades de corriente (microamperios/metro). Con otros algoritmos como el BMA (Bayesian Model Averaging), lo que se obtiene son valores estadísticos obtenidos con la prueba estadística T2 de Hotelling (Trujillo-Barreto, Aubert-Vázquez, & Valdés-Sosa, 2004). La prueba T2 es un análisis estadístico multivariado generalmente aplicado en experimentos neurocientíficos, en este caso para detectar las áreas cerebrales con activación estadísticamente significativa en comparación con el estado promedio de los valores registrados en todos los electrodos. Los visualizadores genéricos no especifican las unidades, sino los valores numéricos, para mostrar las áreas de máxima activación (ver Ilustración 68).

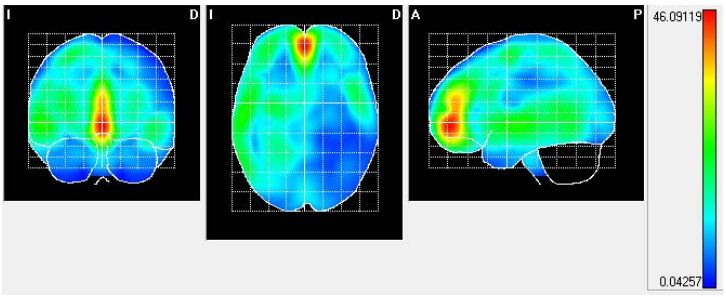


Ilustración 68. Ejemplo de visualizador genérico de mapas cerebrales

En la primera experimentación los mapas de activación cerebral fueron proporcionados por el departamento de Psiquiatría de la Universidad Complutense de Madrid. En la segunda experimentación los mapas se generaron con el software Neuronic Localizador de Fuentes (Neuronic SA). Este software permite seleccionar el algoritmo LORETA y generar los correspondientes mapas cerebrales.

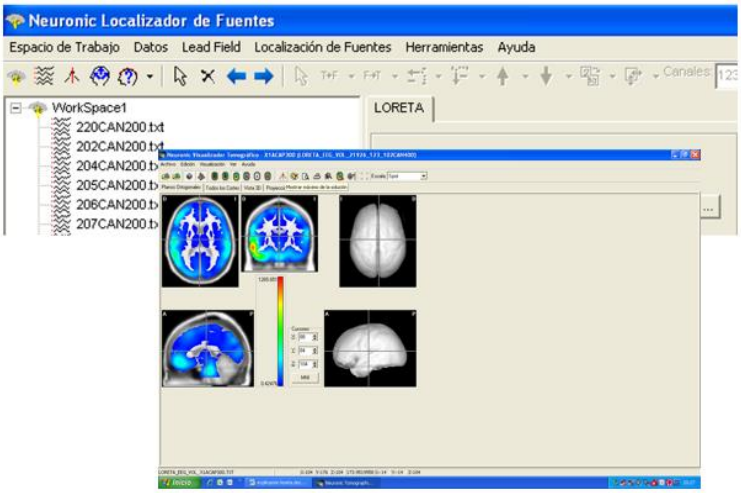


Ilustración 69. Mapas cerebrales generados con LORETA

Además, para cada mapa se puede obtener el detalle de estas áreas ordenadas de mayor a menor activación con la siguiente información (ver Tabla 6):

- AAL: etiqueta correspondiente a la zona activada (por ejemplo: Temporal_Sup_R, correspondiente a lóbulo temporal superior derecho). AAL son las iniciales de Automated Anatomical Labeling.
- Coordenadas X, Y, Z: coordenadas correspondientes a la zona activada de acuerdo con el atlas promedio MNI.
- Activación: valor de activación

AAL	X	Y	Z	Activation
Temporal_Sup_R	62	-18	-4	30.2320
Temporal_Mid_R	62	-18	-8	30.0220

Tabla 6. Detalle de las áreas de máxima activación

Estos mapas se obtienen promediados por condición y permiten comparar las activaciones cerebrales en las distintas condiciones.

3.3 Experimentación 1

3.3.1 Hipótesis de partida

La primera experimentación tiene como objetivo la validación de la Hipótesis 1 propuesta en esta investigación: *El subtitulado accesible textual no transmite la información que aporta la música de forma inmediata a través de la emoción*.

3.3.2 Protocolo experimental

El estudio se realizó en las instalaciones de la Facultad de Medicina, en colaboración con el departamento de Psiquiatría.

3.3.2.1 Participantes

Se reclutaron dos grupos de participantes. El grupo de control estaba compuesto por 16 participantes, con audición normal auto referida, ocho mujeres y ocho hombres, con edades entre 20 y 60 años (media: 39,83, DE: 12,24), 24,1% con bachillerato superior, 27,6% con grado universitario y 48,3% con estudios de postgrado.

El grupo experimental estaba compuesto por siete mujeres y seis hombres, de edad entre 20 y 60 (media: 39,4, DE: 12,21), 38,5% con bachillerato superior, 38,5% con grado universitario y 23,1% con estudios de postgrado. El grado de hipoacusia fue auto referido por los participantes. Cuatro participantes tenían pérdida auditiva moderada y usaban audífonos, cuatro participantes tenían pérdida severa, usaban audífonos y tres de ellos tenían implante coclear, y finalmente un participante tenía pérdida auditiva total e implante coclear.

3.3.2.2 Estímulos

Se preparó un vídeo con secuencias neutras extraídas del documental Samsara[18]. Este documental presenta imágenes de distintos países sobre naturaleza y sociedad, con fondo musical, pero sin diálogos ni textos. Se seleccionaron 40 fragmentos sin cambios de plano de 10 segundos, se añadió un fundido de entrada y de salida de 2 segundos para suavizar las transiciones, y se eliminó la banda sonora original. En cada fragmento se añadió primero un efecto sonoro extraído de una base de datos de audio; los fragmentos y el instante en el que aparecieron fueron asignados al azar (entre los segundos 2 y 8). Para cada efecto sonoro se preparó el correspondiente subtítulo de acuerdo con la norma de subtitulado accesible. El vídeo final se compuso con 80 fragmentos ordenados de forma aleatoria: 40 fragmentos con efecto sonoro sin subtítulo y los mismos 40 fragmentos sin audio, pero con el correspondiente subtítulo. El vídeo se presentaba en 5 tiempos, con 20 segundos de descanso entre cada tiempo.

Ilustración 70. Imagen del documental Samsara

[18] https://www.filmaffinity.com/es/film269079.html

3.3.2.3 Procedimiento

Los participantes fueron citados en sesiones individuales. Se solicitaba que leyeran y firmaran el documento de consentimiento informado, y que rellenaran la encuesta con preguntas sobre edad, nivel educativo, y en su caso grado de pérdida auditiva, uso de audífonos, o implante coclear.

A continuación, se les invitaba a acomodarse en un sillón frente a una pantalla de 17 pulgadas situada a 1,5 metros en línea directa de visión. El sonido llegaba a los participantes a través de altavoces estéreo conectados a la pantalla. Se les pedía que retiraran sus audífonos o la parte extraíble del implante coclear, y se les colocaba el casco EEG. Finalmente, bajo su mano izquierda se colocaba un botón de pulsación conectado al PC de control. Se les explicaba que iban a ver un vídeo y que debían pulsar el botón de control cada vez que sintieran cualquier tipo de emoción durante la proyección. Se apagaban las luces de la habitación, y se lanzaba la proyección del vídeo en el PC de control.

Durante la proyección se registraba la actividad cerebral desde el PC de registro. En el PC de control, se registraba el número total de pulsaciones del botón de control, y el momento en el que ocurrían, que se transmitía al registro EEG donde quedaba grabado como una marca temporal.

3.3.2.4 Condiciones

Se consideraron cuatro condiciones experimentales, en función de los fragmentos del vídeo visualizados por cada grupo. Estas condiciones se resumen en la siguiente tabla (Tabla 7).

Grupo	Condición	Descripción
CONTROL (audición normal)	AUDIO	40 fragmentos con audio
	SUBTÍTULO	40 fragmentos con subtítulo
EXPERIMENTAL (pérdida auditiva muy severa o total)	MUTE	40 fragmentos sin audio ni subtítulos
	SUBTÍTULO	40 fragmentos con subtítulo

Tabla 7. Condiciones experimentales (experimentación 1)

3.3.3 Resultados

3.3.3.1 Registros de pulsaciones

Se utilizó la prueba U de Mann-Whitney para comparar el número de pulsaciones del botón en las diferentes condiciones y grupos. Se seleccionó esta prueba en base a que la prueba de Shapiro-Wilk rechazó la normalidad de la muestra en algunas condiciones, y en otras condiciones el tamaño de las muestras no era suficientemente grande (menos de 20) para asumir una distribución normal. Los resultados se muestran en la Tabla 8 .

Grupo control	Grupo experimental		
AUDIO	**MUTE**		
19,87 ± 9,63	7 ± 4.1		
SUBTÍTULO	**SUBTÍTULO**		
9,81 ± 9,69	7.15 ± 4,86	**P-value****	0.79486
P-value*	**P-value***		
0.00528	0.52218		
< 0.05	> 0.05		

* Valores p de Mann-Whitney comparando condiciones del mismo grupo

** Valores p del estadístico comparando la condición Subtítulo entre grupos

Tabla 8. Comparativa del número de pulsaciones entre condiciones

En el caso del número de pulsaciones en las condiciones AUDIO (19.87 ± 9.63) y SUBTÍTULO (9.81 ± 9.69) en el grupo de control, los resultados muestran una diferencia significativa ($p = 0.00528$) entre ambas condiciones, lo que sugiere que las imágenes con audio producen más reacciones emocionales que las imágenes solas.

Una segunda prueba de Mann-Whitney se utilizó para comprar las condiciones MUTE (7 ± 4.1) y SUBTÍTULO (7.15 ± 4.86) en el grupo experimental, con el resultado de una diferencia no significativa, lo que sugiere que los subtítulos no producen reacciones emocionales adicionales a los estímulos visuales.

Finalmente se utilizó la prueba de Mann-Whitney para comparar el número de pulsaciones en la condición SUBTÍTULO en el grupo de control (9.81 ± 9.69) con la condición SUBTÍTULO en el grupo experimental (7.15 ± 4.86), resultando en una diferencia no significativa ($p = 0.79486$), lo que sugiere que la reacción emocional producida por los subtítulos es similar en ambos grupos.

3.3.3.2 Registros de actividad cerebral

Se analizaron los registros de la activación cerebral justo antes de las marcas temporales correspondientes a las pulsaciones del botón. Se observó la presencia de dos ERPs negativos (ver

Ilustración 71, el primero alrededor de 300ms (etiquetado como NS300), y el segundo alrededor de 100ms (etiquetado como NS100), inmediatamente antes de la respuesta motora de pulsación del botón indicativa de una reacción emocional. Estos ERP aparecían de forma similar en ambos grupos y en todas las condiciones. Los ERPs negativos inmediatamente anteriores a la respuesta motora se considera reflejan el procesamiento emocional y cognitivo de los estímulos y procesos relacionados con la toma de decisión (Shibasaki, Barrett, Halliday, & Halliday, 1980) (Bianchin & Angrilli, 2011)(Duncan et al., 2009). El componente NS100 se relaciona con los procesos cognitivos necesarios para desencadenar la acción motora (Bianchin & Angrilli, 2011), por lo que el análisis se centró en el componente NS300.

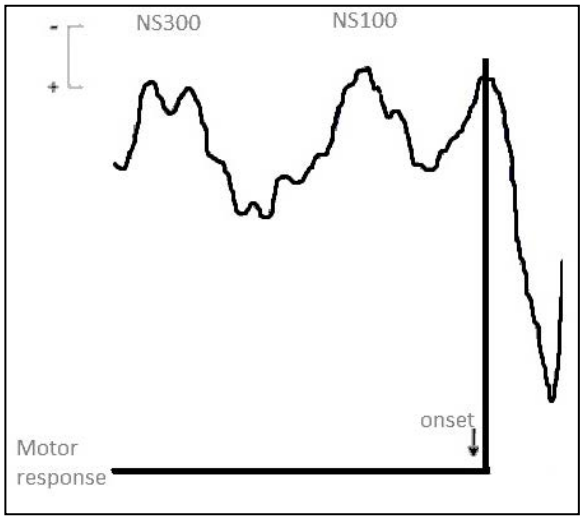
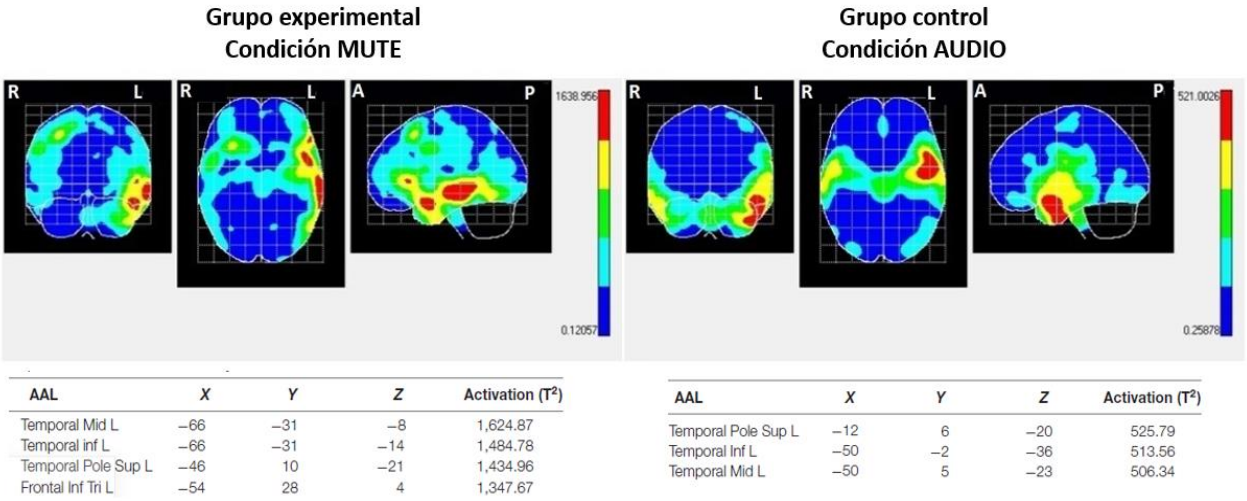


Ilustración 71. Promedio de formas de onda cerebrales antes del inicio de la respuesta motora de pulsación del botón.

A partir de estos registros se generaron los mapas de activación promedio correspondientes al ERP NS300 por cada grupo y condición, obteniendo los siguientes resultados.

Los mapas de activación cerebral correspondientes a NS300 mostraron una alta activación en el lóbulo temporal izquierdo tanto en el grupo de control en la condición AUDIO como en el grupo experimental en la condición MUTE, siendo la magnitud de la activación mayor en el grupo experimental (>1,400) comparada con el grupo de control (~500). Adicionalmente, aparece gran activación en el lóbulo frontal en el grupo experimental, que no aparece en el grupo de control (Ilustración 72).

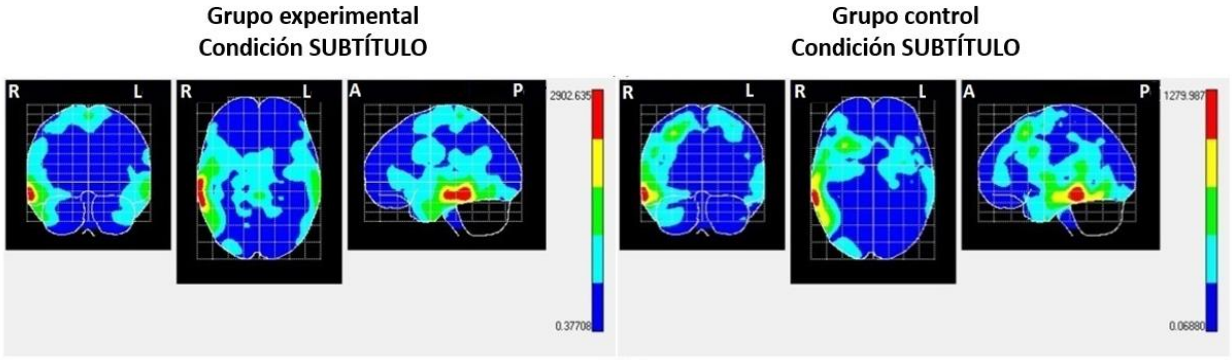


AAL	X	Y	Z	Activation (T^2)
Temporal Mid L	−66	−31	−8	1,624.87
Temporal inf L	−66	−31	−14	1,484.78
Temporal Pole Sup L	−46	10	−21	1,434.96
Frontal Inf Tri L	−54	28	4	1,347.67

AAL	X	Y	Z	Activation (T^2)
Temporal Pole Sup L	−12	6	−20	525.79
Temporal Inf L	−50	−2	−36	513.56
Temporal Mid L	−50	5	−23	506.34

Ilustración 72. Mapas de activación cerebral promedio alrededor de NS300 en las condiciones MUTE y AUDIO. Las áreas de mayor intensidad se muestran en rojo/amarillo.

En la condición SUBTÍTULO, la activación se centra en el lóbulo temporal derecho en ambos grupos, pero también con mayor activación en el grupo experimental (>2,500) comparado con el grupo de control (~1,200) (ver Ilustración 73).



AAL	X	Y	Z	Activation (T^2)
Temporal Inf R	66	−38	−12	2,902.56
Temporal Mid R	66	−22	−13	2,579.89

AAL	X	Y	Z	Activation (T^2)
Temporal Mid R	66	−35	−13	1,265.89

Ilustración 73. Mapas de activación cerebral promedio alrededor de NS300 en la condición SUBTÍTULO. Las áreas de mayor intensidad se muestran en rojo/amarillo.

3.3.4 Discusión

Comparando las condiciones AUDIO y MUTE en ambos grupos se observa primero que la condición AUDIO produce un número significativamente mayor de respuestas emocionales que la condición MUTE, y segundo que ambos grupos activan las mismas áreas temporales. Las áreas temporales inferior y media se han asociado con el procesamiento visual y auditivo, el reconocimiento de objetos y caras, y el reconocimiento de palabras (Pehrs et al., 2017). El polo temporal forma parte del córtex asociativo y está involucrado en la integración sensorial multimodal y contextual, especialmente en el procesamiento social y emocional multimodal (Olson et al., 2007) (Skipper et al., 2011).

La diferencia en las condiciones AUDIO y MUTE, en los mapas NS.300, es la mayor activación de estas áreas en el grupo con discapacidad auditiva. Numerosos estudios consideran que una mayor amplitud en los potenciales registrados se asocia con un mayor esfuerzo de procesamiento cognitivo (Moreno, Casado, & Martín-Loeches, 2016) (Sanchez-Lopez, Silva-Pereyra, & Fernandez, 2016). Adicionalmente, únicamente el grupo con discapacidad auditiva mostró activación en el lóbulo frontal en la condición MUTE. Esta actividad frontal puede asociarse con un mayor consumo de recursos atencionales y cognitivos (Böcker, Baas, Kenemans, & Verbaten, 2001)(Martino, Kumaran, Seymour, & Dolan, 2006).

En cuanto a la condición SUBTÍTULO, comparando ambos grupos, se encontró que ambos grupos activaban las áreas temporales inferior y media (la actividad se centra en procesamiento visual y reconocimiento de palabras), pero una vez más, el grupo con discapacidad auditiva activó estas áreas con mayor intensidad.

3.3.5 Conclusiones

Los resultados indican, por una parte, teniendo en cuenta los datos relativos a las pulsaciones de botón, que los estímulos auditivos producen mayor número de respuestas emocionales que los subtítulos.

Por otra parte, los resultados de los registros EEG muestran que en ambos grupos se activan las mismas áreas temporales de procesamiento de la información. Pero el grupo de personas sordas moviliza estas áreas con mucha más intensidad al ver los vídeos sin sonido ni subtítulos, es decir con mayor esfuerzo de procesamiento, y además moviliza áreas frontales cerebrales relacionadas con la atención y procesos cognitivos de orden superior. Las personas oyentes movilizan estos recursos temporales con menor nivel de intensidad.

La presencia de subtítulos aumenta la activación de las áreas visuales y de procesamiento verbal en ambos grupos, pero con mayor intensidad en el grupo de personas sordas.

Los resultados muestran pues que los subtítulos no producen las mismas reacciones emocionales que los estímulos auditivos, y que cuando un sujeto con pérdida auditiva está viendo un video sin subtítulos, necesita un mayor consumo de recursos cerebrales como consecuencia de la pérdida auditiva. Si además agregamos subtítulos al video, este esfuerzo aumenta y se enfoca en el procesamiento visual y verbal.

Por tanto, los subtítulos no sólo no son capaces de transmitir la emoción subyacente en los sonidos, si no que requieren un mayor esfuerzo cognitivo, muy diferente a la respuesta emocional inmediata que genera la música desde los circuitos internos del cerebro interno.

Para producir esas reacciones emocionales primarias, inmediatas, hay que explorar pues otras formas no verbales de transmitir la información emocional contenida en la música.

3.4 Experimentación 2

3.4.1 Hipótesis de partida

Esta segunda experimentación tiene como objetivo la validación de las Hipótesis 2 y 3 propuestas en esta investigación:

2. *El tacto puede ser un canal de transmisión alternativo de emociones musicales básicas.*
3. *Los parámetros musicales pueden transmitirse de forma similar, aunque más limitada, mediante estimulación vibro-táctil.*

La propuesta concreta de esta experimentación es comprobar que una estimulación vibro táctil rítmica, aplicada en la mano a personas con pérdida auditiva profunda y severa durante el visionado de un material audiovisual, puede activar una respuesta emocional en regiones del cerebro similar a la producida por la banda sonora musical en participantes con audición normal.

3.4.2 Protocolo experimental

El estudio se realizó en las instalaciones de la Facultad de Medicina, en colaboración con el departamento de Psiquiatría durante los meses de mayo-septiembre 2019.

3.4.2.1 Participantes

Para el grupo de control, se reclutaron estudiantes voluntarios sin discapacidad auditiva de la facultad de medicina donde se realizó el estudio. Para el grupo experimental se solicitaron voluntarios con discapacidad auditiva severa a través de asociaciones de personas sordas.

En el grupo experimental se presentaron únicamente participantes mujeres, por lo que en el grupo de control sólo se consideraron mujeres también. Esta circunstancia permitió eliminar la variable género del estudio. El género se ha asociado con importantes diferencias cerebrales en la forma en que se procesan los estímulos emocionales. Por ejemplo, los estímulos desagradables y muy excitantes evocan mayores amplitudes de ERP en las mujeres que en los hombres (Lithari et al., 2010), o los estímulos considerados estéticamente bellos por los participantes producen una actividad cerebral bilateral en las mujeres, mientras que en los hombres sólo en el hemisferio derecho (Cela-Conde et al., 2009).

El grupo de control se estableció con 9 participantes del sexo femenino, con audición normal auto referida, de edad entre 18 y 22 años (media: 19,22, DE: 1,31), todas ellas con título de bachillerato y estudiantes de medicina en la Universidad Complutense de Madrid.

El grupo experimental, se estableció con 7 participantes femeninas con pérdida auditiva severa, de edad entre 37 y 61 años (media: 49,28, DE: 10,76), una con título de bachillerato, 4 con grado universitario y 2 con estudios de postgrado. El grado de hipoacusia fue auto referido por las participantes. Cuatro participantes presentaban una pérdida auditiva muy severa y usaban audífonos (una de ellas tenía un implante coclear en un oído) y tres participantes tenían una pérdida auditiva total (una de ellas tenía un implante coclear). Todas las participantes eran diestras.

La diferencia de la media de edad en ambos grupos se debía a las circunstancias de la experimentación. En la colaboración con la facultad de Medicina y para compartir el equipamiento EEG se debía aprovechar el mismo grupo de control conformado por estudiantes. Esta diferencia de edad es una limitación de esta investigación.

3.4.2.2 Estímulos

Se crearon dos videos sin sonido, de un minuto de duración, que mostraban respectivamente imágenes aéreas de campos de cereales al atardecer (Ilustración 74), e imágenes de organismos y plantas moviéndose en las profundidades del mar (ver Ilustración 75). Las secuencias fueron extraídas de las películas The Straight Story (David Lynch, 1999) y The Tree of Life (Terrence Malick, 2011).

Ilustración 74. Vídeo 1: Vistas aéreas de campos de cereales al atardecer

Ilustración 75. Vídeo 2: Organismos y plantas en el fondo del mar

Estos vídeos se asociaban con estímulos musicales o con estímulos vibro táctiles. Para los estímulos musicales, se seleccionaron dos fragmentos musicales: un extracto del Concierto para violín n. ° 1, op. 6, Rondo, de Nicolo Paganini, con tempo rápido (106 bpm) y escala mayor, y un extracto del Adagio for Strings, Op.11 de Samuel Barber, con tempo lento (60 bpm) y escala menor. Estos estímulos se seleccionaron por haber sido ya utilizados anteriormente en estudios sobre emoción (Krumhansl, 1997) (Koelsch et al., 2013).

Para los estímulos vibro táctiles se utilizó el guante, creado por el Grupo de Displays y Aplicaciones Fotónicas (GDA) de la Universidad Carlos III de Madrid. El ritmo al que se dispararon las activaciones rítmicas correspondía al tempo de las cada una de las dos músicas seleccionadas.

Se creó un tercer vídeo (ver Ilustración 76), también de un minuto de duración y sin sonido, y con escenas neutras mostrando imágenes de un pasillo de colegio (las escenas se obtuvieron de la película Elephant[19]). A este vídeo se añadieron estímulos visuales no verbales consistentes en una combinación de emoticonos, vúmetros y notas musicales. En esta elección se consideró la familiaridad de estos estímulos. Los emoticonos se presentaban en formato alegre o triste.

Ilustración 76. Vídeo 3 asociado a estímulos visuales no verbales

[19] https://www.filmaffinity.com/es/film553137.html

Se utilizaron estímulos musicales tanto positivos como negativos (tristes o alegres), y sus correlatos vibro táctiles, y visuales, para calcular el promedio de todas las reacciones, eliminando el efecto de valencia.

3.4.2.3 Procedimiento

Los participantes se citaban en sesiones individuales. Se solicitaba que leyeran y firmaran el documento de consentimiento informado, y que rellenaran la encuesta con preguntas sobre edad, nivel educativo, y en su caso grado de pérdida auditiva, uso de audífonos, o implante coclear.

A continuación, se les invitaba a entrar en una pequeña habitación, insonorizada y protegida como cámara de Faraday para evitar interferencias en los registros EEG, y a acomodarse en un sillón frente a una pantalla de 25 pulgadas, empotrada en la pared, situada a 1,5 metros en línea directa de visión. El sonido llegaba a los participantes a través de altavoces estéreo conectados a la pantalla. Se les pedía que retiraran sus audífonos o la parte extraíble del implante coclear, y se fijaba el casco EEG a la cabeza y en el caso del grupo experimental el guante en la mano derecha (todas las participantes eran diestras).

Fuera de esta habitación, justo detrás de la pantalla empotrada se disponían los dos PC de sobremesa, uno de control y otro de registro EEG.

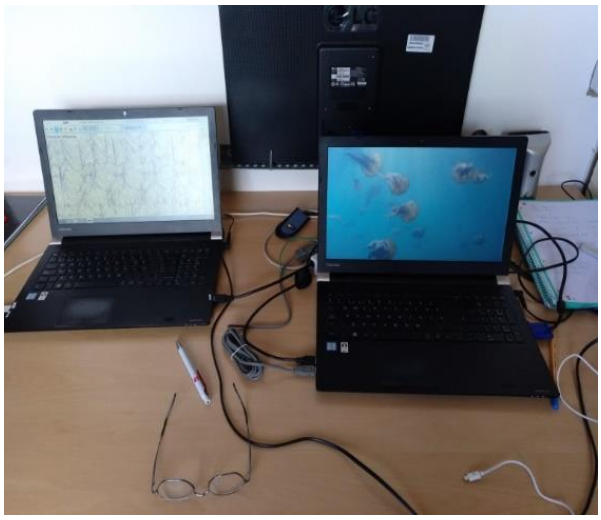


Ilustración 77. Equipos de control y registro

Se explicó a las participantes que iban a ver un video y que se sentaran cómodamente y se relajaran durante su proyección. Se apagaban entonces las luces de la habitación y se lanzaba el video correspondiente.

3.4.2.4 Condiciones

Los videos se organizaron en 4 condiciones diferentes, que se resumen en la siguiente tabla (Tabla 9).

Grupo	Condición	Descripción
CONTROL (audición normal)	AUDIO	Los vídeos 1 y 2 se proyectaban en orden aleatorio, cada vídeo asociado a uno de los dos fragmentos musicales seleccionado aleatoriamente como banda sonora.
EXPERIMENTAL (pérdida auditiva muy severa o total)	MUTE	Los vídeos 1 y 2 se proyectaban sin sonido en orden aleatorio.
	TÁCTIL	A continuación, se proyectaba uno de los vídeos 1 y 2, seleccionado aleatoriamente, sin sonido, mientras simultáneamente el guante háptico producía uno de los 2 estímulos rítmicos vibro táctiles seleccionado de forma aleatoria.
	EMOTICONO	Se proyectaba el vídeo 3 asociado aleatoriamente a un emoticono alegre o triste.

Tabla 9. Condiciones experimentales (experimentación 2)

El grupo de control únicamente vio los vídeos en la condición AUDIO ya que, por las circunstancias de compartición de la experimentación, el mismo grupo de control debía someterse a más condiciones de otras experimentaciones.

3.4.3 Resultados

Se obtuvieron los mapas de la actividad cerebral promedio registrada durante el visionado de cada vídeo (alrededor de 1 minuto). Se generaron los mapas promedio para las distintas condiciones experimentales siguiendo el procedimiento descrito en la sección 3.2.3, con los siguientes resultados.

En la condición AUDIO (grupo de control, video + música), como se muestra en la Ilustración 78 y en la Tabla 10, se obtienen activaciones cerebrales máximas en el lóbulo temporal (áreas superior, media e inferior, área de Heschl, polo temporal), en la ínsula, en el opérculo rolándico y en la circunvolución frontal inferior. Todas estas activaciones se localizan en el hemisferio derecho.

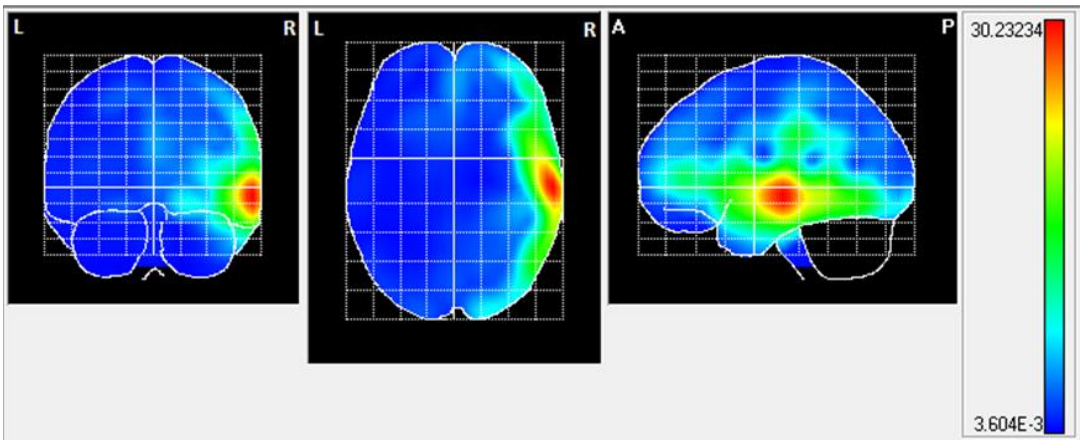


Ilustración 78. Mapas de activación cerebral media en la condición AUDIO. Las áreas de máxima intensidad se muestran en rojo/amarillo.

AAL	X	Y	Z	Activation
Temporal_Sup_R	62	-18	-4	30.2320
Temporal_Mid_R	62	-18	-8	30.0220
Heschl_R	54	-10	4	20.8320
Temporal_Pole_Sup_R	54	2	0	20.7600
Temporal_Inf_R	62	-22	-20	20.6870
Insula_R	50	2	-4	20.2430
Rolandic_Oper_R	54	6	0	19.4980
Frontal_Inf_Oper_R	50	10	0	18.3190
Temporal_Pole_Mid_R	54	2	-16	17.8680
Frontal_Inf_Tri_R	50	22	0	14.5790

Tabla 10. Áreas de máxima activación en la condición AUDIO (coordenadas MNI)

En la condición MUTE (grupo experimental con discapacidad auditiva, video sin audio), tal y como se muestra en la Ilustración 79 y en la Tabla 11 , los picos máximos de activación se localizan en las áreas frontales. También se encontró una activación menos significativa en el polo temporal. La mayoría de las activaciones se encuentran en el hemisferio izquierdo, aunque las áreas frontal superior y medial derecha también muestran activaciones significativas.

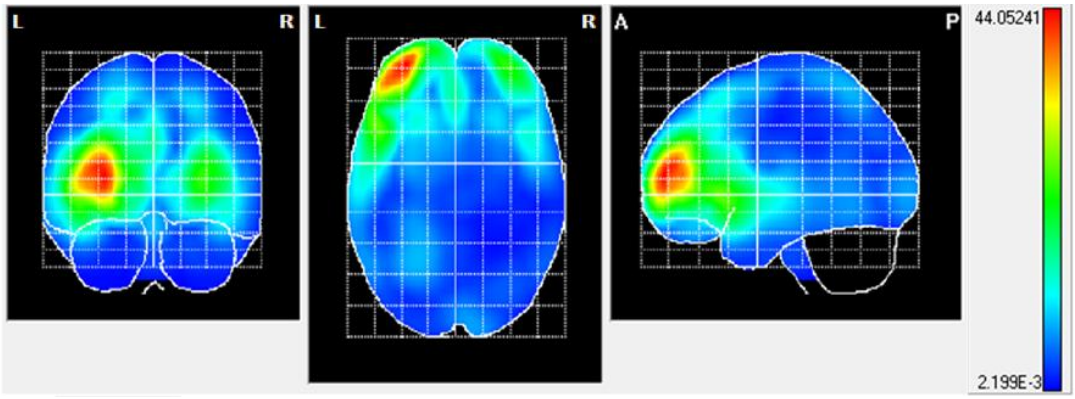


Ilustración 79. Mapas de activación cerebral media en la condición MUTE. Las áreas de máxima intensidad se muestran en rojo/amarillo.

AAL	X	Y	Z	Activation
Frontal_Mid_L	-30	58	12	44.0520
Frontal_Sup_L	-26	58	12	39.5670
Frontal_Inf_Tri_L	-42	46	12	39.2800
Frontal_Sup_Orb_L	-30	58	-4	32.0470
Frontal_Mid_Orb_L	-34	58	-4	31.7150
Frontal_Inf_Orb_L	-46	18	-4	25.3910
Frontal_Sup_R	30	58	8	24.5120
Frontal_Mid_R	30	58	4	23.7370
Frontal_Inf_Oper_L	-50	14	0	23.2600
Temporal_Pole_Sup_L	-50	14	-4	22.5480

Tabla 11. Áreas de máxima activación en la condición MUTE (coordenadas MNI)

En la condición TÁCTIL (grupo experimental con discapacidad auditiva, video sin audio + estimulación vibro táctil), como se muestra en la Ilustración 80 y en la Tabla 12, se obtienen activaciones muy significativas en el lóbulo temporal (áreas superior, medial y polo temporal),

el área frontal inferior, en la ínsula y en el opérculo rolándico. También se encuentran picos menos significativos en el lóbulo frontal medio y las áreas de Heschl. Todas estas activaciones están en el hemisferio izquierdo.

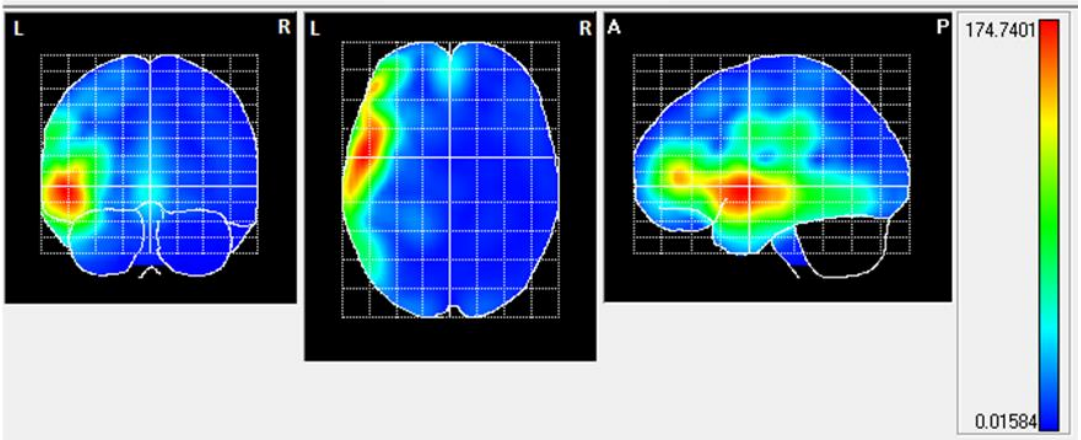


Ilustración 80. Mapas de activación cerebral media en la condición TÁCTIL. Las áreas de máxima intensidad se muestran en rojo/amarillo.

AAL	X	Y	Z	Activation
Temporal_Sup_L	-50	6	-4	174.7400
Temporal_Pole_Sup_L	-54	6	0	170.9190
Frontal_Inf_Oper_L	-50	10	0	167.2420
Rolandic_Oper_L	-54	6	4	161.4850
Insula_L	-46	10	-8	161.3440
Temporal_Mid_L	-58	-6	-8	157.1770
Frontal_Inf_Orb_L	-46	18	-4	156.2220
Frontal_Inf_Tri_L	-50	18	0	153.5690
Frontal_Mid_L	-46	46	0	132.2130
Heschl_L	-58	-10	8	104.0720

Tabla 12. Áreas de máxima activación en la condición TÁCTIL (coordenadas MNI)

En la condición EMOTICONO (grupo experimental con discapacidad auditiva, video sin audio + estímulos visuales no verbales), la máxima activación se localiza en las regiones frontales, relacionadas con la atención y procesos cognitivos (Ilustración 81).

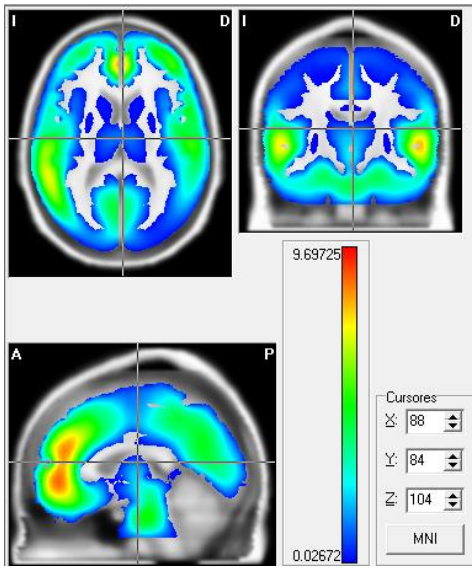


Ilustración 81. Mapas de activación cerebral media en la condición EMOTICONO. Las áreas de máxima intensidad se muestran en rojo/amarillo.

3.4.4 Discusión

Esto resultados muestran que el grupo de control en la condición AUDIO presenta activaciones cerebrales consistentes con la literatura existente (Frühholz et al., 2016), específicamente en las áreas mencionadas en la sección 2.2.5.1 involucradas en el procesamiento de la música afectiva: polo temporal superior, área de Heschl (corteza auditiva), ínsula y circunvolución frontal inferior, y que pueden registrarse con EEG. Además de estas áreas, se encontró activación en la circunvolución temporal medial y el opérculo rolándico.

Los picos principales corresponden a la circunvolución temporal superior derecha, que ha sido identificada como una región de integración multisensorial (Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007) (Menon & Levitin, 2005), y a la circunvolución temporal medial, que está asociada con el procesamiento visual y auditivo (Pehrs et al., 2017). También se encontró una mayor activación de la circunvolución de Heschl, que es parte del lóbulo temporal superior y contiene la corteza auditiva primaria, y en el polo temporal, parte de la corteza de asociación, también involucrada en la integración sensorial multimodal (Olson et al., 2007) (Skipper et al., 2011). La ínsula, que también muestra una activación significativa, generalmente se considera un mediador entre los sistemas cerebrales sensoriales y afectivos en la percepción de sonidos afectivos (Mirz, Gjedde, Sdkilde-Jrgensen, & Pedersen, 2000), traduciendo la

percepción de señales afectivas de los sonidos en emociones subjetivas (Kotz et al., 2012). Los estudios de imágenes funcionales han demostrado la activación de la ínsula durante las tareas de integración audiovisual y la escucha pasiva de música (Wildgruber et al., 2004) (Lang & Bradley, 2010). La activación del opérculo rolándico se ha relacionado con el procesamiento de la información sintáctica de la música agradable (Koelsch et al., 2013).

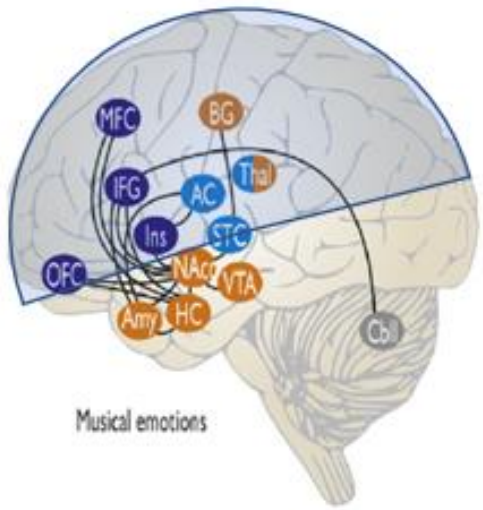


Ilustración 82. Áreas de procesamiento música afectiva (en sombreado las áreas que pueden registrarse con EEG)

En la condición MUTE (participantes con discapacidad auditiva, videos sin audio), la activación se localiza principalmente en el lóbulo frontal, asociada con mayores recursos de atención voluntaria e integración emocional y cognitiva (Bushara et al., 2003) (Böcker et al., 2001). Se activa significativamente la corteza frontal medial, que participa en varias funciones sociales y emocionales relacionadas con la comunicación y la comprensión interpersonales (Amodio & Frith, 2006). Este resultado es coherente con los resultados de la Experimentación 1 que muestran cómo los participantes con discapacidad auditiva movilizan estas áreas con alta intensidad cuando ven un video sin audio. Lamentablemente no se dispone de resultados comparativos con el grupo de control para la condición MUTE, ya que esta condición no pudo establecerse en el grupo de control debido a que ya estaba sometido a otras condiciones de otros experimentos.

En la condición TÁCTIL (grupo experimental con discapacidad auditiva, video sin sonido + estimulación vibro táctil), nuevamente se encuentra una activación máxima en las mismas áreas de la condición AUDIO, que coinciden con las áreas involucradas en el procesamiento de la música afectiva: lóbulo temporal superior, circunvolución frontal inferior, opérculo rolándico, e ínsula. La activación del área de Heschl (corteza auditiva), aunque en menor intensidad, es un resultado interesante en los participantes con discapacidad auditiva. Otra área importante de activación es el lóbulo frontal medial, que se asocia con el procesamiento visual y auditivo (Pehrs et al., 2017). Se observa además una intensidad mucho mayor en la condición TÁCTIL.

En la comparación entre ambas condiciones (ver

Ilustración 83), sorprende la similitud de los mapas, pero con inversión de la lateralidad, probablemente porque los participantes en la condición TÁCTIL recibieron la estimulación vibro táctil a través de su mano derecha y, por tanto, el hemisferio izquierdo sería el encargado de procesarla. Sin embargo, no se ha encontrado explicación al hecho de que la activación en la condición AUDIO se concentre en el hemisferio derecho.

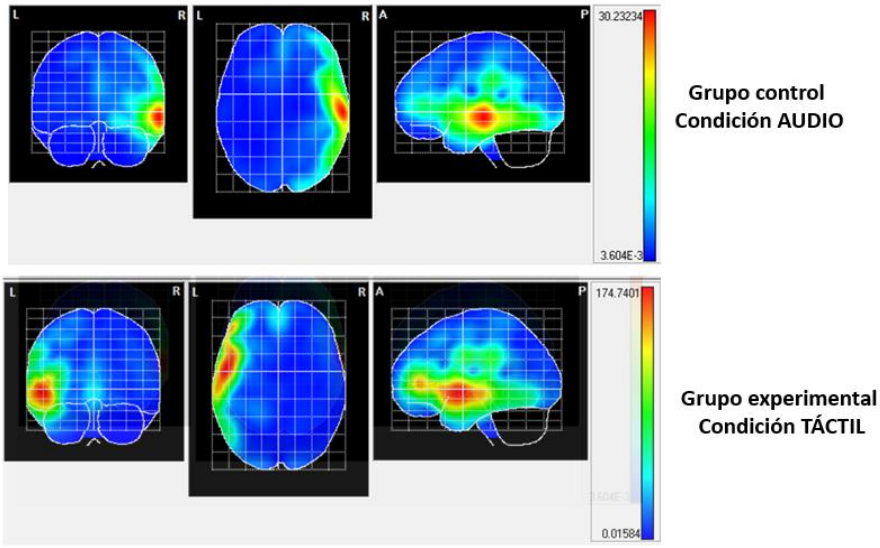


Ilustración 83. Similitudes e inversión de lateralidad entre las condiciones AUDIO/grupo de control y TÁCTIL/grupo experimental.

La activación de la corteza auditiva concuerda con otros estudios que muestran que las vibraciones táctiles activan la corteza auditiva, o mejoran la percepción auditiva musical y del habla (Schürmann, Caetano, Hlushchuk, Jousmäki, & Hari, 2006) (Leder, Spitzer, Milner, Flevaris-Phillips, & Richardson, 1986) (Luo & Hayes, 2019).

En la condición EMOTICONO, las áreas de máxima activación son las áreas frontales, por lo que la inclusión de estímulos visuales más gráficos, aunque no requieren procesamiento verbal, siguen requiriendo de recursos cerebrales cognitivos, alejados del procesamiento emocional.

Hay que señalar por último que en la condición TÁCTIL las participantes con sordera total manifestaron sensaciones muy positivas con las vibraciones del guante, mientras que el resto de las participantes manifestaron en general que el guante no les desagradaba pero que les producía una sensación leve de confusión.

3.4.5 Conclusiones

El resultado más interesante de este estudio es que las redes neuronales emocionales, que pueden ser registradas mediante EEG, involucradas en el procesamiento de la música afectiva (Frühholz et al., 2016) también se activan significativamente en la condición vibro táctil cuando se combinan con los medios visuales.

Aunque este estudio presenta limitaciones, como son el tamaño de la muestra, participantes de un solo sexo, medias de edad distintas en ambos grupos, la limitación de la estimulación vibro táctil a un patrón rítmico, o la limitación de los registros EEG para conocer las activaciones más internas del cerebro, los resultados obtenidos muestran que la integración de estímulos vibro táctiles muy simples con imágenes neutras puede potenciar la activación de áreas similares a las que se activan con la música, acercando así las reacciones cerebrales a las de sujetos oyentes expuestos a una experiencia audiovisual completa.

4 MODELOS CNN PARA EXTRACCIÓN DE LA EMOCIÓN MUSICAL

4.1 Introducción

El objetivo de esta parte de la investigación es determinar un modelo sencillo de clasificación automática para extraer las emociones de la música de películas como base para un futuro framework de subtitulado. La experimentación se basa en la cuarta y última hipótesis de trabajo: *Los modelos de aprendizaje CNN permiten clasificar emociones básicas (alegría, tristeza, miedo) en fragmentos breves de música de película*.

Se pretende establecer una primera aproximación al problema con la premisa de simplicidad, tomando como punto de partida unas condiciones básicas, acordes con las consideraciones neurocientíficas respecto a la emoción (sección 2.3.5):

- Clasificar en base a las emociones básicas de alegría, tristeza y miedo, expresadas en grado intenso
- Utilizar fragmentos musicales del orden de 2 segundos
- Utilizar las bases de datos de películas etiquetadas con rigor científico

Y acordes a las consideraciones en cuanto a los modelos MER (sección 2.4.8):

- No se sabe cuáles son las características de audio significativas para la captura de la emoción musical, ni si son correctos los correspondientes algoritmos desarrollados, pero precisamente los modelos CNN permiten trabajar sin una selección previa de características.
- Los modelos CNN de clasificación de género musical más simples obtienen los mejores resultados con fragmentos musicales de pocos segundos.

Por tanto, la experimentación se basa en desarrollar modelos CNN sencillos para clasificar muestras audio de 2 segundos (duración suficiente para transmitir las emociones que se desea identificar), etiquetadas con rigor científico, en base a las emociones básicas de alegría, tristeza y miedo, expresadas en grado intenso.

4.2 Metodología y materiales

4.2.1 Entorno de desarrollo

Todos los programas se desarrollaron en el lenguaje de programación Python (versión 3.7.6), y con el entorno de desarrollo de Spyder (Scientific Python Development Environment)[20], software abierto bajo licencia MIT[21], utilizado por la familiaridad con el mismo, y por ser un entorno destinado a la programación científica en el lenguaje Python.

[20] https://www.spyder-ide.org/

[21] Licencia libre originada en Massachusetts Institute of Technology

Todas las experimentaciones se realizaron siempre con el mismo equipo (Procesador Intel Core i5 2.50 GHz y 16GB RAM).

Para el tratamiento audio se utilizó la librería Librosa[22] (versión 0.8.0), la librería de referencia en Python para MIR en Python, y que dispone de amplias funcionalidades para obtener distintos tipos de espectrogramas, siendo los espectrogramas las posibles entradas audio "visuales" para las redes CNN.

Para el desarrollo de los modelos CNN se utilizó la librería Keras[23] (versión 2.4.3). Keras es una biblioteca de código abierto (con licencia MIT) escrita en Python, y tiene como objetivo ofrecer unas API sencillas e intuitivas para el desarrollo de modelos complejos de aprendizaje profundo.

El entorno de desarrollo utilizado se resume en la siguiente figura.

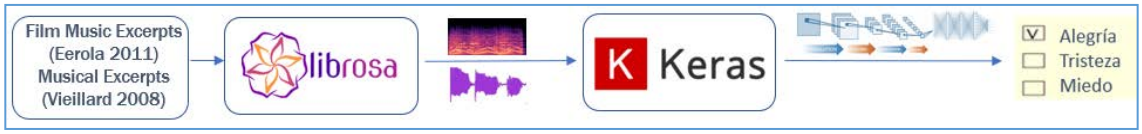


Ilustración 84. Framework para clasificación CNN

4.2.2 Selección de fragmentos musicales

Como datos de entrenamiento se utilizaron las bases de datos, Film Music Excerpts (Eerola & Vuoskoski, 2011) y Musical Excerpts (Vieillard et al., 2008) descritas en la sección 2.2.3, por ser las únicas disponibles etiquetadas en cuanto a emoción de forma rigurosa desde la neurociencia.

Musical Excerpts (Vieillard et al., 2008) se compone de 40 fragmentos, compuestos, específicamente en el género de la música de cine. Los fragmentos están cualificados en base a cuatro emociones: alegría, tristeza, miedo y tranquilidad, 10 para cada tipo de emoción. Las tasas de reconocimientos de cada emoción son respectivamente 99%, 84%, 72% y 94%. Para la experimentación se utilizaron los 40 fragmentos. La duración media es de 12,5 segundos. (Los derechos de autor de estos fragmentos son Copyright Bernard Bouchard 1998, y se permite su uso.)

Film Music excerpts (Eerola & Vuoskoski, 2011) se compone de un primer set de 360 extractos musicales de 60 bandas sonoras de películas seleccionadas por un panel de expertos musicólogos. Las músicas ofrecen ejemplos de las emociones de alegría, tristeza, miedo, ira y tranquilidad en intensidad alta y en intensidad moderada (puntuadas de 1 a 7). Además, un segundo set se compone de los 110 ejemplos más representativos del primer set, evaluados en una segunda ronda por otro grupo de participantes estudiantes (puntuadas de 1 a 9). En total se seleccionaron 57 fragmentos del primer set, y 37 fragmentos del segundo set (no incluidos en el primer set), con las más altas puntuaciones en las emociones de alegría,

[22] https://librosa.org/

[23] https://keras.io/

tristeza, miedo y tranquilidad (puntuaciones >=6). Se descartaron los fragmentos representativos de ira por ser ambiguos ya que también tienen una puntuación alta en la emoción de miedo. La duración media de estos fragmentos es de 16 segundos.

En total, se seleccionaron 134 fragmentos, y un total de 2.074,43 segundos de audio distribuidos de acuerdo con la siguiente tabla (Tabla 13).

Emociones	Fragmentos	Duración Total(segundos)
Miedo	40	589,68
Alegría	31	471,34
Tristeza	34	558,99
Tranquilidad	29	454,42
Total	**134**	**2074,43**

Tabla 13. Distribución de emociones por fragmentos

Como ya se ha comentado (sección 2.2.3), la tranquilidad no se considera una emoción básica, pero es una emoción que se incluye en los estudios sobre música y emoción, en general etiquetada como paz/ternura/tranquilidad, ya que es una emoción que se percibe en la música con frecuencia, lo que no ocurre con otras emociones básicas como el asco (Eerola & Vuoskoski, 2011).

El formato de las muestras era MP3, con tasa de muestreo original de 44.100Hz.

4.2.3 Conjunto de datos de entrenamiento

Los fragmentos musicales seleccionados permitían generar 976 muestras de 2 segundos de duración. Aunque se trata de una cantidad limitada, y además con muestras pertenecientes a mismos fragmentos musicales, con riesgo de sobreajuste, se decidió utilizar únicamente estas muestras científicamente contrastadas y correspondientes a emociones expresadas con intensidad.

Los fragmentos se redujeron a una tasa de muestreo de 16.000 Hz para facilitar el procesamiento (como se indica más adelante, en la sección 4.3.2, se comprobó que los resultados no se veían afectados significativamente por el cambio en la tasa de muestreo), y se dividieron en muestras de 2 segundos.

Para cada muestra de 2 segundos, y utilizando la librería Librosa, se generaron tres tipos de espectrogramas: STFT (espectrogramas de frecuencias), Mel (espectrogramas de frecuencias convertidas a la escala Mel), y CQT (espectrogramas con las frecuencias convertidas a tonos musicales). Se consideraron ventanas superpuestas de 512 muestras (longitud correspondiente a unos 31 milisegundos), con un salto entre ventanas de 256. Librosa ofrece además la funcionalidad de separar el audio en componentes percusivos y en componentes armónicos (en base a que un determinado punto responda mejor a filtros verticales y horizontales respectivamente), por lo que adicionalmente se generaron espectrogramas sobre el componente armónico y el componente percusivo.

Es de destacar que el espectrograma STFT correspondería al análisis de Fourier realizado en el oído a nivel de la membrana basilar, el espectrograma Mel a la percepción no lineal de las frecuencias, y el espectrograma CQT a la percepción de la relación entre frecuencias.

La Ilustración 85 muestra los espectrogramas de frecuencias STFT correspondientes a una muestra de 2 segundos.

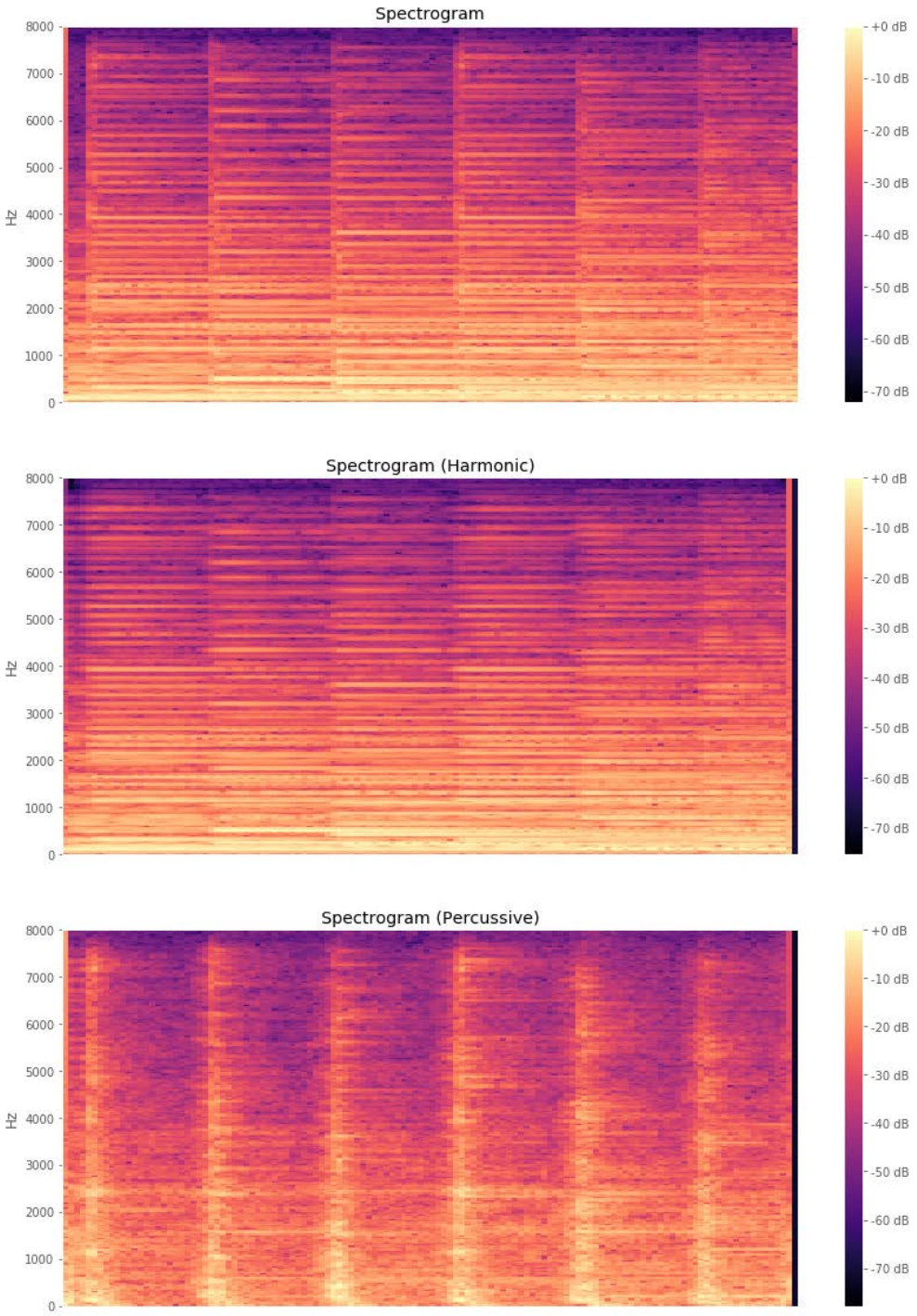


Ilustración 85. Espectrogramas STFT sobre una muestra de 2 seg. a 16KHz.

La Ilustración 86 muestra los espectrogramas Mel (espectrograma de frecuencias donde las frecuencias se convierten a la escala Mel) correspondientes a la misma muestra.

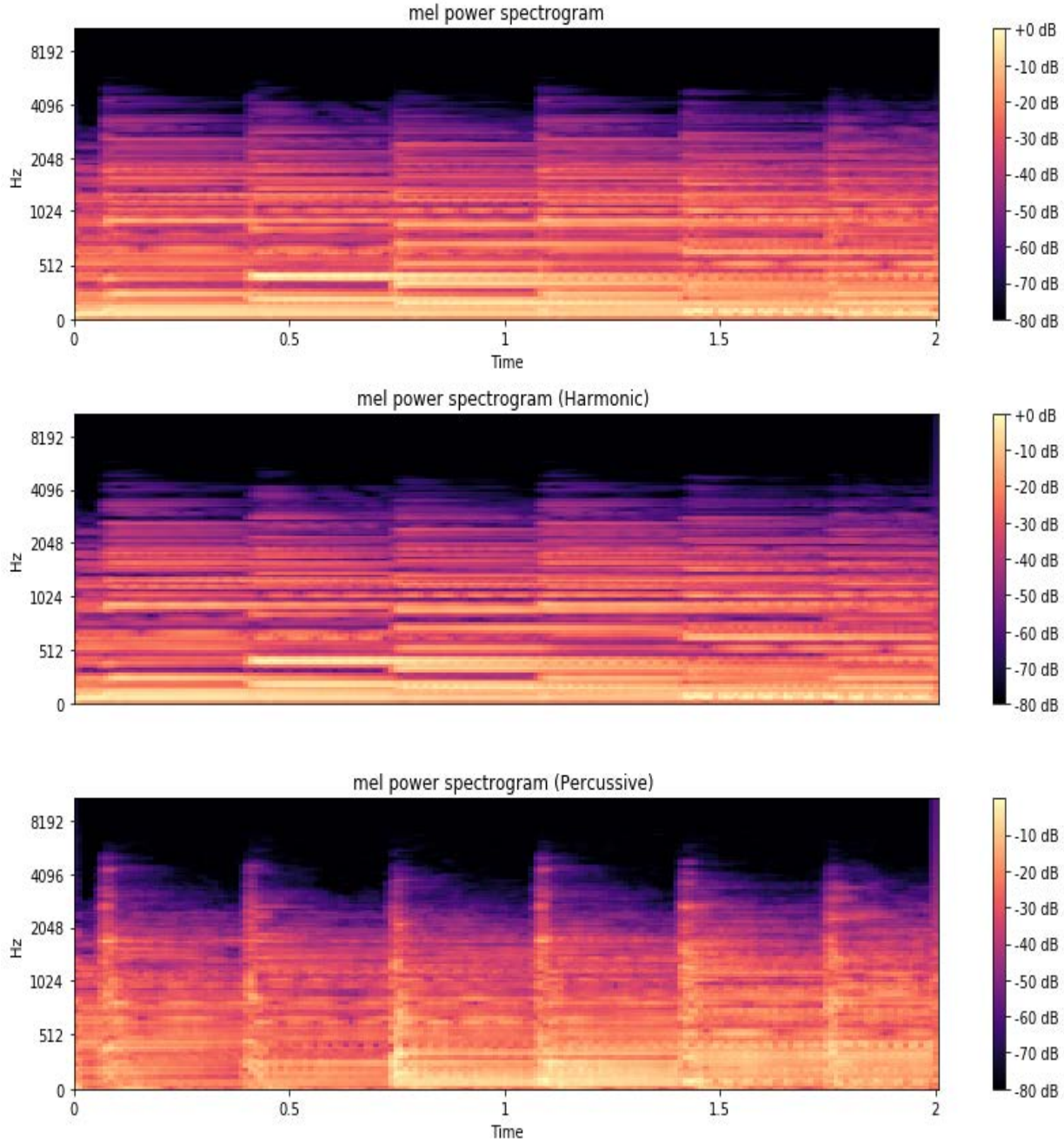


Ilustración 86. Espectrogramas Mel sobre la misma muestra de 2 seg. a 16KHz.

Finalmente, la Ilustración 87 muestra los espectrogramas CQT correspondientes a la misma muestra. En estos espectrogramas las frecuencias se representan en escala logarítmica, correspondiendo a las diferentes notas (C1, C2, C3, C4…), y se muestran los niveles de energía estimados en los distintos tipos de tonos musicales.

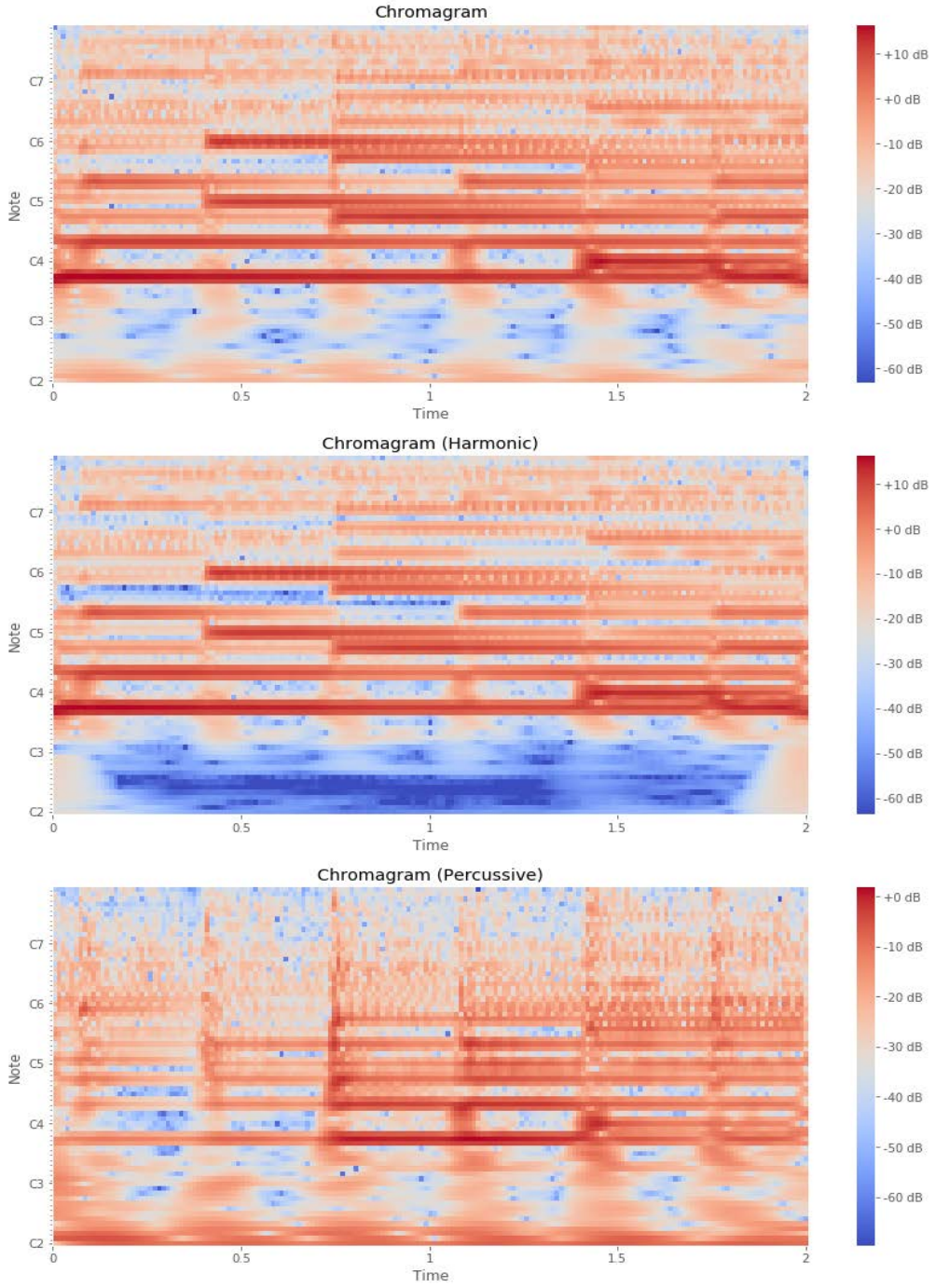


Ilustración 87. Espectrogramas CQT sobre la misma muestra de 2 seg. a 16KHz.

Este conjunto de datos fue el utilizado en todas las experimentaciones. En la Tabla 14 se detalla el tamaño de los datos de entrada correspondientes a cada tipo de espectrograma, considerando 2 segundos de audio a 16.000Hz, y ventanas superpuestas de 512 muestras y salto de 256.

Tipo Espectrograma	Tamaño datos entrada	Total Muestras de 2 segundos
STFT	257x126	976
Mel	128x126	976
CQT	82x126	976

Tabla 14. Tamaño datos de entrada en función del tipo de espectrograma

4.3 Experimentación 1

El modelo CNN utilizado en la primera experimentación fue adaptado a partir de una de las primeras publicaciones (Zhang et al., 2016), descrita en la sección 2.4.7, en la que se utilizaba una red convolucional con pocas capas para la clasificación de género musical.

En esta primera experimentación se buscaba el desarrollo de un modelo CNN base que alcanzara tasas de reconocimiento en línea con el estado del arte, para determinar el tipo de espectrograma más idóneo como dato de entrada y los parámetros de entrenamiento más adecuados, para la clasificación en base a emociones.

4.3.1 Procedimiento

El modelo CNN utilizado de partida fue adaptado a partir del modelo de (Zhang et al., 2016) para obtener una clasificación en base a 4 etiquetas correspondientes a las emociones de alegría, tristeza, miedo y tranquilidad. Además, realizaron distintas pruebas para afinar el modelo, añadiendo y ajustando capas.

La Ilustración 88 resume el modelo resultante. En la capa convolucional se consideran 5 convoluciones con filtros 3x3, y activación ReLU. Se realiza BatchNormalization tras cada convolución, para evitar sobreajustes, y Max Pooling 2x2 cada dos convoluciones. El resultado, redimensionado a un vector unidimensional, se procesa en la capa completamente conectada por 2 redes Dense de 300 y 150 neuronas, con activación ReLU, y una última capa de salida con activación Softmax. En estas capas densas se realiza Dropout para evitar sobreajustes.

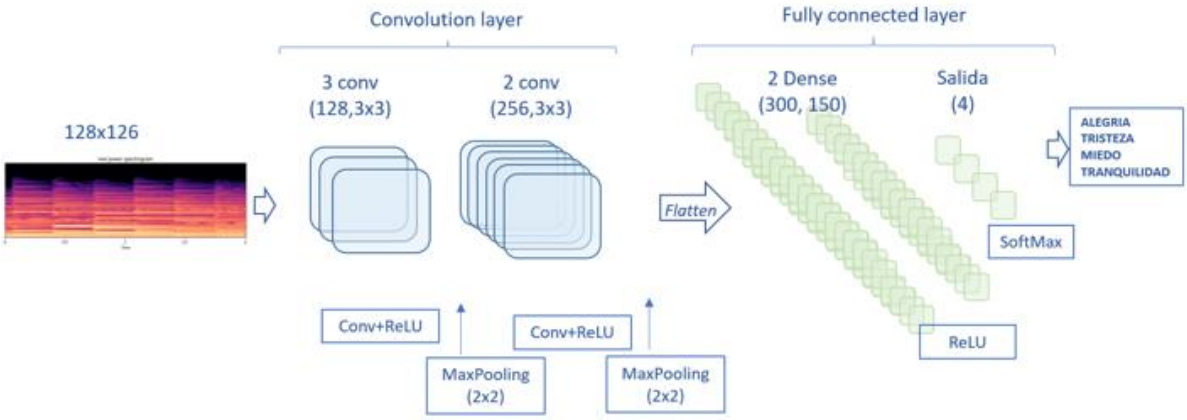


Ilustración 88. Modelo CNN utilizado en experimentación 1

Se utilizaron los espectrogramas STFT, Mel, y CQT y se realizaron distintas pruebas para afinar el modelo. La Tabla 15 resume los parámetros con los que se obtuvieron mejores resultados, y los parámetros seleccionados para las siguientes experimentaciones.

Parámetro	Valores estudiados	Resultados	Selección
Samplig rate	16KHz 32KHz 44.1 KHz	Resultados similares pero mejores tiempos de procesamiento con 16KHz	16KHz
Learnig Rate	1e-2, 1e-3, 1e-4	Mejores resultados con 1e-3	1e-3
Dropout	En capas Conv y Dense	Mejores resultados aplicando Dropout únicamente en las capas densas	Dropout sólo en capas densas.
Optimizador	Stochastic gradient descent (SGD) Adaptive Moment Estimation (Adam)	Ligeramente mejores resultados con SGD.	SGD
Epochs	50,100,200	Mejores resultados con 50 y 100. Se consideran 50 epochs para optimizar el tiempo de procesamiento.	50

Tabla 15. Resumen de parámetros evaluados

Una vez establecido este modelo base se procedió a evaluar los distintos tipos de espectrogramas. Para la evaluación del modelo sobre cada conjunto de espectrogramas se utilizó validación cruzada, considerando varias repeticiones sobre diferentes particiones del conjunto de datos de entrada, es decir las 976 muestras de 2 segundos.

4.3.2 Resultados

La Tabla 16 resume los resultados obtenidos con los distintos tipos de espectrogramas. El valor considerado es el valor Accuracy, es decir el total de muestras correctamente clasificadas de acuerdo con la fórmula: {TP + TN} / {TP + TN + FP + FN}, donde:

- TP = Positivo Verdadero
- TN= Negativo Verdadero
- FP = Falso Positivo
- FN= Falso Negativo

El entrenamiento del modelo se realizó sobre 50 epochs.

	Tipo de Espectrograma								
	CQT	Harmonic	Percussive	Mel	Harmonic	Percussive	STFT	Harmonic	Percussive
	0,74	0,76	0,64	0,74	0,74	0,73	**0,80**	0,79	0,70
	0,77	0,77	0,69	0,76	0,77	0,61	0,77	0,66	0,71
	0,77	0,74	0,72	**0,82**	0,73	0,68	0,75	0,71	0,65
	0,80	0,77	0,63	0,73	0,70	0,63	0,74	0,70	0,73
	0,81	0,77	0,76	0,78	0,77	0,60	0,73	0,73	0,66
Media	0,78	0,76	0,69	0,76	0,74	0,65	0,76	0,72	0,69
STD	0,02	0,01	0,05	0,03	0,03	0,05	0,03	0,04	0,03

Tabla 16. Resultados Accuracy para distintos tipos de espectrogramas.

Los resultados obtenidos están en línea con el estado del arte. Además, se observa que los mejores resultados se obtienen con el tipo de espectrograma CQT, y que separando las componentes armónica y percusiva los resultados no mejoran.

Ligeramente inferiores son los resultados con el espectrograma tipo Mel, y tipo STFT, sin embargo, el tiempo de entrenamiento es mucho mayor con los espectrogramas Mel y sobre todo STFT como se muestra en la siguiente figura.

Datos para 50 epochs		
	Accuracy	*Tiempo procesamiento*
CQT	78%	~38 m
Mel	76%	~60mn
STFT	76%	~144 mn

Tabla 17. Resultados medios de Accuracy y tiempos de procesamiento por espectrograma

Se estudiaron también las métricas de Precisión, Recall Y F1 en la clasificación de las distintas emociones:

- Precisión es la ratio TP/(TP+FP). Representa la calidad la clasificación, la capacidad de no clasificar como positivas muestras negativas.

- Recall es la ratio TP/(TP+FN). Representa la exhaustividad de la clasificación, la capacidad de reconocer todos los ejemplos positivos.

- F1 representa la media armónica entre Precisión y Recall, es un valor entre 0 y 1, siendo 1 la puntuación óptima. El valor F1 asume que precisión y exhaustividad son igualmente importantes.

La Tabla 18 muestra un ejemplo de los resultados medios obtenidos en los espectrogramas CQT, y Mel en la clasificación de las distintas emociones. Se observa nuevamente que los resultados con el tipo de espectrograma CQT y Mel son similares, aunque son mejores en cuanto a la emoción tranquilidad en los espectrogramas CQT.

	Precisión	Recall	F1
CQT			
ALEGRÍA	**0,87**	0,82	0,84
MIEDO	0,85	0,88	0,86
TRISTEZA	0,75	0,74	0,74
TRANQUILIDAD	0,70	0,70	0,70
MEL			
ALEGRÍA	0,88	0,84	0,85
MIEDO	0,76	0,91	0,83
TRISTEZA	0,73	0,77	0,75
TRANQUILIDAD	0,73	0,45	0,54

Tabla 18. Resultados medios (Precisión, Recall, F1) por emoción

También se observa que se clasifican mejor las emociones de Alegría y Miedo que las emociones de tristeza y tranquilidad. En las puntuaciones dadas por los sujetos experimentales en el estudio de Film Music Excerpts de (Eerola & Vuoskoski, 2011), se observa que los fragmentos calificados como tristes con puntuaciones altas, también puntúan en tranquilidad con puntuaciones medias-altas. Es decir, en el procesamiento cerebral, las melodías tristes con ritmo lento, volumen suave, se pueden percibir también como tranquilas, ya que las emociones de tristeza y tranquilidad comparten valores medios en parámetros musicales como tempo, dinámica, articulación y timbre (ver Tabla 3). Estas similitudes pueden estar en el origen de los peores resultados en la clasificación automática de las emociones de tristeza y tranquilidad.

Comparando con los porcentajes de fragmentos identificados por los participantes del estudio de Musical Excerpts de (Vieillard et al., 2008), se observa nuevamente que el clasificador automático identifica peor tristeza y tranquilidad, aunque mejor la emoción de miedo.

	Vieillard
ALEGRÍA	0,99
MIEDO	0,72
TRISTEZA	0,84
TRANQUILIDAD	0,94

Tabla 19. Tasas de reconocimiento en Musical Excerpts (Vieillard et al., 2008)

Hay que considerar que estos resultados se basan en un número limitado de muestras (976 muestras), que además pertenecen a pocos fragmentos musicales (134). Sin embargo, se considera que son de utilidad para determinar el tipo de espectrograma más idóneo como dato de entrada para las CNN.

De acuerdo con estos resultados, y el mejor rendimiento en tiempo de procesamiento de los espectrogramas CQT, en las siguientes experimentaciones se utiliza este tipo de espectrogramas. Curiosamente el espectrograma CQT es el que representa las relaciones entre tonos, paralelamente a como el cerebro procesa las relaciones tonales musicales.

4.4 Experimentación 2

Esta segunda experimentación parte de la publicación de (Won et al., 2020), descrita en la sección 2.4.7, en la que se muestran los resultados de un benchmarking realizado con los modelos CNN más representativos del estado del arte para clasificación del género musical. El objetivo de esta segunda experimentación era adaptar estos modelos para la clasificación

en las emociones de alegría, tristeza, miedo y tranquilidad, y evaluar posibles mejoras respecto al modelo establecido en la experimentación anterior.

4.4.1 Procedimiento

Se seleccionaron los tres modelos que se utilizaron en el benchmarking con las muestras de audio de menor duración (en el benchmarking la duración mínima de los fragmentos era de 3 segundos), los denominadas Short-chunk CNN, Musicnn y Sample-level CNN (ver sección 2.4.7). Short-chunk CNN además fue el que mejor resultados obtuvo en general.

Los modelos fueron adaptados para utilizar como datos de entrada los espectrogramas CQT (en vez de los espectrogramas Mel utilizados en el benchmarking), y clasificar en base a las cuatro emociones de alegría, tristeza, miedo y tranquilidad. Además, se consideraron los parámetros ya ajustados en la experimentación anterior, de acuerdo con la Tabla 15.

Nos referiremos a estos modelos como Short-chunk CNN adaptado, Musicnn adaptado, y Sample-level CNN adaptado. A continuación, se describen las experimentaciones realizadas con cada uno de los modelos.

4.4.1.1 Modelo Short-chunk CNN adaptado

La Ilustración 89 resume este modelo. El modelo es muy similar al modelo desarrollado en la anterior experimentación, pero con más profundidad. En la capa convolucional se consideran 8 convoluciones con filtros 3x3, y activación ReLU. Se realiza BatchNormalization y Max Pooling 2x2 tras cada convolución. El resultado, redimensionado a un vector unidimensional, se procesa en la capa completamente conectada de 512 neuronas, con activación ReLU, y una última capa de salida con activación Softmax. En la capa Dense se realiza Dropout para evitar sobreajustes.

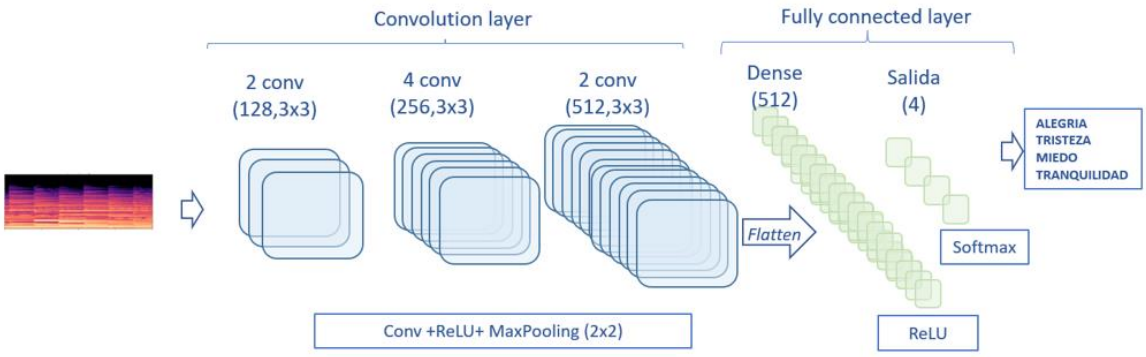


Ilustración 89. Modelo Short Chunk CNN adaptado

El conjunto de datos de entrenamiento para este modelo fueron los 976 espectrogramas CQT en base a los fragmentos audio de 2 segundos con muestreo 16KHz, y tamaño 82x126.

Para la evaluación del modelo, primero se utilizó validación cruzada, considerando distintas repeticiones del entrenamiento sobre diferentes particiones del conjunto de datos de entrada, es decir las 976 muestras de 2 segundos de audio. Al ser esta red más profunda, la evaluación se realizó con 50 epochs y se repitió con 100 epochs.

Posteriormente se realizó una división del conjunto de fragmentos musicales en un conjunto de entrenamiento (75%, 732 muestras) y otro de validación (25%, 244 muestras), antes de dividir los fragmentos en las muestras de 2 segundos, de forma que en los datos de validación no había ninguna muestra de fragmentos pertenecientes al conjunto de entrenamiento. Esta operación se repitió 4 veces para disponer de 4 entrenamientos con conjuntos de validación distintos.

Por último, se repitió todo el proceso considerando únicamente las 3 emociones básicas de alegría, tristeza y miedo, descartando esta vez la emoción de tranquilidad.

4.4.1.1.1 Modelo Short-chunk CNN con ResNet adaptado

Este modelo es análogo al anterior, pero cada capa de convolución se sustituye por un bloque ResNet (con el mismo número de filtros y tamaño de filtros) como el que se muestra en la Ilustración 90.

.

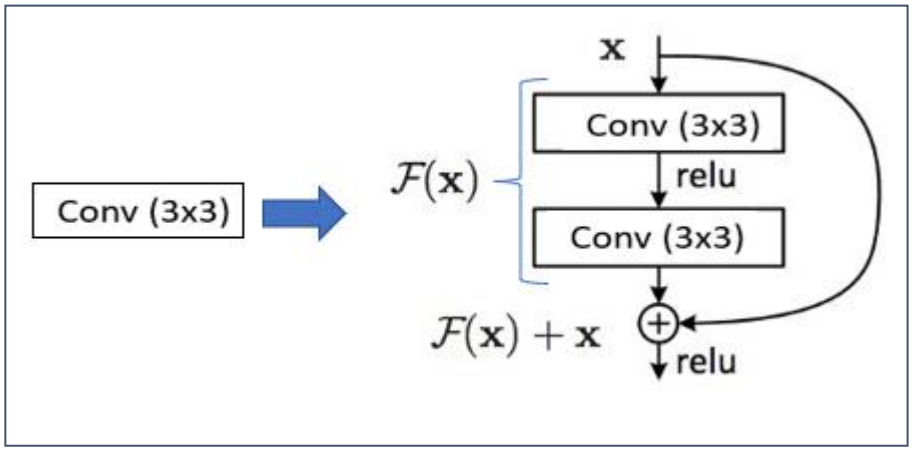


Ilustración 90. Bloque ResNet

Este modelo había obtenido peores resultados en el benchmarking de (Won et al., 2020), por lo que únicamente se realizó validación cruzada, considerando varias repeticiones sobre diferentes particiones del conjunto de datos de entrada, es decir las 976 muestras de 2 segundos de audio con 50 y 100 epochs. Se quería de esta manera verificar si en esta adaptación los resultados también eran peores antes de descartar el modelo.

4.4.1.2 Modelo Musiccnn adaptado

Este modelo utiliza una primera capa convolucional compuesta, como en los módulos Inception, por convoluciones paralelas con distintos filtros horizontales y verticales, concatenándose a continuación sus salidas. Los filtros verticales capturarían las características tímbricas, y los filtros horizontales la evolución temporal.

La Ilustración 91 resume este modelo. En la primera capa convolucional se consideran 5 convoluciones con diferentes filtros horizontales y verticales y activación ReLU. Se realiza BatchNormalization y Pooling tras cada convolución. Las distintas salidas se concatenan y se procesan posteriormente por 3 capas convolucionales adicionales con bloques ResNet. El resultado se procesa en la capa completamente conectada, con activación ReLU, y una última capa de salida con activación Softmax. En la capa Dense se realiza Dropout para evitar sobreajustes

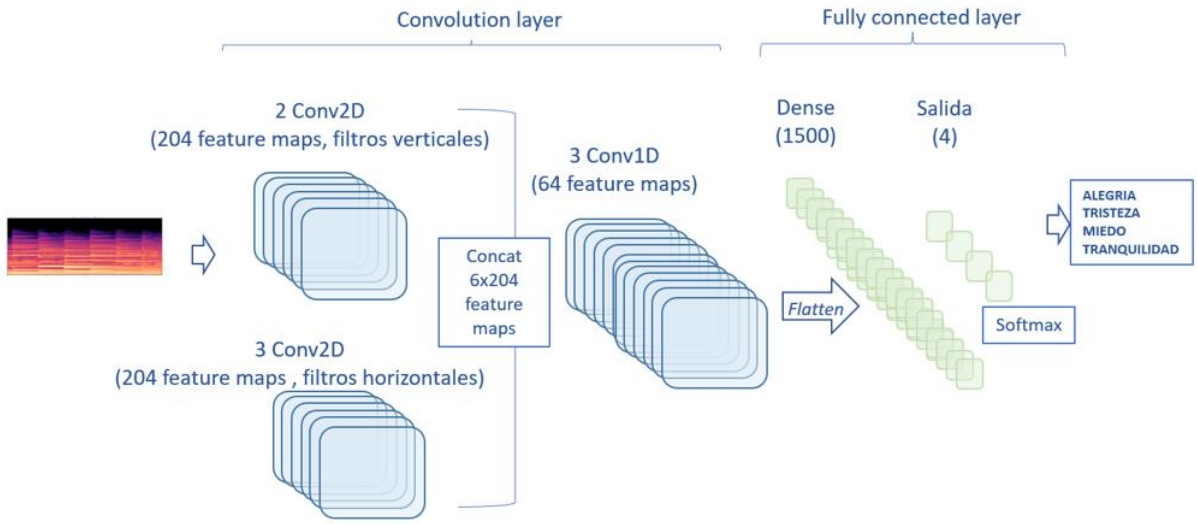


Ilustración 91. Modelo Musiccnn adaptado

Con este modelo se realizó validación cruzada, considerando varias repeticiones sobre diferentes particiones del conjunto de datos de entrada, es decir las 976 muestras de 2 segundos de audio, con 50 y 100 epochs.

4.4.1.3 Modelo Sample-level CNN adaptado

La particularidad de este modelo es que los datos de entrada en este caso son los datos audio a 16KHz directamente sin preprocesar. Se ha querido probar este modelo precisamente por el interés en evaluar este tipo de dato de entrada, que es similar a la que llega al oído humano. El tamaño de las muestras para 2 segundos es de 32000x1.

La Ilustración 92 resume este modelo. En la capa convolucional se consideran 11 convoluciones con filtros 3x3, y activación ReLU. Se realiza BatchNormalization y Max Pooling 2x2 tras cada convolución. El resultado, redimensionado a un vector unidimensional, se procesa en la capa completamente conectada de 512 neuronas, con activación ReLU, y una

última capa de salida con activación Softmax. En la capa Dense se realiza Dropout para evitar sobreajustes.

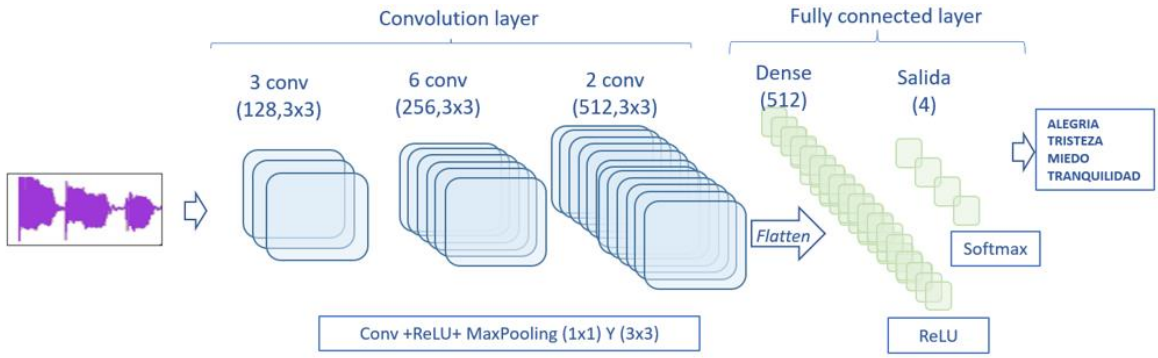


Ilustración 92. Modelo Sample-level CNN adaptado

Para la evaluación del modelo se utilizó validación cruzada, considerando distintas repeticiones sobre diferentes particiones del conjunto de datos de entrada, es decir las 976 muestras de 2 segundos de audio, con 50 y 100 epochs.

4.4.2 Resultados

4.4.2.1 Modelo Short-chunk CNN adaptado

Los resultados de Accuracy de distintas validaciones cruzadas con las 976 muestras de espectrograma CQT con este modelo se muestran en la Tabla 20.

Accuracy	
50 epochs	100 epochs
0,80	0,86
0,59	0,83
0,81	0,79
0,83	0,77
0,82	0,80
0,76	**0,81**

Tabla 20. Resultados Accuracy validación cruzada modelo Short chunk CNN adaptado 4 emociones

Se observa que con 100 epochs se obtienen mejores resultados (media 81%) que con 50 epochs, resultando en una mejora respecto al modelo utilizado en la anterior experimentación (media 78%). Posiblemente este modelo más profundo requiera de más tiempo de entrenamiento.

Los resultados obtenidos con el modelo entrenado y evaluado con el conjunto de entrenamiento (732 muestras) y validación (244 muestras) compuestos por muestras de fragmentos musicales distintos en cada set, y con 100 epochs se muestran en la Tabla 21. Se muestran también las métricas de Precisión, Recall Y F1. El proceso se repitió 4 veces cambiando cada vez el conjunto de entrenamiento y validación.

Accuracy	precision	recall	f1-score
0,60	0,62	0,61	0,58
0.66	0,67	0,68	0,65
0,54	0,59	0,55	0,52
0,62	0,63	0,66	0,62
0,58	**0,63**	**0,62**	**0,59**

Tabla 21. Resultados de validación con muestras de fragmentos musicales no utilizados durante el entrenamiento modelo Short chunk CNN adaptado a 4 emociones

Estos resultados (Accuracy media 58%) muestran que el modelo no generaliza bien en base a estas 4 emociones. Teniendo en cuenta los resultados de la experimentación anterior (sección4.3.2), en los que se vio que en la clasificación automática de las emociones podían producirse confusiones entre Tristeza y Tranquilidad, debido a que compartían parte de los valores medios de los parámetros musicales más significativos en cuanto a su relación con la emoción, se realizó el mismo proceso, pero considerando únicamente las 3 emociones básicas de alegría, tristeza y miedo, descartando los fragmentos representativos de tranquilidad.

4.4.2.1.1 Modelo Short-chunk CNN adaptado (3 emociones)

Los resultados de la validación cruzada del modelo de clasificación considerando 3 emociones básicas se muestran en la Tabla 22. Se observa una mejora notable respecto al modelo adaptado a 4 emociones, alcanzando un valor de Accuracy de 89% con 3 emociones frente a 81% con 4 emociones.

Accuracy (100 epochs)
0,91
0,85
0,92
0,85
0,93
0,89

Tabla 22. Resultados Accuracy validación cruzada modelo Short chunk CNN adaptado 3 emociones

Esta mejora se observa también en la capacidad de generalización, alcanzando un valor medio de Accuracy de 79% (frente a 58% con 4 emociones) con el modelo entrenado y evaluado con el conjunto de entrenamiento (732 muestras) y validación (244 muestras) compuestos por muestras de fragmentos musicales distintos en cada set, y con 100 epochs. El proceso se repitió 4 veces cambiando cada vez el conjunto de entrenamiento y validación. Estos resultados se detallan en la Tabla 23.

Accuracy	precision	recall	f1-score
0,79	0,79	0,79	0,79
0,80	0,79	0,81	0,79
0,81	0,82	0,79	0,8
0,76	0,77	0,77	0,77
0,79	**0,79**	**0,79**	**0,79**

Tabla 23. Resultados (Precisión, Recall, F1) de validación con muestras de fragmentos musicales no utilizados durante el entrenamiento (Short chunk CNN adaptada a 3 emociones)

El modelo muestra puntuaciones similares en precisión y recall, es decir, tanto en la capacidad de no clasificar como positivas muestras negativas, como en la capacidad de reconocer todos los ejemplos positivos. Si además se considera la media de los valores de precisión en función de las 3 emociones y los comparamos con los resultados obtenidos por (Vieillard et al., 2008) en su estudio para la base de datos de Musical Excerpts con fragmentos musicales representativos de emociones básicas intensas, se comprueba que la precisión en el reconocimiento del miedo y la tristeza es muy similar.

		Precisión
Vieillard		Short Chunk 3 emociones
ALEGRÍA	0,99	0,78
MIEDO	0,72	0,75
TRISTEZA	0,84	0,84

Tabla 24. Comparativa resultados (Vieillard et al., 2008) y modelo Short chunk CNN adaptado 3 emociones

4.4.2.2 Modelo Short-chunk CNN con ResNet adaptado

Los resultados medios de Accuracy en la validación cruzada de la clasificación de 4 emociones con 50 y 100 epochs de este modelo (69% y 72%) fueron peores que con el modelo Short-chunk CNN (76% y 81%). Los resultados obtenidos se resumen en la siguiente tabla.

Accuracy	
50 epochs	100 epochs
0,65	0,68
0,61	0,67
0,73	0,70
0,70	0,79
0,80	0,75
0,69	**0,72**

Tabla 25. Resultados Accuracy de validación cruzada modelo Short-chunk CNN con ResNet adaptado a 4 emociones

4.4.2.3 Modelo Musiccnn adaptado

Los resultados medios de Accuracy en la validación cruzada de la clasificación de 4 emociones con 50 y 100 epochs de este modelo (60% y 62%) fueron peores que con el modelo Short-chunk CNN (76% y 81%). Los resultados de Accuracy obtenidos se resumen en la siguiente tabla.

Accuracy	
50 epochs	100 epochs
0,50	0,50
0,67	0,63
0,59	0,61
0,62	0,67
0,66	0,78
0,60	**0,62**

Tabla 26. Resultados Accuracy de validación cruzada modelo Musicnn adaptado a 4 emociones

4.4.2.4 Modelo Sample-level CNN adaptado

Los resultados medios de Accuracy en la validación cruzada de la clasificación de 4 emociones con 50 epochs de este modelo (60%) fueron peores que con el modelo Short-chunk CNN (76%). Los resultados de Accuracy obtenidos se resumen en la siguiente tabla.

Accuracy
50 epochs
0,62
0,65
0,50
0,67
0,59
0,60

Tabla 27. Resultados Accuracy modelo Sample-level CNN adaptado a 4 emociones

4.4.3 Discusión y conclusiones

Las distintas experimentaciones realizadas muestran por un lado que los espectrogramas CQT, curiosamente los que mejor representan las relaciones entre tonos musicales, son los que ofrecen mejores resultados cuando se utilizan como conjunto de datos de entrada al modelo CNN. Además, el tiempo de procesamiento que requieren es mucho menor que el resto de los espectrogramas (Tabla 17).

Por otro lado, las distintas experimentaciones realizadas muestran que los modelos CNN de arquitectura más sencilla, relativamente profundos (8 capas convolucionales), con convoluciones con estructura sencilla:

Filtro(3x3) + BatchNormalization + Activation ReLu + MaxPooling(2,2)

ofrecen mejores resultados que otros modelos más complejos que incluyen por ejemplo bloques ResNet o Inception, tal y como ser resume en la siguiente tabla.

Accuracy (media)		
	50 epochs	*100 epochs*
Short Chunk CNN adaptado	76%	81%
Short Chunk CNN +ResNet adaptado	69%	72%
Musicnn adaptado	60%	62%
Sample Level CNN adaptado	60%	

Tabla 28. Resultados medios Accuracy de validación cruzada en los diferentes modelos evaluados (con 4 emociones)

Así, el modelo Short Chunk CNN adaptado es el que mejores resultados obtiene, y la clasificación mejora cuando se consideran únicamente las 3 emociones básicas de alegría,

tristeza y miedo, al eliminar la emoción de tranquilidad, de acuerdo con los resultados obtenidos:

- Modelo de clasificación en 4 emociones: Accuracy media 81% en validación cruzada, y 58% en la validación con muestras de fragmentos musicales distintos.

- Modelo de clasificación en 3 emociones: Accuracy media 89% en validación cruzada y 79% en la validación con muestras de fragmentos musicales distintos.

Como ya se ha visto (sección 2.2.3), la tranquilidad no se considera una emoción básica, pero se percibe con frecuencia como emoción musical, y se caracteriza por parámetros musicales en parte comunes a la emoción básica de tristeza, por lo que en la clasificación automática puede dar lugar a confusión.

Hay que tener en cuenta la limitación de la muestra utilizada en los resultados obtenidos, sobre todo, por el riesgo de sobreajuste al tratarse de grupos de muestras pertenecientes a mismos fragmentos musicales. Aun así, se puede considerar que la arquitectura CNN sencilla puede ser una arquitectura eficaz en la clasificación de la emoción en fragmentos de audio de 2 segundos. El modelo muestra ser eficaz en las emociones básicas de alegría, tristeza, y miedo que son precisamente las más interesantes de identificar en el caso del subtitulado de la música de películas, acercándose a los resultados de los experimentos neurocientíficos, con 2 segundos de muestra, tiempo suficiente para transmitir la emoción de forma inmediata.

Además, frente a otros modelos presenta la gran ventaja de no requerir una selección previa de las características de las muestras de audio, ni el soporte de datos adicionales a las muestras de audio, lo que hace más sencilla su aplicación. Por tanto, puede ser una base sencilla y eficaz para la extracción de la emoción de la música de cara a un subtitulado accesible, junto con el uso de espectrogramas CQT.

5 CONCLUSIONES Y TRABAJOS FUTUROS

5.1 Conclusiones

A lo largo de esta investigación se ha podido avanzar en los objetivos inicialmente establecidos alcanzando todos y cada uno de los planteados en el capítulo 1 de la presente memoria.

En primer lugar, se planteaba la necesidad de establecer un marco científico en el que sustentar la investigación. A raíz del estudio del estado del arte se han podido establecer unos puntos de partida científicos, que han permitido orientar el desarrollo de las investigaciones propuestas, así como sugerir nuevos trabajos futuros. Estos puntos se detallan en el capítulo 2 y se resumen a continuación.

Respecto a la emoción musical, los puntos de interés más destacados para esta investigación se detallan en la sección 2.1 y se resumen en:

- La emoción está asociada a circuitos neuronales primarios.
- Las emociones básicas de alegría, tristeza, miedo, son universales e inmediatas.
- El cerebro procesa relaciones entre frecuencias y le *gustan* las relaciones matemáticas sencillas.
- Las emociones básicas e intensas de alegría, tristeza y miedo son las más claramente identificables en la música y se asocian a valores claros de los parámetros modo y tempo principalmente.

Respecto a la percepción vibro táctil, presentada en la sección2.2, los puntos de interés más destacados son:

- Existencia de una percepción multimodal vibro táctil-auditiva cerca de las áreas sensoriales primarias.
- Parámetros como el ritmo o el timbre se procesarían por un mismo mecanismo perceptivo común al oído y al tacto.
- El rango de percepción de la frecuencia táctil es de 5-1000 Hz frente a 20-20000 Hz en la percepción auditiva, y el intervalo mínimo detectado es de tercera menor, por lo que hay que adapta la señal acústica a estos rangos, que cubren las frecuencias fundamentales de la voz humana y las octavas más frecuentes de la música.
- Existen dispositivos que transmiten características rítmicas satisfactoriamente de forma vibro táctil, pero la transmisión de la información tonal aún no está resuelta.
- La reacción a la estimulación vibro táctil es similar en sujetos con y sin discapacidad auditiva, lo que facilita el reclutamiento de sujetos experimentales y amplia la aplicabilidad de la técnica en todos los sujetos.

Respecto los modelos MER de clasificación automática de le emoción en la música planteados en la sección 2.3, se han encontrado muchas limitaciones que se resumen en:

- No existe un framework común de experimentación.

- No hay consenso en la literatura sobre las características de audio significativas para la captura de la emoción.
- No existen algoritmos de aprendizaje automático robustos para capturar las relaciones música-emoción.
- Las redes neuronales CNN ofrecen un enfoque más sencillo que enfoques estadísticos basados en regresión, utilizando espectrogramas de las muestras audio como imágenes de entrada.
- Las arquitecturas CNN más sencillas son las que ofrecen mejores resultados en la clasificación de género musical.

En segundo lugar, se quería desarrollar una propuesta de investigación que permitiera aportar ideas base para el desarrollo de un framework de subtitulado accesible de la música de películas alternativo al canal tradicional normativo. Este framework incluiría una funcionalidad de extracción automática de la emoción musical y una funcionalidad de transmisión de estas emociones a través del canal vibro táctil. Respecto a este objetivo, los resultados de las experimentaciones realizadas han permitido validar las hipótesis de partida y aportar las ideas base que se resumen a continuación.

- **Hipótesis 1**: *El subtitulado accesible textual no transmite la información que aporta la música de forma inmediata a través de la emoción.*

  Las medidas EEG realizadas muestran que, por el contrario, el subtitulado textual incrementa la activación cerebral en zonas de procesamiento visual y verbal y no las zonas mas emocionales del cerebro tal y como se demuestra en la sección 3.3.

- **Hipótesis 2**: *El tacto puede ser un canal de transmisión alternativo de emociones musicales básicas.*

  Las medidas EEG realizadas muestran que las zonas cerebrales involucradas en el procesamiento de la música, al menos las que se pueden medir mediante EEG, se activan significativamente durante una proyección audiovisual acompañada de suave estimulación vibro táctil basada en un patrón rítmico tal y como se refleja en la sección 3.4.

- **Hipótesis 3**: *Los parámetros musicales pueden transmitirse de forma similar, aunque más limitada, mediante estimulación vibro-táctil.*

  Los resultados muestran que una simple estimulación vibro táctil, reproduciendo un patrón simple de tempo durante una proyección audiovisual, activas zonas cerebrales análogas a las que produce la misma proyección con música. En la sección 3.4 se observan estos resultados.

- **Hipótesis 4**: *Los modelos de aprendizaje CNN permiten clasificar emociones básicas (alegría, tristeza, miedo) en fragmentos breves de música de película.*

  Los modelos de aprendizaje CNN, basados en arquitecturas sencillas utilizando espectrogramas CQT como imágenes de entrada a la red, obtienen resultados de clasificación de las emociones básicas de alegría, tristeza y miedo, en fragmentos musicales de 2 segundos representativos de emociones básicas intensas, similares a los obtenidos en los estudios neurocientíficos con oyentes. En el capítulo 4 se resumen las principales características de las CNN, su aplicación a la clasificación de la música así como los resultados de la experimentación.

El objetivo inicial de esta investigación era aportar ideas base para el desarrollo de un framework de *subtitulado accesible* de la música de películas alternativo al canal textual. La reacción positiva a una suave y sencilla estimulación rítmica vibro táctil, alienta a continuar en la investigación del canal vibro táctil que parece ser capaz de aportar nuevas soluciones, como alternativa a los subtítulos tradicionales, para transmitir la información emocional contenida en la banda sonora audiovisual, y así producir la intención emocional del autor en sujetos con discapacidad auditiva. Mientras que los modelos de aprendizaje CNN, basados en arquitecturas sencillas, presentan una solución simple y eficaz como base de la clasificación automática de fragmentos musicales en base a emociones básicas.

5.2 Trabajos futuros

Las propuestas de investigación desarrolladas han permitido establecer unas bases iniciales de trabajo de cara al desarrollo de un framework de subtitulado accesible de la música y han abierto la posibilidad de continuar las investigaciones en varias direcciones.

Por un lado, habría que profundizar en el estudio de la actividad cerebral inducida por la estimulación vibro táctil, estudiando la activación de zonas cerebrales más profundas, que no han podido ser analizadas con los registros EEG. Para ello es necesario la utilización de técnicas fMRI de resonancia magnética. El objetivo sería realizar una experimentación análoga a la experimentación descrita en la sección 3.4, midiendo la activación cerebral mediante fMRI en las condiciones experimentales de silencio, audio y estimulación vibro táctil, tanto en participantes sin discapacidad auditiva como en participantes con discapacidad auditiva. El objetivo sería ampliar el estudio ya realizado con EEG, explorando las zonas más internas del cerebro (marcadas en rojo en la Ilustración 93), para verificar si se activan con la estimulación vibro táctil de la misma forma que con la música, como sugiere la experimentación realizada con EEG.

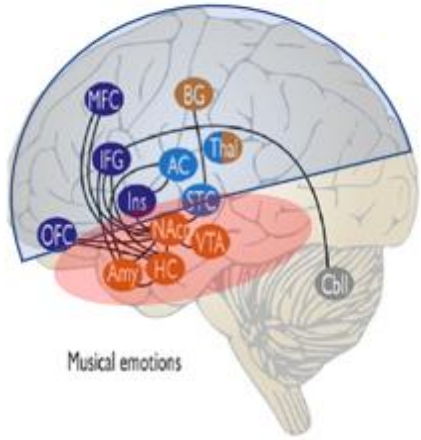


Ilustración 93. Zonas cerebrales activadas por la música. En azul, zonas medibles mediante EEG. En rojo zonas medibles mediante fMRI.

Otra línea de trabajo, que apenas se ha esbozado en esta investigación, es la de establecer patrones vibro táctiles correspondientes a los parámetros musicales asociados a la emoción. Para iniciar la investigación desde una base sencilla, se propone realizar una experimentación que explore si los intervalos musicales consonantes se producen de forma análoga a nivel vibro táctil. Es decir, los intervalos de octava, quinta y cuarta justa que *gustan* el oído, *¿gustan* también a nivel vibro táctil? Se propone aquí una experimentación sencilla basada en la configuración experimental utilizada por (Kuroki et al., 2017), en la que se presenta a los sujetos experimentales distintas frecuencias vibro táctiles en dos dedos de la misma mano, o en dos dedos en manos distintas, mediante un actuador piezoeléctrico. El objetivo de la investigación de (Kuroki et al., 2017) era ver cuál era la frecuencia percibida resultante, obteniendo que los participantes reportaban consistentemente una frecuencia promediada en base a la frecuencia e intensidad de las frecuencias presentadas. Lo que se propone comprobar, con un entorno experimental análogo, es si a nivel vibro táctil se perciben consonantes las frecuencias presentadas con relación (F,2F), (F,3/2F) y (F,3/4F), correspondientes a los intervalos de octava, quinta y cuarta justa, en dedos de la misma mano, y en dedos de distintas manos.

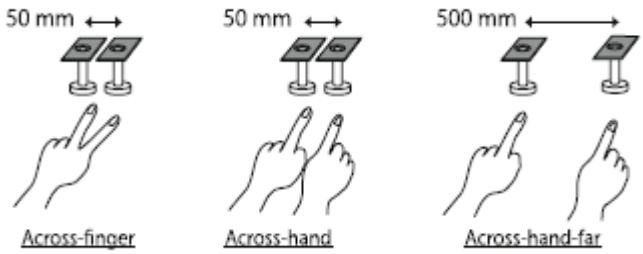


Ilustración 94. Esquema de la experimentación realizada por (Kuroki et al., 2017)

Para esta experimentación los participantes serían participantes con audición normal, ya que se ha visto que la reacción a la estimulación vibro táctil es similar en sujetos con y sin discapacidad auditiva. Debido a la dificultad de reclutar a participantes con discapacidad auditiva, sería más interesante establecer las primeras conclusiones con los participantes con audición normal, y ajustar las condiciones experimentales antes de realizar la experimentación con participantes con discapacidad auditiva.

Por último, respecto a los modelos CNN, hay que trabajar en dos frentes inmediatos. Por un lado, optimizar los modelos para obtener mejores resultados de generalización. Y por otro trabajar en la obtención de una base de datos de música de películas más amplia que permita entrenar los modelos de forma eficiente. Esta base de datos podría obtenerse a partir de películas subtituladas de forma accesible y con calidad, extrayendo las bandas sonoras y el correspondiente subtitulado accesible de la música producido por subtituladores profesionales, lo que sería equivalente a un etiquetado de calidad de un conjunto de datos muy representativo. Finalmente, sería interesante estudiar los filtros generados por los modelos CNN, y analizar los parámetros que estos modelos destacan como características audio más representativas de la emoción musical.

5.3 Aportaciones

Las aportaciones fundamentales de esta tesis (incluidas en las publicaciones de relevancia realizadas) son:

- En primer lugar, llevar a cabo una revisión bibliográfica profunda de distintos campos que se han relacionado para crear un marco teórico que sustentara esta investigación multidisciplinar.
- Fundamentar que la transmisión de la información de la emoción de la música a través de la piel en personas con discapacidad auditiva es una excelente alternativa al subtitulado textual, o en emoticonos, de la música y la experimentación ha validado esta hipótesis.
- Desarrollar un clasificador de la música eficaz para determinar una taxonomía de diferentes emociones (tristeza, alegría, miedo y tranquilidad) de manera automática en cortos fragmentos musicales de películas, que ha sido validado a través de la experimentación.
- Desarrollar un procedimiento para alimentar un guante vibro táctil con vibraciones rítmicas y demostrar que parámetros musicales como el ritmo son percibidos emocionalmente por personas con discapacidad auditiva. Se ha realizado una primera experimentación que valida esta hipótesis y abre un campo científico para incorporar nuevos parámetros musicales y patrones emocionales asociados.
- Adicionalmente se ha creado un *framework* de experimentación que será útil en futuros trabajos y que ha servido para integrar todas las hipótesis y permitir que las personas con discapacidad auditiva puedan percibir de manera táctil la emoción de la música.
- Se ha contribuido científicamente con dos artículos indexados, un capítulo de libro, tres ponencias en congresos del ámbito de la accesibilidad y la informática.

Este trabajo es el comienzo de una nueva línea de investigación y proporciona herramientas a los investigadores para que se pueda lograr una integración de la transmisión háptica en un espacio futuro de Subtitulado Multimodal.

6 BIBLIOGRAFÍA

Adachi, M., & Trehub, S. E. (1998). Children's expression of emotion in song. Psychology of Music, 26(2), 133-153. doi:10.1177/0305735698262003

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature, 372(6507), 669-672. doi:10.1038/372669a0

AENOR. (2012). Norma UNE 153010:2012. subtitulado para personas sordas y personas con discapacidad auditiva. Retrieved from https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0049426

Al-Nafjan, A., Hosny, M., Al-Ohali, Y., & Al-Wabil, A. (2017). Review and classification of emotion recognition based on EEG brain-computer interface system research: A systematic review. Applied Sciences, 7(12) doi:10.3390/app7121239

Altozano, J. (2021). Por qué no tienes oído absoluto (pero sí color absoluto). Retrieved from https://www.youtube.com/watch?v=w192BHvVri4

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. Nature Reviews Neuroscience, 7(4), 268-277. doi:10.1038/nrn1884

Anna Aljanaki, Yi-Hsuan Yang, & Mohammad Soleymani. (2017). Developing a benchmark for emotional analysis of music. PLoS ONE 12(3): e0173392, doi:10.1371/journal.pone.0173392

Ant-neuro. (2021). Electrode layouts. Retrieved from https://www.ant-neuro.com/products/waveguard/electrode-layouts

Baijal, A., Kim, J., Branje, C., Russo, F., & Fels, D. I. (March 2012). Composing vibrotactile music: A multi-sensory experience with the emoti-chair. Paper presented at the EEE Haptics Symposium (HAPTICS), 509-515. doi:10.1109/HAPTIC.2012.6183839

Balkwill, L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music Perception, 17(1), 43-64. doi:10.2307/40285811

Balkwill, L., Thompson, W. F., & Matsunaga, R. (2004). Recognition of emotion in japanese, western, and hindustani music by japanese listeners1. Japanese Psychological Research, 46(4), 337-349. doi:10.1111/j.1468-5584.2004.00265.x

Bamiou, D., Musiek, F. E., & Luxon, L. M. (2003). The insula (island of reil) and its role in auditory processing: Literature review. Brain Research Reviews, 42(2), 143-154. doi:10.1016/S0165-0173(03)00172-3

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 40(2), 531-539. doi:10.3758/BRM.40.2.531

Bengio, Y., & Lecun, Y. (1995). Convolutional networks for images, speech, and time-series. Handbook of brain theory and neural networks (pp. 255-258) MIT Press, Cambridge, Massachusetts.

Bianchin, M., & Angrilli, A. (2011). Decision preceding negativity in the IOWA gambling task: An ERP study. Brain and Cognition, 75(3), 273-280. doi:10.1016/j.bandc.2011.01.005

BIAP. (1996). BIAP recommendation 02/1: Audiometric classification of hearing impairments. . Retrieved from https://www.biap.org/en/recommandations/recommendations/tc-02-classification/213-rec-02-1-enaudiometric-classification-of-hearing-impairments/file. Accessed February 4,2020

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion, 19(8), 1113-1139. doi:10.1080/02699930500204250

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion Proceedings of the National Academy of Sciences Sep. 2001. doi:rg/10.1073/pnas.191355898

Böcker, K. B. E., Baas, J. M. P., Kenemans, J. L., & Verbaten, M. N. (2001). Stimulus-preceding negativity induced by fear: A manifestation of affective anticipation. International Journal of Psychophysiology, 43(1), 77-90. doi:10.1016/S0167-8760(01)00180-5

BOE. (2010). Ley 7/2010, de 31 de marzo, general de la comunicación audiovisual. Retrieved from https://www.boe.es/buscar/pdf/2010/BOE-A-2010-5292-consolidado.pdf

BOE. (2019). Real decreto 94/2019 de 1 de marzo, por el que se establece el curso de especialización en audiodescripción y subtitulación y se fijan los aspectos básicos del currículo. Retrieved from https://www.boe.es/boe/dias/2019/03/22/pdfs/BOE-A-2019-4153.pdf

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., . . . Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. Paper presented at the International Society for Music Information Retrieval (ISMIR'13), 493-498. Retrieved from http://repositori.upf.edu/handle/10230/32252

Bradley, M. (2009). Natural selective attention: Orienting and emotion. Psychophysiology, 46(1), 1–11. doi:10.1111/j.1469–8986.2008.00702.x

Bradley, M. M., Hamby, S., Löw, A., & Lang, P. J. (2007). Brain potentials in perception: Picture complexity and emotional arousal. Psychophysiology, 44(3), 364–373. doi:10.1111/j.1469–8986.2007.00520.x

Bradley, M. M., & Lang, P. J. (2017). International affective picture system. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), Encyclopedia of personality and individual differences (pp. 1–4). Cham: Springer International Publishing. doi:10.1007/978–3–319–28099–8_42–1

Branje, C., & Fels, D. I. (2014). Playing vibrotactile music: A comparison between the vibrochord and a piano keyboard. International Journal of Human–Computer Studies, 72(4), 431–439. doi:10.1016/j.ijhcs.2014.01.003

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron (Cambridge, Mass.), 30(2), 619–639. doi:10.1016/S0896–6273(01)00303–8

Bresin, R., & Friberg, A. (2011). Emotion rendering in music: Range and characteristic values of seven musical variables. Cortex, 47(9), 1068–1081. doi:10.1016/j.cortex.2011.05.009

Brown, D. R., & Cavanagh, J. F. (2017). The sound and the fury: Late positive potential is sensitive to sound affect. Psychophysiology, 54(12), 1812–1825. doi:10.1111/psyp.12959

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. Trends in Cognitive Sciences, 4(6), 215–222. doi:10.1016/S1364–6613(00)01483–2

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. Nature Neuroscience, 6(2), 190–195. doi:10.1038/nn993

Caetano, G., & Jousmäki, V. (2006). Evidence of vibrotactile input to human auditory cortex. NeuroImage, 29(1), 15–28. doi:10.1016/j.neuroimage.2005.07.023

Cela-Conde, C. J., Ayala, F. J., Munar, E., Maestú, F., Nadal, M., Capó, M. A., . . . Marty, G. (2009). Sex-related similarities and differences in the neural correlates of beauty. Proceedings of the National Academy of Sciences, 106(10), 3847–3852. doi:rg/10.1073/pnas.0900304106

Cesya. (2014). Seguimiento del subtitulado y audiodescripción en la TDT. Retrieved from https://www.cesya.es/sites/default/files/documentos/InformeAccesibilidadTDT2014.pdf

Cinzia, D. D., & Vittorio, G. (2009). Neuroaesthetics: A review. Current Opinion in Neurobiology, 19(6), 682–687. doi:10.1016/j.conb.2009.09.001

Cohen, A. L., Fair, D. A., Dosenbach, N. U. F., Miezin, F. M., Dierker, D., Van Essen, D. C., . . . Petersen, S. E. (2008). Defining functional areas in individual human brains using resting functional connectivity MRI. NeuroImage, 41(1), 45–57. doi:10.1016/j.neuroimage.2008.01.066

Convention on the rights of persons with disabilities (CRPD). (2006). Retrieved from https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

Convento, S., Wegner-Clemens, K. A., & Yau, J. M. (2019). Reciprocal interactions between audition and touch in flutter frequency perception. Multisensory Research, 32(1), 67–85. doi:10.1163/22134808-20181334

Craig, A. D. (2002). How do you feel? interoception: The sense of the physiological condition of the body. Nature Reviews Neuroscience, 3(8), 655–666. doi:10.1038/nrn894

Craig, A. D. (2004). Human feelings: Why are some more aware than others? Trends in Cognitive Sciences, 8(6), 239–241. doi:10.1016/j.tics.2004.04.004

Crommett, L. E., Pérez-Bellido, A., & Yau, J. M. (2017). Auditory adaptation improves tactile frequency perception. Journal of Neurophysiology, 117(3), 1352–1362. doi:10.1152/jn.00783.2016

Culurciello, E. (2017). Neural network architectures. Retrieved from https://towardsdatascience.com/neural-network-architectures-156e5bad51ba

Dalla, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. Cognition, 80(3), B1–B10. doi:10.1016/S0010-0277(00)00136-0

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. Science (New York, N.Y.), 313(5787), 684–687. doi:10.1126/science.1128356

Debevc, M., Milošević, D., & Kožuh, I. (2015). A comparison of comprehension processes in sign language interpreter videos with or without captions. Plos One, 10(5), e0127577. doi:10.1371/journal.pone.0127577

Díaz, J. (2006). Competencias profesionales
del subtitulador y el audiodescriptor . Retrieved from
https://www.cesya.es/sites/default/files/documentos/informe_formacion.pdf

Dolgin, K. G., & Adelson, E. H. (1990). Age changes in the ability to interpret affect in sung and instrumentally-presented melodies. Psychology of Music, 18(1), 87-98. doi:10.1177/0305735690181007

Donnelly, K. J. (2005). The spectre of sound: Music in film and television. London: British Film Institute. doi:10.5040/9781838711009

Duncan, C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P., Näätänen, R., . . . Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 120(11), 1883-1908. doi:10.1016/j.clinph.2009.07.045

D'Ydewalle, G., Praet, C., Verfaillie, K., & Rensbergen, J. V. (1991). Watching subtitled television: Automatic reading behavior. Communication Research, 18(5), 650-666. doi:10.1177/009365091018005005

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. Paper presented at the 10th International Society for Music Information Retrieval Conference (ISMIR 2009),

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. Frontiers in Psychology, 4 doi:10.3389/fpsyg.2013.00487

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. Psychology of Music, 39(1), 18-49. doi:10.1177/0305735610362821

Ekman, P. (1992a). Are there basic emotions? Psychological Review, 99(3), 550-553. doi:10.1037/0033-295X.99.3.550

Ekman, P. (1992b). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169-200. doi:10.1080/02699939208411068

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. Paper presented at the IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, 1813-1817 vol.3. doi:10.1109/NSSMIC.1993.373602

Farnsworth, B. (2019). EEG (electroencephalography): The complete pocket guide. Retrieved from https://imotions.com/blog/eeg/

Feng, Y., Zhuang, Y., & Pan, Y. (2003). Popular music retrieval by detecting mood. Paper presented at the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion 375-376. doi:10.1145/860435.860508

Filipic, S., Tillmann, B., & Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. Psychonomic Bulletin & Review, 17(3), 335-341. doi:10.3758/PBR.17.3.335

Fleuret, F. (2021). Deep learning course of UNIGE/EPFL. Retrieved from https://fleuret.org/dlc/materials/dlc-handout-6-1-benefits-of-depth.pdf

Fogassi, L., & Luppino, G. (2005). Motor functions of the parietal lobe. Current Opinion in Neurobiology, 15(6), 626-631. doi:10.1016/j.conb.2005.10.015

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., . . . Koelsch, S. (2009). Universal recognition of three basic emotions in music. Current Biology, 19(7), 573-576. doi:10.1016/j.cub.2009.02.058

Frühholz, S., Trost, W., & Kotz, S. A. (2016). The sound of emotions—Towards a unifying neural network perspective of affective sound processing. Neuroscience & Biobehavioral Reviews, 68, 96-110. doi:10.1016/j.neubiorev.2016.05.002

Fuentes, V., González, I. & Ruiz, B. (2007). Subtitulado en tiempo real. sistemas y tecnología . Retrieved from https://www.cesya.es/sites/default/files/documentos/Subtitulado_tiempo_real.pdf

Fulford, R., Ginsborg, J., & Goldbart, J. (2011). Learning not to listen: The experiences of musicians with hearing impairments. Music Education Research, 13(4), 447-464. doi:10.1080/14613808.2011.632086

Gabrielsson, A. (2001). Emotions in strong experiences with music. Music and emotion: Theory and research (pp. 431-449). New York, NY, US: Oxford University Press.

Gabrielsson, A., Lindström, E. (2010). The role of structure in the musical expression of emotions. Music and emotion: Theory, research, applications (pp. 367-400) Oxford University Press. doi:10.1093/acprof:oso/9780199230143.003.0014

Gabrielsson, A., & Juslin, P. N. (2003). Emotional expression in music. Handbook of affective sciences (pp. 503-534). New York, NY, US: Oxford University Press.

Gerdes, A. B. M., Wieser, M. J., Bublatzky, F., Kusay, A., Plichta, M. M., & Alpers, G. W. (2013). Emotional sounds modulate early neural processing of emotional pictures. Frontiers in Psychology, 4, 741. doi:10.3389/fpsyg.2013.00741

Gianotti, L. R. R., Faber, P. L., Schuler, M., Pascual-Marqui, R., Kochi, K., & Lehmann, D. (2008). First valence, then arousal: The temporal dynamics of brain electric activity evoked by emotional stimuli. Brain Topography, 20(3), 143-156. doi:10.1007/s10548-007-0041-2

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. Neurosurgery Clinics of North America, 22(2), 133-139. doi:10.1016/j.nec.2010.11.001

Gorzelańczyk, E. J., Podlipniak, P., Walecki, P., Karpiński, M., & Tarnowska, E. (2017). Pitch syntax violations are linked to greater skin conductance changes, relative to timbral violations – the predictive role of the reward system in perspective of cortico-subcortical loops. Frontiers in Psychology, 8 doi:10.3389/fpsyg.2017.00586

Gosselin, N., Peretz, I., Johnsen, E., & Adolphs, R. (2007). Amygdala damage impairs emotion recognition from music. Neuropsychologia, 45(2), 236-244. doi:10.1016/j.neuropsychologia.2006.07.012

Gray, J. M., Young, A. W., Barker, W. A., Curtis, A., & Gibson, D. (1997). Impaired recognition of disgust in huntington's disease gene carriers. Brain, 120(11), 2029-2038. doi:10.1093/brain/120.11.2029

Gulliver, S. R., & Ghinea, G. (2003). How level and type of deafness affect user perception of multimedia video clips. Universal Access in the Information Society, 2(4), 374-386. doi:10.1007/s10209-003-0067-5

Gunther, E., & O'Modhrain, S. (2003). Cutaneous grooves: Composing for the sense of touch. Journal of New Music Research, 32(4), 369-381. doi:10.1076/jnmr.32.4.369.18856

Hailstone, J. C., Omar, R., Henley, S. M. D., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. Quarterly Journal of Experimental Psychology, 62(11), 2141-2155. doi:10.1080/17470210902765957

Hajcak, G., MacNamara, A., & Olvet, D. M. (2010). Event-related potentials, emotion, and emotion regulation: An integrative review. Developmental Neuropsychology, 35(2), 129-155. doi:10.1080/87565640903526504

Han, B., Rho, S., Dannenberg, R., & Hwang, E. (2009). SMERS: Music emotion recognition using support vector regression. Paper presented at the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), .651-656.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Paper presented at the EEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. doi:10.1109/CVPR.2016.90

Hopkins, C., Maté-Cid, S., Fulford, R., Seiffert, G., & Ginsborg, J. (2016). Vibrotactile presentation of musical notes to the glabrous skin for adults with normal hearing or a hearing impairment: Thresholds, dynamic range and high-frequency perception. Plos One, 11(5), e0155807. doi:10.1371/journal.pone.0155807

Hopkins, C., Mate-Cid, S., Seiffert, G., Fulford, R., & Ginsborg, J. (2013). Inherent and learnt abilities for relative pitch in the vibrotactile domain using the fingertip. 20th International Congress on Sound and Vibration 2013, ICSV 2013, 4, 3207-3214.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-excitation network. arXiv.Org > Computer Science, arXiv: 1709.01507v4

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 mirex audio mood classification task: Lessons learned. Paper presented at the Ninth International Conference on Music Information Retrieval (ISMIR 2008),

Huang, J., Gamble, D., Sarnlertsophon, K., Wang, X., & Hsiao, S. (2012). Feeling music: Integration of auditory and tactile inputs in musical meter perception. PloS One, 7(10), e48496. doi:10.1371/journal.pone.0048496

Ito, T., Cacioppo, J., & Lang, P. (1998). Eliciting affect using the international affective picture system: Trajectories through evaluative space. Personality and Social Psychology Bulletin, doi:10.1177/0146167298248006

Jack, R., McPherson, A., & Stockman, T. (2015). Designing tactile musical devices with and for
deaf users: A case study . Paper presented at the International Conference on the Multimodal Experience of Music,

Jones, L. A., & Lederman, S. J. (2007). In Oxford Scholarship Online (Ed.), Human hand function doi:10.1093/acprof:oso/9780195173154.001.0001

Juslin, P. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology: Human Perception and Performance, 26(6), 1797-1812. doi:10.1037/0096-1523.26.6.1797

Juslin, P. N., & Sloboda, J. A. (2001). Music and emotion: Theory and research Oxford University Press. doi:10.1037/1528-3542.1.4.381

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), 770-814. doi:10.1037/0033-2909.129.5.770

Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: Modelling listeners' judgements of composed and performed features. Music Analysis, 29(1-3), 334-364. doi:10.1111/j.1468-2249.2011.00323.x

Kallinen, K. (2005). Emotional ratings of music excerpts in the western art music repertoire and their self-organization in the kohonen neural network. Psychology of Music, 33(4), 373-393. doi:10.1177/0305735605056147

Kastner, M. P., & Crowder, R. G. (1990). Perception of the major/minor distinction: IV. emotional connotations in young children. Music Perception, 8(2), 189–201. doi:10.2307/40285496

Kawabata, H., & Zeki, S. (2004). Neural correlates of beauty. Journal of Neurophysiology, 91(4), 1699–1705. doi:10.1152/jn.00696.2003

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. Neuron, 48(2), 373–384. doi:10.1016/j.neuron.2005.09.018

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. Journal of Neuroscience, 21(16), RC159. doi:10.1523/JNEUROSCI.21-16-j0002.2001

Koelsch, S. (2014). Brain correlates of music-evoked emotions. Nature Reviews Neuroscience, 15(3), 170–180. doi:10.1038/nrn3666

Koelsch, S., Fritz, T., Cramon, D. Y. v., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. Human Brain Mapping, 27(3), 239–250. doi:10.1002/hbm.20180

Koelsch, S., Skouras, S., Fritz, T., Herrera, P., Bonhage, C., Küssner, M. B., & Jacobs, A. M. (2013). The roles of superficial amygdala and auditory cortex in music-evoked fear and joy. NeuroImage, 81, 49–60. doi:10.1016/j.neuroimage.2013.05.008

Korzeniowski, F., Nieto, O., McCallum, M., Won, M., Oramas, S., & Schmidt, E. (2020). Mood classification using listening data. Arxiv.Org/Abs/2010.11512v1,

Kosonen, K., & Raisamo, R. (2006). Rhythm perception through different modalities. Paper presented at the EuroHaptics,

Kotz, S. A., Kalberlah, C., Bahlmann, J., Friederici, A. D., & Haynes, J. (2012). Predicting vocal emotion expressions from the human brain. Human Brain Mapping, 34(8), 1971–1981. doi:10.1002/hbm.22041

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. NeuroImage, 37(4), 1445–1456. doi:10.1016/j.neuroimage.2007.06.020

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale, 51(4), 336–353. doi:10.1037/1196-1961.51.4.336

Kuroki, S., Watanabe, J., & Nishida, S. (2017). Integration of vibrotactile frequency information beyond the mechanoreceptor channel and somatotopy. Scientific Reports, 7(1), 1-13. doi:10.1038/s41598-017-02922-7

Lang, P., Bradley, M., & Cuthbert, B. (1997). Motivated attention: Affect, activation, and action. In P. J. Lang, R. F. Simons, & M. T. Balaban (Ed.), Attention and orienting: Sensory and motivational processes (pp. 97-135) Lawrence Erlbaum Associates Publishers.

Lang, P. J., & Bradley, M. M. (2010). Emotion and the motivational brain. Biological Psychology, 84(3), 437-450. doi:10.1016/j.biopsycho.2009.10.007

Larsen, J. T., Berntson, G. G., Poehlmann, K. M., Ito, T. A., & Cacioppo, J. T. (2008). The psychophysiology of emotion. Handbook of emotions, 3rd ed (pp. 180-195). New York, NY, US: The Guilford Press.

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. Journal of Personality and Social Psychology, 61(1), 132-140. doi:10.1037/0022-3514.61.1.132

Lartillot, O. (2019). MIRtoolbox 1.7.2 user's manual. University of Oslo, Norway:

Lartillot, O., & Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. Paper presented at the 10th Int. Conference on Digital Audio Effects,

Law, E., West, K., Mandel, M. I., Bay, M., & Downie, J. S. (2009). Evaluation of algorithms using games: The
case of music tagging. Paper presented at the Event10th International Society for Music Information Retrieval Conference, ISMIR 2009, 387-392.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. doi:10.1109/5.726791

Leder, S. B., Spitzer, J. B., Milner, P., Flevaris-Phillips, C., & Richardson, F. (1986). Vibrotactile stimulation for the adventitiously deaf: An alternative to cochlear implantation. Archives of Physical Medicine and Rehabilitation, 67(10), 754-758. doi:10.1016/0003-9993(86)90010-9

Lenc, T., Keller, P. E., Varlet, M., & Nozaradan, S. (2018). Neural tracking of the musical beat is enhanced by low-frequency sounds. Proceedings of the National Academy of Sciences, 115(32), 8221-8226.

Li, T., Chan, A., & Chun, A. (2010). Automatic musical pattern feature extraction
using convolutional neural network. Paper presented at the International MultiConference of Enineers and Computer Scientists IMECS 2010,

Lifshitz, K. (1966). The averaged evoked cortical response to complex visual stimuli. Psychophysiology, 3(1), 55–68. doi:10.1111/j.1469-8986.1966.tb02680.x

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. The Behavioral and Brain Sciences, 35(3), 121–143. doi:10.1017/S0140525X11000446

Lindquist, M. A. (2008). The statistical analysis of fMRI data. Statistical Science, 23(4), 439–464. doi:10.1214/09-STS282

Lithari, C., Lithari, C., Frantzidis, C., Frantzidis, C., Papadelis, C., Papadelis, C., . . . Bamidis, P. (2010). Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. Brain Topography, 23(1), 27–40. doi:10.1007/s10548-009-0130-5

Lonsdale, A. J., & North, A. C. (2011). Why do we listen to music? A uses and gratifications analysis. British Journal of Psychology, 102(1), 108–134. doi:10.1348/000712610X506831

Lucía, M. J., Revuelta, P., García, Á, Ruiz, B., Vergaz, R., Cerdán, V., & Ortiz, T. (2020). Vibrotactile captioning of musical effects in audio-visual media as an alternative for deaf and hard of hearing people: An EEG study. IEEE Access, 8, 190873–190881. doi:10.1109/ACCESS.2020.3032229

Luo, X., & Hayes, L. (2019). Vibrotactile stimulation based on the fundamental frequency can improve melodic contour identification of normal-hearing listeners with a 4-channel cochlear implant simulatio. Frontiers in Neuroscience, 13, 1145. doi:10.3389/fnins.2019.01145

Market, G. (2020). Types of data & measurement scales: Nominal, ordinal, interval and ratio. Retrieved from https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Martino, B. D., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. Science, 313(5787), 684–687. doi:10.1126/science.1128356

Matsumoto, Y., Maeda, S., & OJI, Y. (2002). Influence of frequency on difference thresholds for magnitude of vertical sinusoidal whole-body vibration. Industrial Health, 40(4), 313–319. doi:10.2486/indhealth.40.313

Menon, V., & Levitin, D. J. (2005). The rewards of music listening: Response and physiological connectivity of the mesolimbic system. NeuroImage, 28(1), 175–184. doi:10.1016/j.neuroimage.2005.05.053

Merchel, S., & Altinsoy, M. E. (2013). Auditory-tactile music perception. Paper presented at the Meetings on Acoustics 2013, , 19 doi:10.1121/1.4799137

Mirz, F., Gjedde, A., Sdkilde-Jrgensen, H., & Pedersen, C. B. (2000). Functional brain imaging of tinnitus-like perception induced by aversive auditory stimuli. NeuroReport, 11(3), 633-637.

Molnar-Szakacs, I., & Overy, K. (2006). Music and mirror neurons: From motion to 'e'motion. Social Cognitive and Affective Neuroscience, 1(3), 235-241. doi:10.1093/scan/nsl029

Montagu, J. (2017). How music and instruments began: A brief overview of the origin and entire development of music, from its earliest stages. Frontiers in Sociology, 2 doi:10.3389/fsoc.2017.00008

Moreno, E. M., Casado, P., & Martín-Loeches, M. (2016). Tell me sweet little lies: An event-related potentials study on the processing of social lies. Cognitive, Affective, & Behavioral Neuroscience, 16(4), 616-625. doi:10.3758/s13415-016-0418-3

Morley, J. W., & Rowe, M. J. (1990). Perceived pitch of vibrotactile stimuli: Effects of vibration amplitude, and implications for vibration frequency coding. The Journal of Physiology, 431(1), 403-416. doi:10.1113/jphysiol.1990.sp018336

Oatley, K., & Johnson-laird, P. N. (1987). Towards a cognitive theory of emotions. Cognition and Emotion, 1(1), 29-50. doi:10.1080/02699938708408362

Oey, H., & Mellert, V. (2004). Vibration thresholds and equal vibration levels at the human fingertip and palm. Paper presented at the 5th
International Congress on Acoustics,

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. Biological Psychology, 77(3), 247-265. doi:10.1016/j.biopsycho.2007.11.006

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007a). The enigmatic temporal pole: A review of findings on social and emotional processing. Brain, 130(7), 1718-1731. doi:10.1093/brain/awm052

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007b). The enigmatic temporal pole: A review of findings on social and emotional processing. Brain, 130(7), 1718-1731. doi:10.1093/brain/awm052

Panda, R. (2019). Emotion-based analysis and classification of audio music. http://hdl.handle.net/10316/87618: Universidad de Coimbra.

Panda, R., & Paiva, R. P. (2012). Music emotion classification: Analysis of a classifier ensemble approach. Paper presented at the 5th International Workshop on Music and Machine Learning – MML'2012,

Panda, R., Malheiro, R. M., & Paiva, R. P. (2020). Audio features for music emotion recognition: A survey. IEEE Transactions on Affective Computing, , 1-1. doi:10.1109/TAFFC.2020.3032373

Pandrea, A. G., Gómez Cañón, J. S., & Herrera Boyer, P. (2020). Cross-dataset music emotion recognition: An end-to-end approach. Paper presented at the International Society of Music Information Retrieval Conference (ISMIR 2020),

Panksepp, J., & Bernatzky, G. (2002). Emotional sounds and the brain: The neuro-affective foundations of musical appreciation. Behavioural Processes, 60(2), 133-155. doi:10.1016/S0376-6357(02)00080-3

Papadogianni-Kouranti, M., Egermann, H., & Weinzierl, S. (2015). Auditive and audiotactile music perception of cochlear implant users. Paper presented at the Fortschritte Der Akustik - DAGA 2015, 1203-1205. doi:10.14279/depositonce-8774

Papetti, S., & Saitis, C. (2018). In Papetti, Stefano, Saitis, Charalampos (Ed.), Musical haptics Springer International Publishing.

Papetti, S., Järveläinen, H., Giordano, B. L., Schiesser, S., & Fröhlich, M. (2017a). Vibrotactile sensitivity in active touch: Effect of pressing force. IEEE Transactions on Haptics, 10(1), 113-122. doi:10.1109/TOH.2016.2582485

Papetti, S., Järveläinen, H., Giordano, B. L., Schiesser, S., & Fröhlich, M. (2017b). Vibrotactile sensitivity in active touch: Effect of pressing force. IEEE Transactions on Haptics, 10(1), 113-122. doi:10.1109/TOH.2016.2582485

Paquette, S., Peretz, I., & Belin, P. (2013). The "Musical emotional bursts": A validated set of musical affect bursts to investigate auditory affective processing. Frontiers in Psychology, 4, 509. doi:10.3389/fpsyg.2013.00509

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain. International Journal of Psychophysiology, 18(1), 49-65. doi:10.1016/0167-8760(84)90014-x

Pehrs, C., Deserno, L., Bakels, J., Schlochtermeier, L. H., Kappelhoff, H., Jacobs, A. M., . . . Kuchinke, L. (2014). How music alters a kiss: Superior temporal gyrus controls fusiform-amygdalar effective connectivity. Social Cognitive and Affective Neuroscience, 9(11), 1770-1778. doi:10.1093/scan/nst169

Pehrs, C., Zaki, J., Schlochtermeier, L. H., Jacobs, A. M., Kuchinke, L., & Koelsch, S. (2017). The temporal pole top-down modulates the ventral visual stream during social cognition. Cerebral Cortex, 27(1), 777-792. doi:10.1093/cercor/bhv226

Perego, E., Del Missie, F., & Bottiroli, S. (2015). Dubbing versus subtitling in young and older adults: Cognitive and evaluative aspects. Perspectives: Studies in Translatology, 23(1), 1-21. doi:10.1080/0907676X.2014.912343

Peretz, I. (2010). Towards a neurobiology of musical emotions. Handbook of music and emotion: Theory, research, applications (pp. 99-126). New York, NY, US: Oxford University Press.

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. Cognition, 68(2), 111-141. doi:10.1016/S0010-0277(98)00043-2

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature, 331(6157), 585-589. doi:10.1038/331585a0

Petisco, J. M. (2015). Comunicación no verbal: La figura de paul ekman. Retrieved from https://nonverbalbehavior.blogspot.com/2015/02/la-figura-de-paul-ekman.html

Polo, N. (2019). Las voces femeninas y su investigación. Retrieved from https://sottovoce.hypotheses.org/1656

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(3), 715-734. doi:10.1017/S0954579405050340

Punset, E. (2011). Redes nº 105: Música, emociones y neurociencia. () RTVE. Retrieved from https://www.youtube.com/watch?v=wl0ep6vdWaQ

Rahman, M. S., Barnes, K. A., Crommett, L. E., Tommerdahl, M., & Yau, J. M. (2020). Auditory and tactile frequency representations are co-embedded in modality-defined cortical sensory systems. NeuroImage, 215, 116837. doi:10.1016/j.neuroimage.2020.116837

Revuelta, P., Ortiz, T., Lucía, M. J., Ruiz, B., & Sánchez-Pena, J. M. (2020). Limitations of standard accessible captioning of sounds and music for deaf and hard of hearing people: An EEG study. Frontiers in Integrative Neuroscience, 14, 1. doi:10.3389/fnint.2020.00001

Roberts, L. (2020). Understanding the mel spectrogram. Retrieved from https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Rodríguez, C. (2021). Accesibilidad audiovisual: El subtitulado para sordos. Retrieved from https://tatutrad.net/accesibilidad-en-el-contenido-audiovisual-el-subtitulado-para-sordos/

Romero-Fresco, P., & Eugeni, C. (2020). Live subtitling through respeaking. In Ł Bogucki, & M. Deckert (Eds.), The palgrave handbook of audiovisual translation and media accessibility () Palgrave Macmillan, Cham.

Rosebrock, A. (2018). Keras Conv2D and convolutional layers. Retrieved from https://www.pyimagesearch.com/2018/12/31/keras-conv2d-and-convolutional-layers/

Rovan, J., & Hayward, V. (2000). Typology of tactile sounds and their synthesis in gesture-driven computer music performance. In M. Wanderley, & M. Battier (Eds.), Trends in gestural control of music (pp. 297-320) IRCAM.

Russo, F. A., Ammirante, P., & Fels, D. I. (2012). Vibrotactile discrimination of musical timbre. Journal of Experimental Psychology: Human Perception and Performance, 38(4), 822-826. doi:10.1037/a0029046

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience, 14(2), 257-262. doi:10.1038/nn.2726

Sanchez-Lopez, J., Silva-Pereyra, J., & Fernandez, T. (2016). Sustained attention in skilled and novice martial arts athletes: A study of event-related potentials and current sources. PeerJ (San Francisco, CA), 4, e1614. doi:10.7717/peerj.1614

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. NeuroImage, 28(4), 848-858. doi:10.1016/j.neuroimage.2005.06.023

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology, 32(1), 76-92. doi:10.1177/0022022101032001009

Scherer, K. R., & Zentner, M. R. (2001). Emotional effects of music: Production rules. Music and emotion: Theory and research (pp. 361-392). New York, NY, US: Oxford University Press.

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. European Journal of Personality, 14(4), 325-345. doi:10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I

Schupp, H., Cuthbert, B., Bradley, M., Hillman, C., Hamm, A., & Lang, P. (2004). Brain processes in emotional perception: Motivated attention. Cognition and Emotion, 18(5), 593-611. doi:10.1080/02699930341000239

Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., & Hari, R. (2006). Touch activates human auditory cortex. NeuroImage, 30(4), 1325-1331. doi:10.1016/j.neuroimage.2005.11.020

Schutz, M. (2017). Acoustic constraints and musical consequences: Exploring composers' use of cues for musical emotion. Frontiers in Psychology, 8, 1402. doi:10.3389/fpsyg.2017.01402

Senabre, M. (2018). Breve historia de los sistemas de afinación. Retrieved from https://musicologiaempirica.wordpress.com/2018/05/28/breve-historia-de-los-sistemas-de-afinacion/

Seo, Y., & Huh, J. (2019). Automatic emotion-based music classification for supporting intelligent IoT applications. Electronics, 8(2), 164. doi:10.3390/electronics8020164

Sermanet, P., & LeCun, Y. (2011). Traffic sign recognition with multi-scale convolutional networks. Paper presented at the The 2011 International Joint Conference on Neural Networks, 2809-2813. doi:10.1109/IJCNN.2011.6033589

Sherman, S., & Guillery, R. (2001). Exploring the thalamus Academic Press. doi:10.1016/B978-0-12-305460-9.X5013-8

Shibasaki, H., Barrett, G., Halliday, E., & Halliday, A. M. (1980). Components of the movement-related cortical potential and their scalp topography. Electroencephalography and Clinical Neurophysiology, 49(3), 213-226. doi:10.1016/0013-4694(80)90216-3

Skipper, L. M., Ross, L. A., & Olson, I. R. (2011). Sensory and semantic category subdivisions within the anterior temporal lobes. Neuropsychologia, 49(12), 3419-3429. doi:10.1016/j.neuropsychologia.2011.07.033

Sojo, D., & Clavijo, F. (2020). La escala diatónica. Retrieved from https://aulamatematica.neocities.org/musicaymatematicas/la_escala_diatnica.html

Sorana. (2020). A short intuitive explanation of convolutional recurrent neural networks. Retrieved from https://www.analyticsvidhya.com/blog/2020/11/a-short-intuitive-explanation-of-convolutional-recurrent-neural-networks/

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. Paper presented at the 12th International Society for Music Information Retrieval Conference (ISMIR 2011),

Sprengelmeyer, R., Rausch, M., Eysel, U. T., & Przuntek, H. (1998). Neural structures associated with recognition of facial expressions of basic emotions. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265(1409), 1927-1931. doi:10.1098/rspb.1998.0522

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Hömberg, V., . . . Rowland, D. (1996). Loss of disgust: Perception of faces and emotions in huntington's disease. Brain, 119(5), 1647-1665. doi:10.1093/brain/119.5.1647

Stanford University. (2021). Stanford university CS231n: Convolutional neural networks for visual recognition. Retrieved from http://cs231n.stanford.edu/

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America, 8, 185-190. doi:10.1121/1.1915893

Sturm, B. L. (2013). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv:1306.1461, doi:10.1080/09298215.2014.894533

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. Psychological Bulletin, 110(1), 67-85. doi:10.1037/0033-2909.110.1.67

Teie, D. (2016). A comparative analysis of the universal elements of music and the fetal environment. Frontiers in Psychology, 7 doi:10.3389/fpsyg.2016.01158

Thayer, R. E. (1989). The biopsychology of mood and arousal. New York, NY, US: Oxford University Press.

Thomas, L., & LaBar, K. (2005). Emotional arousal enhances word repetition priming. Cognition and Emotion, 19(7), 1027-1047. doi:10.1080/02699930500172440

Thompson, W. F., Russo, F. A., & Sinclair, D. (1994). Effects of underscoring on the perception of closure in filmed events. Psychomusicology: A Journal of Research in Music Cognition, 13(1-2), 9-27. doi:10.1037/h0094103

Trainor, L. J., Tsang, C. D., & Cheung, V. H. W. (2002). Preference for sensory consonance in 2- and 4-month-old infants. Music Perception, 20(2), 187-194. doi:10.1525/mp.2002.20.2.187

Tranchant, P., Shiell, M. M., Giordano, M., Nadeau, A., Peretz, I., & Zatorre, R. J. (2017). Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. Frontiers in Neuroscience, 11 doi:10.3389/fnins.2017.00507

Trujillo-Barreto, N. J., Aubert-Vázquez, E., & Valdés-Sosa, P. A. (2004). Bayesian model averaging in EEG/MEG imaging. NeuroImage, 21(4), 1300-1319. doi:10.1016/j.neuroimage.2003.11.008

Ujjwalkarn. (2016). An intuitive explanation of convolutional neural networks . Retrieved from https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Vieillard, S., & Gilet, A. (2013). Age-related differences in affective responses to and memory for emotions conveyed by music: A cross-sectional study. Frontiers in Psychology, 4, 711. doi:10.3389/fpsyg.2013.00711

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. Cognition and Emotion, 22(4), 720-752. doi:10.1080/02699930701503567

Wanderley, M. M., Battier, M. & Arfib, D. (2000). Trends in gestural control of music.

Wang, J., Lo, H., Jeng, S., & Wang, H. (2010). Audio classification using semantic transformation and classifier ensemble. Paper presented at the Mirex 2010,

Wang, T. (2015). A hypothesis on the biological origins and social evolution of music and dance. Frontiers in Neuroscience, 9 doi:10.3389/fnins.2015.00030

Wang, Y., Jiang, M., Huang, Y., Sheng, Z., Huang, X., Lin, W., . . . Lv, B. (2020). Physiological and psychological effects of watching videos of different durations showing urban bamboo forests with varied structures. International Journal of Environmental Research and Public Health, 17(10) doi:10.3390/ijerph17103434

Wei, Y., Wu, Y., & Tudor, J. (2017). A real-time wearable emotion detection headband based on EEG measurement. Sensors and Actuators A: Physical, 263, 614-621. doi:10.1016/j.sna.2017.07.012

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., & Ackermann, H. (2004). Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. Cerebral Cortex, 14(12), 1384-1389. doi:10.1093/cercor/bhh099

Wissmath, B., Weibel, D., & Groner, R. (2009). Dubbing or subtitling? effects on spatial presence, transportation, flow, and enjoyment. Journal of Media Psychology: Theories, Methods, and Applications, 21(3), 114-125. doi:10.1027/1864-1105.21.3.114

Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020). Evaluation of CNN-based automatic music tagging models. arXiv>Electrical Engineering and Systems Science, Retrieved from arXiv:2006.00751

Wyse, L. Nanayakkara, S. Seekings, P. Ong, S. H. Taylor, E. A. (2012). Palm-area sensitivity to vibrotactile stimuli
above 1 kHz. Paper presented at the The International Conference on New Interfaces for Musical Expression 2012,

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (June 2008). What is the best segment duration for music mood analysis? Paper presented at the 17-24. doi:10.1109/CBMI.2008.4564922

Yang, X., Dong, Y., & Li, J. (2018). Review of data features-based music emotion recognition methods. Multimedia Systems, 24(4), 365-389. doi:10.1007/s00530-017-0559-4

Yang, Y., & Chen, H. H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology Vol. 3, no. 3, 3(3), 1-30. doi:10.1145/2168752.2168754

Young, G. W., Murphy, D., & Weeter, J. (2015). Vibrotactile discrimination of pure and complex waveforms. Paper presented at the 12th Sound and Music Computing Conference, 359-362. doi:10.13140/RG.2.1.3205.3608

Zald, D. H., & Pardo, J. V. (2002). The neural correlates of aversive auditory stimulation. NeuroImage, 16(3, Part A), 746-753. doi:10.1006/nimg.2002.1115

Zhang, M., Ge, Y., Kang, C., Guo, T., & Peng, D. (2018). ERP evidence for the contribution of meaning complexity underlying emotional word processing. Journal of Neurolinguistics, 45, 110-118. doi:10.1016/j.jneuroling.2016.07.002

Zhang, W., Lei, W., Xu, X., & Xing, X. (2016). Improved music genre classification with convolutional neural networks. Paper presented at the Interspeech 2016,