UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Programa de Maestría y Doctorado en Música

Facultad de Música
Instituto de Ciencias Aplicadas y Tecnología
Instituto de Investigaciones Antropológicas

Hacia una escucha automática de la espontaneidad: relaciones complejas entre la libre improvisación, la escucha y los sistemas de aprendizaje automático

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN MÚSICA (Tecnología Musical)

PRESENTA
Lic. Aarón Arturo Escobar Castañeda

TUTOR PRINCIPAL
Dr. Hugo Solís García
(Universidad Autónoma Metropolitana Unidad Lerma)

Ciudad de México, junio 2018

Quiero dar mis mas sinceros agradecimientos:

A Hugo Solís, Roberto Kolb, Fernando Nava, la coordinación del posgrado en música de la UNAM, a mis profesores.

Especialmente a Rossana Lara, Cinthya García, Felipe Orduña, Carles Tardio, Iván Paz.

A David, Diego, Nonis, Eduardo, Nicolás.

A Jorge David, Marcela.

A Luanne, Pilar, Arturo.

A Lupita Velázquez, Felipe Lara.

A Galo, Gabriel, Patricio, Diego, Homero.

A José Luis, Diego, Nefi, Fabián, Héctor, Lucila, Gil, Emmanuel, Ross, Luis, Rolando, Miguel, Jorge, Homero, Rafael, Toño, Andy.

A la escena de improvisadores libres de la ciudad de México.

A Wade Matthews, Chefa Alonso, Okkyung Lee, Clare Cooper, Eddie Prévost, Derek Bailey, Tetuzi Akiyama, Fernando Vigueraz, Remi Alvarez, Juan Pablo Villa, Wilfrido Terrazas, Eli Kesler, Nicolas Collins, Yan Leguay, Mike Majkowski, Maja Ratkje, Nick Collins, Otomo Yoshihide.

Índice general

Algunos antecedentes sobre la investigación

Desde hace varios años me ha parecido importante explorar mis capacidades como compositor e improvisador, más allá del campo técnico-teórico de la música estudiado durante el periodo de la licenciatura en Composición. El gran aporte que encontré fue durante el servicio social que realicé en el ICAT (Instituto de Ciencias Aplicadas y Tecnología, UNAM) el cual me permitió conocer una perspectiva epistemológica distinta: las ciencias de la complejidad. El estudio de este enfoque me llevó a plantear un macro-sistema conformado por múltiples agentes interactuando entre sí: músicos/improvisadores, público, espacio acústico y un sistema sonoro interactivo; de manera tal que la retroalimentación de todos sus elementos provocara la emergencia de texturas que colocan al escucha en un voraz y continuo flujo de intercambio energético. Al seguir atentamente las interacciones del sistema, me percaté de algunas situaciones de monotonía. Para poder romperla fue necesaria la intervención de un agente externo (yo, en este caso) que modificara algunos de los parámetros del sistema en tiempo real. Estas modificaciones podían cambiar el nivel de saturación, el grado de densidad, el volumen e incluso la tímbrica de las

salidas producidas por el sistema al interactuar con un ensamble de músicos. Dicha situación me llevó a plantear en la maestría la idea de programar una máquina que fuera capaz de autorregular sus propiedades y, a través de ello, reaccionar de forma inteligente a los estímulos sonoros percibidos en un momento dado. Fue justo aquí donde decidí desarrollar un sistema autónomo que pudiera interactuar con otros improvisadores a partir de una escucha y modos de reacción basados en inteligencia artificial.

Introducción

La presente investigación aborda, desde una estructura rizomática, sin jerarquías epistémicas balanceadas, la relación humano-máquina en la generación de modelos de escucha basados en tecnologías de aprendizaje automático. Específicamente, la creación de un sistema de escucha artificial aplicado al análisis de la improvisación libre. Lo anterior involucra, por un lado, una revisión de distintos proyectos similares además de una indagación profunda sobre descriptores de señales de audio para interpretar algunas de las cualidades acústicas que considero relevantes en dicha práctica artística; por otro, un rastreo para determinar cuál de los distintos algoritmos de aprendizaje de máquinas puede ser el más adecuado para analizar las relaciones sonoras producidas en la libre improvisación.[1] A través de un

[1] Al referirme a la libre improvisación o improvisación libre tengo presente las tendencias y prácticas estilísticas desarrolladas a mediados de los años sesenta en Europa y Estados Unidos, inspiradas en la música electroacústica, el free jazz y la música académica contemporánea. En la improvisación libre se prioriza la experimentación con instrumentos u objetos encontrados, el ruidismo, el uso de sonidos tenidos (*drones*, pedales), las texturas (sean densas o ligeras), el empleo prolongado o deliberado de silencios así como las interacciones descentralizadas entre los involucrados. Estas van mucho más allá de la jerarquía establecida entre el compositor y los intérpretes, además, se vincula con la práctica de la escucha

análisis basado en la detección y clasificación de componentes en las señales de audio, es posible estudiar estructuras y formas arquetípicas en las que distintos autores de diferentes latitudes se han aproximado a la improvisación libre, posibilitando la generación de modelos computacionales para describir algunas cualidades comunes de dicha práctica.[2] Asimismo, la tesis analiza qué sucede en el contexto de la improvisación libre, entendida como un sistema artístico que, aunque estilísticamente acotado –al igual que el jazz, el rock, el blues o cualquier género musical, todavía sigue abriendo nuevas perspectivas para aproximarse a la creación de música y la generación sonora, partiendo de formas de escucha más atentas, descentralizadas y conscientes hacia los fenómenos sonoros. Cabe mencionar que ésta es una investigación de largo aliento y se espera que los modelos de improvisación generados sean útiles en un futuro para aplicarse a un sistema sonoro interactivo (pendiente dentro de los alcances de esta investigación), en el que la máquina sea capaz de generar una estética derivada de la comunicación automática con improvisadores libres en contextos específicos.

Por otro lado, realizar un análisis de la improvisación libre basado en el aprendizaje y la escucha automática, implica necesariamente el uso de herramientas tecnológicas que históricamente y en las prácticas

atenta y una casi escasa o nula planeación sobre lo que se va a interpretar. En el capítulo 4 se desarrolla el tema de la libre improvisación y se profundiza sobre algunas discusiones relacionadas con la noción de libertad en esta práctica.

[2]Debido a la cercanía que tengo con esta práctica desde hace varios años, tengo un interés particular por analizar la improvisación libre frente a otras formas de improvisación u otras músicas, por otro lado me interesa posibilitar la generación de modelos arquetípicos que describan las formas en las que se hace la improvisación libre y de este modo problematizar la noción de libertad y sus implicaciones dentro de la práctica.

actuales se enmarcan en el ámbito de la guerra, la vigilancia y el control. En este sentido cabe preguntarse si es posible a través del arte desmarcararse de las lógicas originarias de la tecnología implicada en muchos proyectos similares, o más bien estamos legitimándolas inconsciente e irremediablemente al utilizarlas.

Las aproximaciones actuales basadas en la detección y clasificación de componentes en las señales de audio, sean altura, timbre, amplitud, brillo, centro espectral, cromagrama, tienden a limitarse a explorar muy pocas áreas de la vasta esfera musical y sonora existente. Pareciera haber una resistencia a entablar un diálogo con otros campos estéticos como el arte sonoro, el paisaje sonoro, la música académica contemporánea, el free jazz o la libre improvisación.[3] Lo que sí hay son varios grupos de trabajo y diversas compañías (como GoogleBrain, Shazam, Spotify, Facebook, Defense Innovation Marketplace, entre otras) que están interesados en generar y perfeccionar sistemas inteligentes capaces de reconocer momentos sonoros importantes con fines comerciales, militares, de seguridad, de vigilancia, control y asistencia.[4] Un ejemplo de particular interés es la creación de sistemas de asistencia al compositor basados en la identificación de elementos del lenguaje de músicas tonales, estos son aplicados a la generación, casi automática, de nueva música acotada a los estándares comerciales. Estos sistemas son alimentados por corpus gigantescos de diversos géneros musicales para crear variaciones dentro de un estilo específico, con duraciones, formas, instrumentaciones y temáticas de-

[3] https://bit.ly/2JuN1lq, Fecha de consulta 9 de junio 2018

[4] Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard. Machine listening techniques as a complement to video image analysis in forensics. *IEEE International Conference on Image Processing*, 2016,

https://engt.co/2rqtSb0, Fecha de consulta 28 de enero 2017.

terminadas. Con todo el potencial tecnológico actual estas iniciativas dejan de preocuparse por crear algo nuevo, que revolucione el estancamiento musical actual en que vivimos, y más bien se interesan en hacer alusión a estilos musicales completamente consolidados; optan por conservar los gustos ya probados de los públicos, ajustándose a los estándares enajenantes de las prácticas musicales industriales.[5]

Indudablemente, los temas del aprendizaje y la escucha automática relacionados con la inteligencia artificial poseen un alto contenido ideológico, político y económico. Comúnmente, éstas tecnologías son entendidas en favor del "progreso humano"; sin embargo, pocas veces dentro de la práctica (artística o de desarrollo tecnológico) se cuestiona su proliferación enmarcada en los ámbitos de lógicas empresariales.[6] De ahí, se han generado sistemas oligopólicos de la conectividad informacional (Facebook, Twiter, Apple, y diversas compañias de telefonia movil) que atienden la estructura social en red a través de dispositivos de vigilancia, todos ellos capacitados para extraer (segmentar), analizar (describir numéricamente), clasificar y producir la generalización de modelos descriptivos, destinados a las nuevas formas de seccionar por clases las diversas subjetividades, patrones de consumo y hábitos sociales para la implementación de nuevas estrategias de dominación y control.

[5] https://magenta.tensorflow.org/blog/, http://www.flow-machines.com/leadsheetgeneration/, Fecha de consulta 9 de junio 2018
http://www.flow-machines.com/flowcomposer-composing-with-ai/, Fecha de consulta 9 de junio 2018
https://bit.ly/2cTSWQo, Fecha de consulta 9 de junio 2018

[6] Juan Martín Prada. *Prácticas artísticas e Internet en la época de la redes sociales.* AKAL, 2015 p. 25.

Desde este marco surge la necesidad por analizar las implicaciones socio-culturales de estos temas, además de intentar contrastar y cuestionar desde la práctica artística y específicamente desde la improvisación libre las prácticas y utilidades originarias del aprendizaje y escucha automática. "De lo que se trata, pues, es de ser algo más conscientes del fascinante y omnicomprensivo capitalismo basado en la conectividad, abriendo si es posible, el camino a formas de vida en red alternativas, menos parasitadas por los intereses del nuevo capitalismo informacional".[7] Además, la aproximación aquí planteada aporta nuevas perspectivas al campo conocido como Music Information Retrival, generalmente concentrado en analizar músicas tonales, con estructuras armónicas claras, y con tendencia a la estabilidad rítmica de la música, aunque también hay proyectos que se focalizan en el estudio de músicas de tradiciones diferentes a las europeas.[8] A diferencia de estas perspectivas, la investigación plantea tratar con un fenómeno poco estudiado por estos campos debido a la liminalidad aún presente en la libre improvisación pese a ser un fenómeno musical que surgió en los años sesenta.[9] "La investigación en humanidades

[7] Idem p. 27.

[8] http://compmusic.upf.edu/

[9] Cabe mencionar que en la historia de la música han existido muchas formas de improvisación tanto en la cultura occidental como en otras culturas, en general mediadas por progresiones armónicas definidas, exploraciones en escalas delimitadas, por ejemplo la pentatónica, ritmos con tendencia a la estabilidad o formas conscientemente delimitadas. Al hablar de improvisación libre no me refiero a estas otras tradiciones ligadas a la improvisación, sino más bien a la tradición europea y estadounidense de la libre improvisación representada por agrupaciones como Spontaneous Music Ensemble, The Music Improvisation Company, The Art Ensemble of Chicago y AMM o a solistas como Evan Parker, Derek Bailey, Anthony Braxton, Peter Brötzmann, John Zorn, George Lewis, Henry Kaiser o Fred

puede proporcionar el análisis necesario sobre el papel de los factores subjetivos y los contextos sociales y culturales en los que funcionarán las aplicaciones tecnológicas. El conocimiento de estos factores debe ser incorporado a los sistemas de recuperación de música y sistemas de música interactivos. [...] [Además], las humanidades proveen una rica fuente de teorías, conceptos y tradiciones que son altamente relevantes e inspiradores para nuevos estudios empíricos y aplicaciones tecnológicas".[10]

El sistema de improvisación automática propuesto se divide en seis fases que incluyen los estados de interacción que he podido detectar a partir de mi actividad como improvisador, y en discusión con el grupo de improvisación *Ruido 13* con el que actualmente toco. Cada uno de estos estados posibles en la improvisación mantiene un equilibrio homeostático[11] y dan coherencia a las interacciones generadas entre varios músicos. Estos estados son: escuchar, imitar, proponer, acompañar, romper y *solear*). Por cuestiones del tiempo que estipula

Frith. Para mayor información y contexto respecto a otras tradiciones de improvisación recomiendo consultar los libros de Derek Bailey *Improvisation. Its Nature and Practice in Music*, Wade Matthews *Improvisando: la libre creación musical*, o la tesis de Licenciatura en composición en la UNAM de Hugo Nefi Herrada titulada *La libre improvisación y la creación musical: memoria e imaginación como generadores de la creatividad sonora.*

[10]Marc Leman, Federico Avanzini, Alain de Cheveigné, and Emmanuel Bigand. The societal contexts for sound and music computing: Research, education, industry, and socio-culture. *Journal of New Music Research*, 36(3):149–167, 2007 p, 152.

[11]La homeostasis es el estado estable de un sistema, incluye el automantenimiento, la autoregulación y el equilibrio; cada variable del sistema está controlada por mecanismos reguladores que en conjunto garantizan el correcto funcionamiento del sistema.

la maestría y debido a la complejidad en la realización del sistema, en esta investigación solo se concretó la primera de ellas, la escucha.

El estado de escucha es el primer paso para la generación del sistema de improvisación automática por varias razones. La escucha es el acto de atender a uno o varios objetos sonoros específicos que serán interpretados de formas diversas de acuerdo con la información almacenada por el agente (humano) o sistema digital. Una escucha atenta involucra su relación con el presente, activa la memoria de los objetos sonoros pasados y ambas contribuyen a la realización momentánea de una consciencia activa de escucha e interpretación; y en algunos casos, involucra una planeación subjetiva, vaga o concreta de lo que podría suceder en el futuro.[12]

Para continuar planteo las siguientes preguntas de investigación: ¿cómo crear un sistema de escucha capaz de analizar la improvisación libre y qué posibilita esta aproximación respecto al campo de las aplicaciones estandarizadas de escucha y aprendizaje automático? Esta pregunta central se relaciona con otras que se convirtieron en guías de la investigación: ¿qué se entiende por el acto de "reconocer" en términos de una máquina? ¿Qué implica crear, en términos de desarrollo tecnológico, un sistema basado en el reconocimiento de patrones audibles? ¿Cuál es el contexto originario y actual de esta tecnología?

Mis investigaciones en torno a la improvisación libre proponen responder: ¿cómo ocurren las interacciones en esta práctica, cuáles son los recursos que se utilizan? Interrogando su supuesta libertad, ¿qué es lo que la improvisación libre enmarca y posibilita para generar la estética característica que la define? Además, ¿qué supone para

[12]Pierre Schaeffer. *Tratado de los objetos musicales.* Alianza música, 1996 pp. 63, 64.

la práctica de la improvisación libre el desarrollo de un sistema que analiza modelos arquetípicos de este estilo y qué aporta el análisis de escucha a dicha práctica?

Para realizar esta investigación se planteó una tesis en la que interactuaran diversos tonos: el técnico, el crítico, el reflexivo, el práctico, el anecdótico y el descriptivo. Alrededor de los cuatro capítulos que comprende esta tesis, se encontrarán unos tonos más presentes que otros debido a la complejidad y especialización que ciertos temas requieren. También se propuso que cada capítulo fuera consistente y autosuficiente por si mismo de manera que no requiriera de los otros para ser comprensible.

El primer capítulo aborda los antecedentes de las máquinas que escuchan,[13] particularmente las discusiones relacionadas con los intereses de la civilización occidental por emular la percepción humana de la realidad a través de los autómatas y posteriormente las máquinas digitales. Además se abordan las implicaciones sobre el uso de tecnologías de aprendizaje y escucha automática donde se discuten algunos modos de producción actuales basados en este paradigma así como algunas reflexiones en torno al uso y aplicaciones de estas tecnologías dentro de los campos de la recuperación de información musical, el análisis y la recopilación de datos destinados a la vigilancia, el control y la guerra. Cabe señalar que a través de esta aportación busqué desmarcarme de hacer una investigación basada en una posición afirmativa y pragmática sobre la ciencia y la tecnología implicadas en el desarrollo de mi sistema y proyectos similares con salidas artísti-

[13]Llamadas también en esta tesis de manera indistinta escucha de máquinas o escucha automática, términos proveniente del inglés que también suelen aparecer como machine listening, listening machine o computer audition

cas. No quise hacer una tesis puramente técnica ya que eso es lo más que he encontrado dentro del campo de Music Information Retrival y en la literatura sobre procesamiento de señales digitales de audio, además de que, admito, no tener las herramientas suficientes ni el estudio formal desde la ciencia para hacerlo. Mi tesis en ese sentido, está planteada desde la investigación en términos artísticos. Considero que en la literatura académica hay un enorme vacío y una falta de argumentos críticos que cuestionen las tecnologías implicadas en este tipo de prácticas, y por lo general carecen de un posicionamiento por parte del investigador, que considero fundamental dada la intersección del ámbito tecnológico con el campo del arte. Por ello, veo sumamente necesario abogar por un estudio mucho más crítico que promueva el pensamiento humanista y cuestione las formas tradicionales y aparentemente neutrales en las que nos aproximamos al cruce entre el arte, la ciencia y la tecnología. En ese sentido, las auto-reflexiones planteadas en el primer capítulo y a lo largo de la tesis, son un aporte que considero sumamente valioso en mi tesis, de gran pertinencia para el área de tecnología musical y de enorme relevancia para el momento histórico en que vivimos.

El capítulo dos presenta cinco proyectos que emplean el aprendizaje y la escucha automática para la creación de sistemas interactivos en contextos improvisatorios. Al mismo tiempo analiza las aproximaciones metodológicas y estrategias seguidas para el desarrollo de estos proyectos. Estas fueron de gran utilidad para el desarrollo de la máquina que escucha presentada en esta tesis ya que son antecedentes directos a la creación de sistemas de reconocimiento de momentos sonoros con un enfoque hacía la improvisación. Junto con ello se genera una reflexión sobre la creación de sistemas interactivos para la impro-

visación, considerando las adecuaciones que los músicos realizan a su práctica a partir de la interacción con estos sistemas, al atender sus alcances y limitaciones.

El capítulo tres aborda ampliamente desde una perspectiva más técnica las definiciones, conceptos, métodos y herramientas de la escucha automática, incluye los descriptores de audio utilizados para el reconocimiento de elementos sonoros en la libre improvisación, tales como MFCCs, contraste espectral, centoide espectral y *onsets*; asimismo, se ofrece una mirada sobre el aprendizaje automático y algunas definiciones sobre los conceptos de aprendizaje supervisado y no supervisado así como los algoritmos de clasificación Multilayer Perceptron (Percepción multicapa o redes neuronales artificiales), DL4MlpClassifier (Aprendizaje Profundo) y K-Means.

Finalmente el cuarto capítulo aborda algunas nociones y discusiones sobre qué es lo libre en la improvisación libre, a través de quiénes, dónde y cómo se desarrolla esta práctica. Estos planteamientos permiten esbozar algunas tendencias estilísticas en la improvisación libre que son una guía para la delimitación del sistema de escucha de improvisaciones libres debido a que posibilita su identificación y segmentación, primero desde mi escucha, luego desde la escucha de la máquina. Esta escucha de la máquina incluye dos metodologías que permitieron dar cuenta de las posibilidades que distintos descriptores de audio, así como distintos algoritmos de clasificación, pueden tener para la clasificación tímbrica de distintas improvisaciones solistas. La primera metodología usó el aprendizaje supervisado, en ésta se empleó como herramienta de extracción de características de audio la librería SCMIR y SuperColllider, para el aprendizaje de máquinas y la clasificación de estos datos se usó Weka; asimismo fueron empleados los

algoritmos de clasificación de percepción multicapa o redes neuronales artificiales, Aprendizaje Profundo y K-Means. En la segunda metodología se creó un sistema basado en aprendizaje sin supervisión para la clasificación de elementos tímbricos de la improvisación libre con el lenguaje de programación Python. En ésta se empleó un segmentador automático que analiza los momentos de inicio de fragmentos de las improvisaciones y genera pequeños archivos, después estos fragmentos son analizados con la librería *Librosa* obteniendo varios descriptores de audio. Entre los descriptores disponibles en esta librería, dada su amplia variedad de aplicaciones actualmente, se tiene la hipótesis de que el descriptor MFCCs puede ser lo suficientemente robusto para generar clasificaciones de improvisaciones libres congruentes con mi escucha. Una vez analizados estos fragmentos de audio se procedió a analizarlos con el algoritmo K-Means empleado con la librería TensorFlow. Después se creó un modelo de evaluación auditivo, su utilidad se encuentra en la posibilidad de analizar qué tanto el sistema está siendo congruente con lo que estoy esperando escuchar dentro de las clasificaciones de improvisaciones libres. Finalmente se obtuvo un modelo visual divido por clases de timbres, amplitudes o densidades que despliega en forma temporal y da cuenta de cómo fueron empleados los materiales en una improvisación libre, generando una descripción de perfiles arquetípicos de la misma. Estas metodologías sirvieron para comparar las diferencias entre los dos tipos de aproximaciones al aprendizaje de máquinas: el aprendizaje supervisado y el no-supervisado. Se concluyó que el aprendizaje supervisado, especialmente en la percepción multicapa, obtiene altos índices de certeza en la clasificación, en contraposición con el aprendizaje basado en el algoritmo K-Means el cual baja hasta un 20 % en relación con el anterior.

Se decidió trabajar en la segunda aproximación debido a que en la aproximación con Weka no fue posible obtener resultados sonoros para comparar de forma auditiva cómo opera el sistema de clasificación. Una vez explicadas las formas de aplicar estas técnicas se realizaron 2 experimentos que dan cuenta de las posibilidades y limitantes de estas dos metodologías. Para realizar estos experimentos se crearon 2 bases de datos de improvisadores solistas, que fueron dispuestas para su análisis. Estos experimentos demuestran: 1. Que el aprendizaje supervisado es más consistente que el no supervisado. 2. Que los descriptores MFCCs son más que suficientes para modelar la escucha de la máquina y aproximarse a una clasificación de improvisaciones libres basadas en mi escucha.

Capítulo 1

Marco histórico-cultural

1.1. Introducción

En esta sección se plantea rastrear el origen de las máquinas que escuchan destacando las primeras incursiones en la creación de máquinas capaces de producir sonidos, y mediadas por una escucha y observación humanas. Por medio de la hibridación humano-máquina se han logrado grandes avances técnicos y científicos: al jugar con la mera ilusión de la reproductibilidad del mundo a través de sistemas mecánicos, y al "prestarle" nuestros oídos a la máquina, ésta sería capaz de reproducir o imitar fragmentos de realidad que jamás podría escuchar, recurriendo únicamente a una "memoria mecánica". Esta memoria mecánica mediada por el oído humano —el cual es entendido también como un mecanismo que puede ser usado de manera instrumental para una variedad de fines, y además como "la fuente y objeto de la reproducción sonora"—,[1] sería capaz de reproducir o

[1] Jonathan Sterne. *The Audible Past: Cultural Origins of Sound Reproduction.* Duke University Press, 2003. pp. 33, 34.

imitar sonidos humanos o animales, ruidos o música; extractos de la realidad anclados en una nueva. Ésta, construida desde la obsesión humana por imitar el mundo a través del mecanismo, de formas cada vez más fieles y prístinas. Estas aproximaciones volvieron al sonido un objeto de manipulación mecánica y posteriormente digital, de modo que su difusión fuese posible en cualquier momento y lugar.

Estos temas se relacionan con el proyecto de esta investigación en el sentido de que el sistema propuesto tampoco escucha por sí mismo desde un inicio, sino que necesita de la mediación de mi oído, mi escucha e incluso mi inclinación estética para adquirir una suerte de sentido de escucha. Es así que los resultados expuestos en el capitulo 4, que corresponde a mi desarrollo de la máquina que escucha, pasan siempre por esta mediación oído(escucha)-máquina-oído(escucha) creando una retroalimentación entre la información que entra y regresa desde y hacia mi para poder optimizar el sistema de escucha automática.

Para esta investigación sería demasiado exhaustivo trazar todos los momentos de la historia donde este afán fuertemente enraizado en la ilusión y la simulación está presente en la vida humana. Bastará con centrarnos en los desarrollos tecnológicos partiendo del siglo XVI en Europa, especialmente en las máquinas que conservan un vínculo con la emisión sonora y que están envueltas por los hallazgos y observaciones anatómicas. Es importante señalar que esta afición por la simulación del mundo afectó y sigue afectando notablemente diferentes campos como las ciencias, la filosofía, la literatura, la tecnología, la industria de la guerra, entre otros.

1.2. Antecedentes de las máquinas que escuchan

Un precedente de las máquinas que escuchan son los autómatas que comenzaron a desarrollarse desde tiempos muy remotos, muchos de ellos sin una utilidad más que la de entretener a sus creadores realizando tareas muy sencillas como moverse de un lado a otro o emitir patrones sonoros repetitivos. Autores como Derek J. De Solla Price ubican sus orígenes en las cuevas prehistóricas, [2] donde los humanos con el afán de simular y preservar la realidad, comienzan a fijarla a través de la pintura y la plástica, incluso a simular movimientos de animales, humanos y el mundo anímico que los rodeaba. Prueba de ello son las pinturas de la cueva Chauvet al sur de Francia descubierta en 1994, las cuales Werner Herzog califica —en su documental *La cueva de los sueños perdidos*— como una especie de proto-cinema. Otro ejemplo podría ser la estatua de Memóm construida por Amenhotep la cual era capaz de emitir sonidos cuando la iluminaban los rayos del sol.[3]

Los autómatas fueron diseñados expresamente para imitar funcionamientos humanos y de la naturaleza, desde simples movimientos, escritura, música, sonidos de vocales, sonidos de animales hasta sus respectivos funcionamientos biológicos, como es el caso del pato de Jacques de Vaucanson. Dicho pato era capaz de comer granos, digerirlos y defecarlos. Cada fenómeno, acción o proceso de la vida diaria

[2] Derek J. de Solla Price. Automata and the origins of mechanism and mechanistic philosophy. *Technology and Culture*, 5(1):9–23, 1964. pp. 10.

[3] http://automata.cps.unizar.es/Historia/Webs/automatas_en_la_historia.htm

y experiencia natural podría ser recreada a través de la imitación de la fuente original. En siglo XVII Francis Bacon muestra al mundo su encanto hacia los autómatas con su libro *La Nueva Atlantis*, donde comenta:

> [Hemos encontrado] casas sonoras donde practicamos y demostramos todo el sonido y su generación... Representemos e imitemos todos los sonidos articulados y cartas, y las voces y las notas de las bestias y los pájaros, tenemos ciertas ayudas, que aumentan notablemente el alcance del grandioso oído".[4]

Otro invento sumamente interesante que atañe directamente a la historia del *cyborg* (la mezcla física del humano y la máquina) fueron las prótesis construidas por el médico Ambroise Paré quien fue pionero en las técnicas de cirugía en el campo de batalla. Estas prótesis construidas cerca de 1560, eran capaces de simular miembros artificiales como brazos o piernas para personas que los habían perdido en la guerra.[5] Este ejemplo resulta interesante debido a que es una muestra de cómo las estrategias y técnicas de simulación han sido útiles desde tiempos remotos en situaciones de guerra. Más adelante, René Descartes quien construyó una mujer autómata activada por magnetos llamada Francine, concebía al cuerpo humano como una máquina donde el espíritu o el alma tenían un lugar muy importante ya que lo diferenciaban de otras máquinas, sin el alma el humano sería

[4]Francis Bacon. *Bacon's Advancement of Learning and the New Atlantis*. Lulu Enterprises Incorporated, 2010. p. 294. De aquí en adelante todas las traducciones son mías.

[5]Derek J. de Solla Price. Automata and the origins of mechanism and mechanistic philosophy. *Technology and Culture*, 5(1):9–23, 1964. p. 22

equiparable a un animal o una planta, concebidos por Descartes como máquinas sin alma. El surgimiento de los autómatas fue indispensable para concebir la reproductibilidad técnica del mundo natural y social, el cual apunta al desarrollo de muchas de las tecnologías que actualmente seguimos usando, como por ejemplo, las bocinas, que podrían encontrar su antecedente directo en las máquinas que hablan, construidas entre 1770 y 1790.

Vaucanson, Christian Gottleib Kratzenstein, y Louis de Castel fueron pioneros en simular la voz humana. Partiendo de los estudios anatómicos de la época, fue posible emular la forma en la que las vocales son emitidas por los diferentes órganos involucrados en su producción. Posteriormente, Vaucanson creó intérpretes musicales automáticos que integraban elementos básicos de la ejecución de un instrumento. Para que un autómata lograra tocar melodías en una flauta real tenía que imitar el movimiento de la embocadura, la presión de aire mediante fuelles y el movimiento de los dedos mediante la automatización mecánica.

Un siglo más tarde, otro tipo de máquinas que escuchan comenzaron a surgir, partiendo de experimentos como el fonoautógrafo patentado en 1857 por Éduard-Léon Scott de Martinville. Desde entonces, comenzaban a hacerse evidentes las necesidades por entender el funcionamiento del aparato auditivo humano y más específicamente la forma en la que el sonido llegaba a la membrana auditiva timpánica. Los avances del conocimiento sobre el funcionamiento del oído humano dieron lugar a nuevas apariciones de herramientas tecnológicas capaces de “imitar” de forma mecánica dichas funcionalidades. Léon Scott experimentó creando varios diafragmas hechos de membrana animal y materiales sintéticos; su máquina simulaba el movimiento

de la membrana timpánica así como de los pequeños huesos del oído. Por medio de una boquilla donde se trasmitían las vibraciones de la voz a la membrana, que a su vez conectaba una aguja, era posible realizar una transducción del sonido en un conjunto de trazos impresos en vidrio o papel ahumado. A estos trazos visuales –que eran una representación de las formas de onda propagadas en el aire–, los llamaba grabaciones o fonoautógramas, éstos eran impresos con tal precisión que rápidamente fueron adoptados por la comunidad científica. La visualización del sonido en términos científicos permitía su estudio así como su cuantificación, medición y comparación. "El sonido visual tiene una relación simbiótica con la cuantificación. De acuerdo con las técnicas aceptadas por la ciencia, el sonido tenía que ser visto para ser cuantificado, medido, y grabado; al mismo tiempo, algunas nociones abstractas y cuantificadas del sonido tenían que estar dispuestas de antemano para que su visibilidad tuviera algún significado científico".[6] Más adelante, Graham Bell y Clarence Blake idearon variantes del fonoautógrafo que permitieron tener un conocimiento mayor acerca de los mecanismos funcionales del oído y la escucha. Al emplear literalmente oídos humanos en su construcción, se tenía la idea de estar aproximándose de manera más profunda a la verdadera forma en la que los humanos percibimos el sonido y consecuentemente abrir camino a los estudios de la acústica y psicoacústica del sonido, instrumentos de reproducción y a tecnologías posteriores como el teléfono, fonógrafo, la radio o el micrófono.

El oído humano comenzaba a ser concebido puramente como un artefacto capaz de ser utilizado como el medio y el mecanismo de la

[6] Jonathan Sterne. *The Audible Past: Cultural Origins of Sound Reproduction.* Duke University Press, 2003.p. 45.

Figura 1.1: El humano pensado como máquina que escucha

reproducción sonora.[7] Por sí mismo, el oído es un aparato capaz de traducir vibraciones producidas en el aire y convertirlas en sonido. Desde esta aproximación comienza a entenderse al oído como un aparato timpánico, como una máquina de escucha, capaz de capturar y rastrar cualquier fuente audible.

Esta nueva capacidad de captar, modelar y encapsular el sonido en un artefacto como el fonógrafo de Edison en 1877, o el disco gramófono introducido por Emile Berliner, diez años más tarde, deviene en un nuevo paradigma –que posteriormente se constatará en la música concreta de Pierre Schaeffer– donde el sonido es entendido como un fenómeno, como un objeto de conocimiento que necesariamente tiene que descontextualizarse de la fuente que lo produce para ser

[7] Ibid. p. 42.

analizado puramente desde su auralidad y visualización. Este cambio inevitablemente trae consigo la apertura de los estudios de la acústica moderna y los estudios de la escucha desarrollados a principios del siglo XX.

Ernst Chladni considerado como el padre de la acústica, enmarca la idea de la objetivación del sonido y encuentra necesaria su visualización para poder estudiarlo certeramente en términos científicos, ello a través de lo que hoy se conoce como las figuras de Chladni o cymatics (cimática) —término introducido por Hans Jenny en 1967—, una forma para crear figuras de arena producidas por la excitación de un plato al resonar de acuerdo a las diferentes alturas emitidas por un violín (ejemplos figuras 2.2 y 2.3). Chladni hace una revisión exhaustiva de las teorías del sonido hasta el momento, incluido el cálculo de la velocidad del sonido en diferentes medios de propagación.[8]

Resulta pertinente hablar de la visualización sonora debido a las posibilidades –y limitantes– que permitió dicha aproximación; el desarrollo de las tecnologías sonoras, su estudio y análisis surgieron a partir de este cambio de paradigma, que actualmente permea todas las aplicaciones que usamos para controlar, generar, manipular, analizar, transmitir y reproducir el sonido.

Más adelante, la era de las máquinas del siglo XIX desató inevitablemente procesos de reconfiguración de la escucha a partir de los nuevos sonidos, producto de la tecnología moderna. Los nuevos aspectos culturales incorporan formas científicas y estéticas de escuchar donde la relación del escucha con su ambiente y las circunstancias sociales dictan quién tiene acceso a escuchar determinados conteni-

[8] Ibid., p. 44.

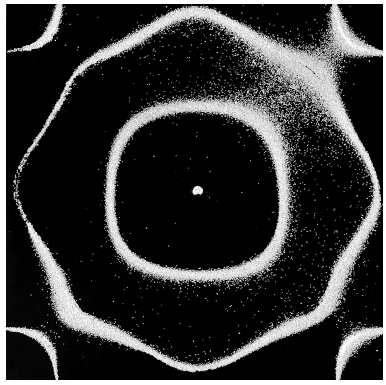

Figura 1.2: Estructura cimática producida con arena y la vibración sonora. Extraído de http://doorofperception.com/2013/11/cymatics/ Fecha de consulta 18 de febrero 2018.

dos. La evolución de la escucha tiene más que ver con la idea de la civilización en constante construcción y progreso, que con la idea del sonido que produce la diversidad del mundo natural. Incluso algunos músicos se negaban a imitar los sonidos de la naturaleza, ya que supuestamente denotaban la falta de creatividad del compositor. Los sonidos escuchados en esa época fueron el resultado de la mediación tecnológica y de la abstracción de la naturaleza del tiempo-espacio, que han sido reconocidos como aspectos de la cultura moderna que impregnan al arte y la ciencia. De acuerdo con Emiliy Thompson,[9] la modernidad ha sido vista y leída en detalle desde múltiples perspectivas pero pocas veces ha sido escuchada. En ese sentido, Douglas Kahn menciona que la cantidad de sonidos nuevos que comenzaron a

[9]Emily Thompson. *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933.* ACLS Humanities E-Book. MIT Press, 2004. p. 10.

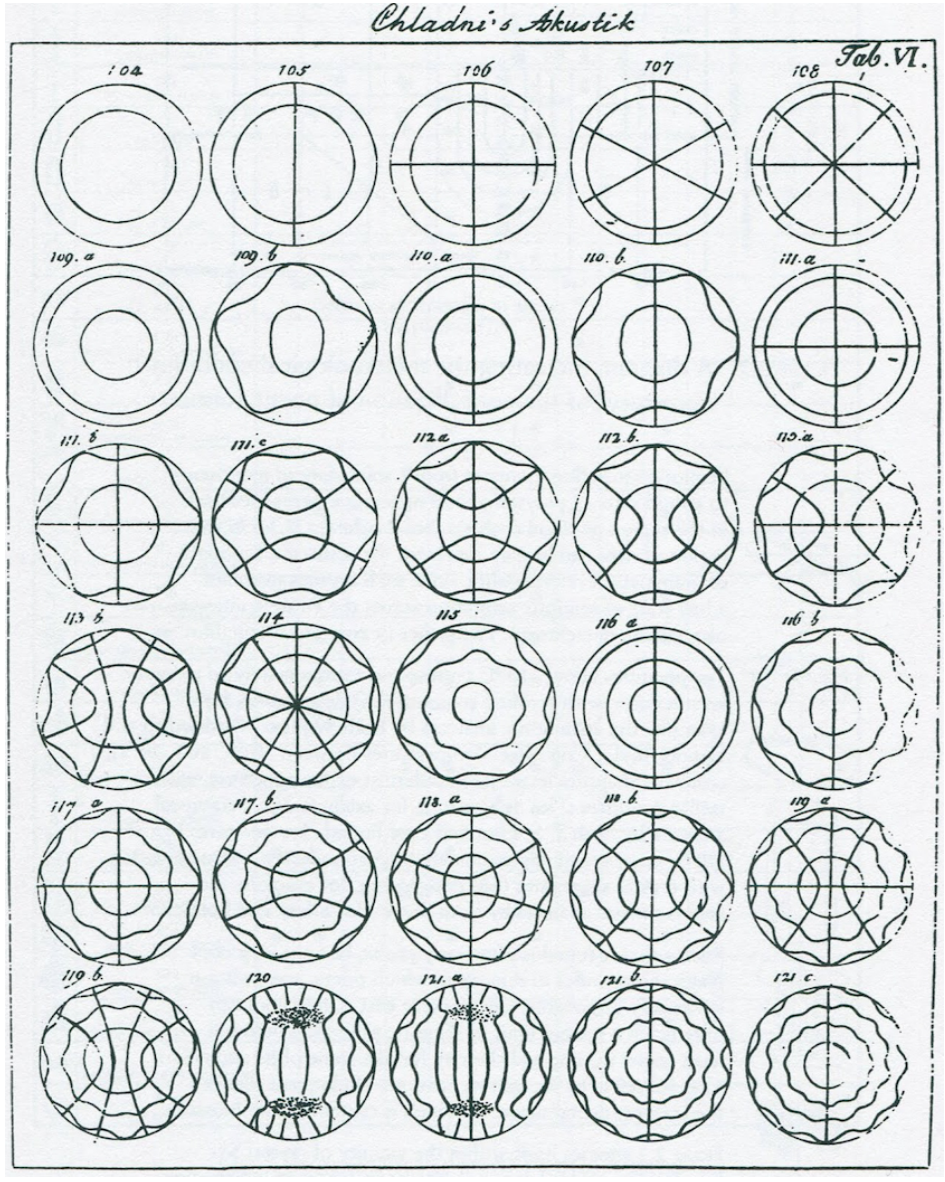

Figura 1.3: Dibujos de Ernst Chladni basados en sus experimentos con la cimática.

existir a partir de la era industrial causaban un énfasis y atención en la escucha como nunca antes se había dado, produciendo una atención aguda hacia estos nuevos sonidos.

Con la finalidad de normalizar y asegurarse de obtener el proceso más "puro" de grabación y reproducción sonora, la ciencia descubre formas para modificar y diseñar espacios arquitectónicos, controlando sus cualidades sonoras para adaptarlas a las necesidades de las salas de concierto de la época. Estos nuevos descubrimientos –que parten de la posibilidad de visualizar el sonido– han estado siempre apegados a las limitaciones de nuestra percepción, lo cual delimita bastante el espectro de posibilidades que las máquinas podrían abarcar, haciendo más simple su estudio. Uno de estos descubrimientos fue la fórmula o ecuación de Wallace Clement Sabine que permite predecir las cualidades acústicas y medir los índices de absorción de diferentes cuartos de acuerdo con diferentes tipos de materiales, es decir, calcular su reverberación. Sabine usó esta formula para medir las refracciones acústicas del Boston's Symphony Hall construido en 1900 y considerado uno de los mejores recintos por sus cualidades acústicas. Estos nuevos enfoques de la acústica dan lugar a la Acoustical Society of America, la cual sigue vigente hasta nuestros días.

La compulsión general por controlar el sonido –e inevitablemente las formas de escucha– a través del desarrollo tecnológico, deviene en la domesticación de la escucha del auditorio, en este caso, en favor de una escucha crítica y exigente.[10] A través de los métodos de grabación y reproducción de la pureza y cristalidad del sonido, se instaura un nuevo criterio para ejercer la elección entre un buen y un mal sonido, entre lo que merece ser escuchado y lo que no. Esta búsqueda

[10]Idem p. 2.

de pureza del sonido proviene en parte de las preocupaciones que el (creciente) ruido exterior produce; así, los sonidos circundantes de las ciudades son silenciados y la eficiencia por eliminar lo innecesario se convierte en una aspiración para las casas productoras de música y otros contenidos sonoros. El sonido, entendido como señal, encarna el ideal de la eficiencia, siendo "extirpado" de los elementos que se consideran innecesarios. En esta reformulación el sonido es paulatinamente disociado del espacio en el que se desenvuelve hasta convertirse en un objeto en sí mismo, lo que posibilita pensar al sonido como señal y posteriormente como dato. Cuando el sonido se convierte en señal se establece el criterio de evaluación, localizado en los medios tecnológicos eléctricos que lo captan, fijan, reproducen y transmiten. Al evaluar la fuerza de la señal contra las intromisiones del ruido eléctrico, es como se juzga un buen sonido de un mal sonido. El sonido moderno fue concebido como un producto de los medios que puede ser también comercializado; para ello había que afinar simultáneamente los oídos de la audiencia a los estándares del mercado. Lo anterior demuestra nuevamente la maestría técnica del humano-máquina frente a otras "bestias-máquinas",[11] el control sobre su ambiente físico y cultural transformando irremediablemente las relaciones entre sonido, escucha, espacio y tiempo, formas de socialidad y consumo.

Otro rasgo importantísimo en las primeras décadas del siglo XX, fue el surgimiento de las vanguardias que comenzaban a entender al arte de formas novedosas, tal es el caso de los futuristas, especialmente Filippo Tommaso Marinetti y Luigi Rusolo, quienes asumen el automatismo como parte del progreso tecnológico, la máquina comenzaba

[11] Derek J. de Solla Price. Automata and the origins of mechanism and mechanistic philosophy. *Technology and Culture*, 5(1):9–23, 1964.

a concebirse como una extensión de la condición humana que inevitablemente la mejora, piensan que desde los desarrollos de las máquinas debe desarrollarse una estética que esté a la altura de las circunstancias de la época. Desde ahí, se propusieron analizar el significado cultural del ruido producido por las máquinas, y demostraron que, a través de la poesía y la construcción de los *intonarumori* (dispositivos acústicos generadores de ruido), las posibilidades que los músicos y los ingenieros podrían tener para crear una nueva cultura alrededor del ruido del mundo moderno eran posibles.[12] A través de esto podríamos decir que el desarrollo de la cultura es más que un fluir de logros tecnológicos y un "porvenir" imparable, la cultura es intrínsecamente inseparable de la tecnología. La tecnología es una fuente de inspiración para crear una nueva cultura alrededor de ella, es un sistema complejo que todo el tiempo se retroalimenta a sí mismo.[13]

Las máquinas que hablan, suenan y tocan son un antecedente de las actuales máquinas que escuchan. Sus respectivos surgimientos se desarrollaron esencialmente de los hallazgos anatómicos tanto en humanos como en otras especies. Estas iniciativas objetivadas en las máquinas se oponen a la efimeralidad del presente, posibilitando el resurgimiento infinito del pasado y la ubicuidad del acontecimiento. Apropiándose de momentos sonoros, cualidades acústicas, físicas y espaciales, es posible reproducir casi cualquier cosa en cualquier momento y lugar. Los autómatas entendidos como mecanismos que simulan la realidad, podrían ser la prueba más fehaciente de que la naturaleza puede ser reproducible desde una perspectiva mecanicista

[12]Luigi Russolo. *The art of noise.* A Great Bear Pamphlet, 1913.

[13]Emily Thompson. *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933.* ACLS Humanities E-Book. MIT Press, 2004. p. 135.

de la vida. El sonido y otros fenómenos físicos, biológicos artísticos y culturales han sido finalmente encapsulados para el control y la manipulación de la realidad, de nosotros mismos, del espacio y el tiempo.

Más adelante, un descubrimiento sumamente interesante para entender lo que ahora son las máquinas que escuchan es la aplicación de la escala de Mel al análisis de señales de audio. Básicamente consiste en hacer un mapeo entre las frecuencias y las alturas percibidas de acuerdo al sistema auditivo humano, que no percibe de manera lineal sino de forma logarítmica.[14] Este mapeo es espaciado en una curva a intervalos fijos de forma lineal en las frecuencias debajo de los mil Hz, y de forma logarítmica en las frecuencias agudas arriba de los mil Hz. El nombre de Mel fue otorgado por Stevens, Volkmann y Neuman en 1937, derivado de la palabra melodía, y sirve para indicar los rangos melódicos en los que el oído puede identificar de manera más certera los intervalos musicales. La escala de Mel parte de los experimentos perceptuales aplicados a diferentes escuchas que juzgan subjetivamente las distancias interválicas, difiriendo de las escalas musicales temperadas, las cuales no se construyeron haciendo análisis subjetivos, sino más bien responden a una necesidad musical para homogeneizar las distancias interválicas, y, de ese modo, poder realizar inflexiones y modulaciones a tonalidades más lejanas de la tonalidad inicial.[15] La escala de Mel es útil en la delimitación de miles de frecuencias que están en el aire y que de nada sirve procesar si nuestra percepción tiene un límite naturalmente biológico; de este modo posibilita ob-

[14]Beth Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

[15]S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

tener análisis más apegados a lo que los humanos somos capaces de percibir.

Analizar e interpretar sonido a partir de una máquina ha sido una tarea recurrentemente ejecutada desde el siglo XX. Por ejemplo, en la Segunda Guerra Mundial los alemanes encriptaban todos sus mensajes con el código Enigma, los cuales circulaban libremente en el aire transmitidos mediante ondas sonoras, accesibles para cualquiera que tuviera una radio. Alan Turing al estudiar los sistemas de descifrамiento y encriptación de la época se dio a la tarea de analizar los canales de información nazis con la primer computadora eletromecánica construida en la historia, dicho hallazgo fue determinante en la planeación de estrategias de guerra por parte de Inglaterra ya que le permitía saber con antelación todos los movimientos militares alemanes, lo que los llevó posteriormente a la victoria.

A principios de la década de los cincuenta surge la computadora digital CSIRAC (Council for Scientific and Industrial Research Automatic Computer), esta fue probablemente la primer computadora capaz de reproducir música de forma digital a través de una bocina a partir de la síntesis sonora, este hallazgo representa el origen de la música por computadora. El matemático Tomas Cherry la programó para tales fines e ideó un sistema de hojas perforadas que podía ser usado por cualquiera que conociera la notación musical convencional, de manera que las hojas perforadas podían almacenar y contener la información musical escrita para ser reproducida por CSIRAC.[16]

Más adelante, en los años noventa del siglo XX, los desarrolladores de software pudieron integrar algoritmos mucho más complejos debi-

[16]Paul Doornbusch. Computer sound synthesis in 1951: The music of csirac. *Computer Music Journal*, 28:10–25.

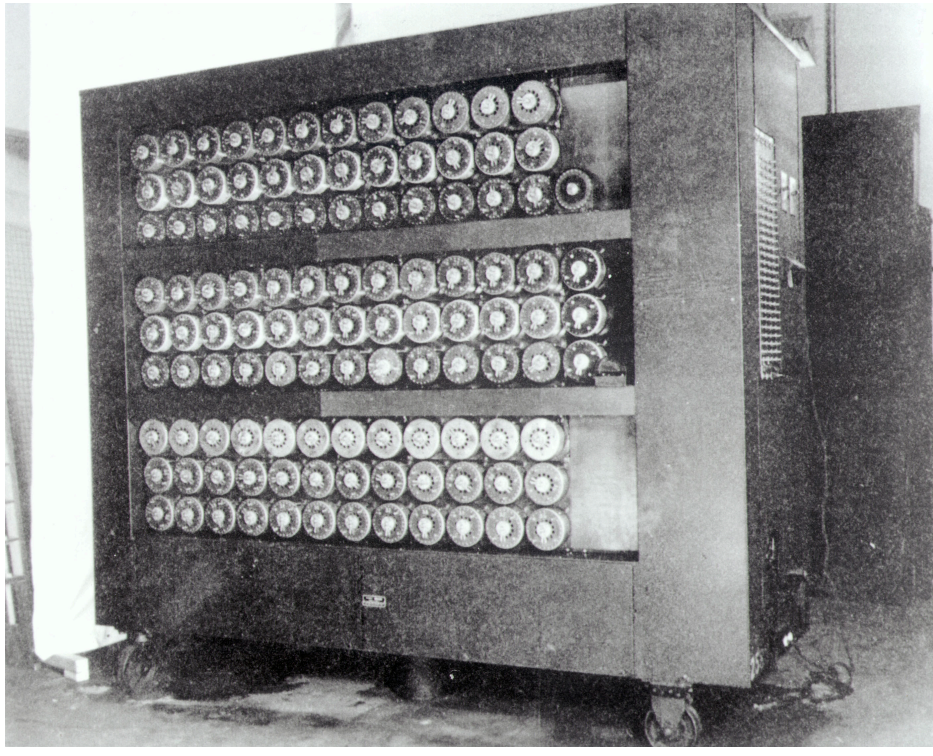

Figura 1.4: Máquina de Turing para decodificar y analizar el código Enigma, 1943

do al rendimiento que los microprocesadores aportaban a las computadoras que para ese momento ya eran capaces de procesar audio en tiempo real, es decir, que podía ser generado, ejecutado y manipulado al momento. El poder de procesamiento de las computadoras fue creciendo enormemente en pocos años y las capacidades de manipulación de los medios digitales se vieron enormemente afectadas, la forma en la que se genera la música y el sonido comenzaría a estar mediada completamente por la capacidad de procesamiento y algoritmos súper sofisticados.

Las máquinas que escuchan actualmente se constituyen de software y hardware y en teoría podrían simular el complejo sistema auditivo humano, con el objetivo de extraer información significativa de las señales de audio analizadas. Julius Smith, profesor de música e ingeniero eléctrico de Stanford, comenta que "analizar audio certeramente involucra varios campos de conocimiento: ingeniería eléctrica (análisis de espectro, filtración y transformación de audio); psicoacústica (percepción sonora); ciencias cognitivas (inteligencia artificial y neurociencia); acústica (física y producción sonora); y música (armonía, ritmo, y timbre)".[17] En pocas palabras, podríamos decir que es un campo transdisciplinar que busca extraer datos significativos y generar ciertas conclusiones acerca del sonido entendido como señal digital.

El ingeniero Paris Smaragdis del MIT se refiere a ellas como software que aprovecha las cualidades del sonido para ubicar personas dentro de un cuarto, monitorizar maquinaria para evitar posibles fallas, o activar cámaras para grabar un determinado momento en el que ocurre algo importante.[18] Las máquinas que escuchan no solo son

[17] http://web.media.mit.edu/~tristan/Classes/MAS.945/technical.html Fecha de consulta 28 de junio 2017

[18] https://bit.ly/2JvMmA5 Fecha de consulta 28 de junio 2017

capaces de discernir entre fenómenos acústicos sino también pueden diferenciar entre información más abstracta que responde a contextos humanos específicos; por ejemplo, análisis de datos enfocados en aplicaciones médicas, seguridad y vigilancia todas con finalidades comerciales y de control. También pueden detectar mecanismos cualitativos como los fenómenos psicoacústicos de percepción sonora humana como los desarrolla Eric Scheirer en su tesis doctoral Music-Listening Systems.[19] Asimismo, pueden distinguir cualidades acústicas y contextuales; la voz en distintos idiomas, el timbre de un instrumento, su amplitud, forma temporal o envolvente, ritmo, tonalidad musical, centro espectral e incluso formas estructurales musicales.

Estos algoritmos de identificación de características sonoras y musicales no serían capaces de discernir entre distintos tipos de fenómenos sin un componente esencial que es el cerebro de la máquina conocido como inteligencia artificial entendida también en esta tesis como aprendizaje de máquinas. Una forma de inteligencia artificial está inspirada en las múltiples y complejas interconexiones de las neuronas de cerebros biológicos, capaces de relacionar una información con otra y reaccionar de acuerdo con ciertas reglas de operación. En ciencias de la computación se conocen como redes neuronales artificiales y su característica más importante es que pueden aprender de forma conjunta a través de ciertas entradas que el usuario o el programador inserta en la red. Este tipo de programación se ha vuelto muy popular en la detección de características de sistemas de información altamente complejos. Dependiendo de la complejidad del problema a resolver, las redes neuronales pueden ir de unas pocas a miles de

[19] Eric D. Scheirer. *Music-Listening Systems*. PhD thesis, 2000.

unidades neuronales y conexiones sinápticas, una configuración por mucho menor a la complejidad existente en un cerebro humano.[20]

La inteligencia artificial no se limita solamente a las implementaciones con redes neuronales, sino que hay muchas más aproximaciones y perspectivas, algunas de estas serán discutidas con detenimiento en el capitulo 3. Algunas implementaciones de inteligencia artificial han permitido a algunos creadores como David Cope, Lejaren Hiller, Clarence Barlow, Francois Pachet, Geoge Lewis o Hugo Solís,[21] entre otros, generar proyectos capaces de reconocer a través de diferentes tipos de análisis estructuras musicales para generar nuevas composiciones, sistemas interactivos y herramientas de análisis musical.

Como se ha señalado, a partir de varios métodos desarrollados de forma mecánica, los avances en la anatomía humana, las tecnologías de grabación, manipulación, reproducción, transmisión de sonido y los avances en la inteligencia artificial que las máquinas que escuchan fueron paulatinamente configuradas y encapsuladas en distintos soportes digitales como programas y librerías dedicadas al análisis de señales digitales de audio.

[20] Actualmente no se requiere trabajar con miles o millones de redes neuronales primero por la enorme capacidad computacional que requeriría procesar millones de redes, segundo, porque trabajar con pocas unidades neuronales puede ser más que suficiente para resolver un problema de forma satisfactoria.

[21] http://hugosolis.net/hugosolisWP/impi-2/.

1.3. Implicaciones sobre el uso de tecnologías de aprendizaje y escucha automática

En los siguientes dos apartados presento de manera no exhaustiva pero si analítica algunos usos y aplicaciones sobre las tecnologías descritas anteriormente y cómo son empleadas en diferentes contextos socio-culturales. El objetivo del apartado es contextualizar a través de temas y discusiones actuales diferentes problemáticas e inquietudes personales que el desarrollo de las tecnologías de escucha automática, aprendizaje de máquinas e inteligencia artificial, conllevan.

Los temas aquí expuestos son interdependientes; se vinculan entre sí en un punto de inflexión común para todos, esto es, el desarrollo tecnológico de sistemas de escucha e inteligencia artificial aplicada a diferentes fines, ya sean musicales, políticos, militares, comerciales o artísticos. La tecnología usada en todos ellos es la misma: sistemas de organización de enormes cantidades de datos, generados a partir de diferentes algoritmos para deducir información clave sobre un contexto específico. El objetivo principal, la segmentación de distintos conjuntos para su identificación y clasificación, que van desde poblaciones de personas, animales, recursos naturales y digitales, todos ellos ligados al análisis y procesamiento de caudales de información de tipo textual, sonora o visual.

1.3.1. Modos de producción basados en la escucha y el aprendizaje automático: FlowComposer

A continuación se pondrá a discusión un caso muy particular y controversial: el proyecto llamado FlowComposer, herramienta para la asistencia composicional aplicada a la creación musical comercial dentro la industria del entretenimiento. Se intentará poner a discusión dicho proyecto para contrastarlo con perspectivas como "la máquina que escucha" planteada en esta tesis, o los proyectos abordados en el siguiente capítulo. Lo interesante, desde mi perspectiva, es que en todos estos trabajos el aprendizaje y la escucha automática son temas centrales. Considero que tanto mi proyecto como los referidos en el apartado *Algunas aproximaciones al aprendizaje y escucha automática aplicadas a la improvisación*, abren el panorama de acción a otros modos de producción posibles no centrados en intereses económicos, la eficiencia productiva, ni la minimización de gastos por parte de las empresas, sino que están interesados en la colaboración descentralizada, la disolución entre productor y usuario, la experimentación estética y la interactividad a través del desarrollo tecnológico.

El caso de FlowComposer es sumamente interesante ya que se relaciona con los ámbitos económico y de consumo musical, fue desarrollado por la empresa Sony en colaboración con Francois Pachet, Vincent Degroote et al.[22] FlowComposer, es un sistema basado en el

[22]Importante señalar que Francois Pachet también ha estado involucrado en la creación de sistemas inteligentes aplicados a la improvisación en estilos como el jazz y el blues. Continuator es uno de sus proyectos desarrollados en el laboratorio de ciencias de la computación Sony en París. Este sistema es capaz de reconocer patrones melódicos, dinámicas, secuencias de acordes, tonalidad, ritmo, pulso, e

conocimiento de 13000 obras musicales[23] y aprendizaje de máquinas que "puede componer en cualquier estilo"[24] y humor musical. Desde una perspectiva que promueve la colaboración compositiva, FlowComposer fue diseñado para componer música de forma autónoma para ayudarnos a ser "más creativos". Al contar con la asistencia de un compositor superinteligente capaz de realizar "combinaciones únicas

incluso ciertas imprecisiones del ejecutante al tocar. Continuator puede responder a (continuar) las entradas que un músico le proporciona, adecuando su estilo al del músico. Pachet detalla que Continuator es más un sistema "reflejante" que "flexible", busca sincronizarse en forma de eco con los materiales musicales introducidos por el usuario, de manera tal que su sistema logra pasar la prueba de Turing, siendo que para un escucha ordinario podría pasar desapercibido el momento en el que la máquina o el músico tocan. Esto debido a varias razones relacionadas con cómo el material subsecuente es reinterpretado a través de los procesos generados con cadenas de Markov y además parámetros como proporción temporal, rítmica, dinámica, agógica, tímbrica y particularmente el estilo musical son los iguales a lo que el humano toca. En las audiciones de los ejemplos de la página web de Pachet se puede notar que en improvisaciones con instrumentos MIDI, resulta difícil determinar si la máquina o el humano está tocando. En otro ejemplo en donde se implementó un sistema mecánico para que Continuator tocara directamente en un piano acústico mientras un instrumentista improvisa, resulta más sencillo identificar las diferencias ya que el trabajo de fraseo, articulación, dinámica, uso de pedal y flexibilidad agógica son demasiado rigurosos en la interpretación del piano mecánico. Pese a sus limitantes, Continuator resulta ser un proyecto realmente interesante, pudiendo acceder a los ámbitos de la pedagogía infantil e incluso a la práctica de la libre improvisación.

[23] http://lsdb.flow-machines.com/ Fecha de consulta 7 de febrero 2018

[24] http://www.flow-machines.com/ai-music/ Fecha de consulta 7 de febrero 2018

de transferencia de estilo, técnicas de optimización e interacción [...]", seria posible ser más productivo.[25]

Algunos medios de divulgación[26] han demostrado que este sistema ha sido ampliamente usado por Spotify para integrar música de artistas falsos a su base de datos y de esta forma ahorrar millones de dólares en regalías debido a la enorme aceptación y a los millones de reproducciones que tiene esta música en la plataforma. Spotify está ahorrando en pagos que podrían ser para artistas a los que se comisionara la creación de música en un estilo determinado destinada a formar parte de sus listas de reproducción. Desde 2012 Pachet ha creado canciones en el estilo de los Beatles, Duke Ellington y otros artistas de la escena del pop, y en años recientes comenzó a trabajar en la creación automatizada de música ambiental de forma masiva. Lo anterior se dio a conocer por una investigación llevada a cabo por Music Business Worldwide la cual informa que Spotify tiene muchos artistas y compositores inexistentes en el mercado. Es importante señalar que al realizar una búsqueda por internet no hay ningún rastro de estos artistas. Independientemente de esto, los artistas desconocidos de Spotify componen música para listas de reproducción basadas en temáticas como *Peaceful Piano*, *Ambient Chill*, entre otras, precisamente el tipo de música que Pachet estaba creando con su compositor basado en inteligencia artificial. Cabe mencionar que es un caso controversial ya que dicho medio de divulgación ha mostrado una lista de más de 50 autores desconocidos;[27] otros medios dicen que varios de estos artistas desconocidos son pseudónimos que un solo artista

[25] http://www.flow-machines.com/ai-makes-pop-music/ Fecha de consulta 7 de febrero 2018

[26] https://bit.ly/2tcOqJw Fecha de consulta 7 de febrero 2018

[27] https://bit.ly/2sXcYWz Fecha de consulta 7 de febrero 2018

usa como estrategia de mercado, pero el hecho es que mucha de la música escuchada en estas listas de reproducción no tienen ninguna plataforma de reproducción más allá de Spotify, además, no hay información sobre ellos ni en ni internet y ni siquiera en su plataforma, y lo más importante es que este tipo de música tiene varias características en común: tímbricas de instrumentos digitales, tendencia a pedales o continuos atmosféricos, canciones que duran no más de 4 minutos, formas cíclicas poco creativas, ritmos de cuatro cuartos, tempo estable, repeticiones exactamente iguales dentro de círculos armónicos claramente definibles y sin presencia de modulaciones o inflexiones a otras tonalidades. Esto sugiere que se trata de música generativa compuesta por una máquina. Lo que dice Spotify es que la función de Pachet en la empresa es desarrollar herramientas de asistencia para la composición, además niegan que mucha de su música haya sido creada por artistas falsos para evitar pagar regalías, y que más bien están publicadas bajo seudónimos pero que fueron escritas por personas reales a las cuales se les paga.

Otro caso similar es la compañía Amper Music, la cual ha creado y puesto a disposición de millones de usuarios un compositor, intérprete y productor automatizado con la capacidad de crear música de fondo con tan solo especificar algunos parámetros, humores y estilos musicales, esté sistema es capaz de crear musica original para diversos fines. [28]

Pero surge la pregunta que Pouge señala en su artículo, "¿Por qué Spotify, o cualquier servicio de música, no podría comenzar a usar IA para generar música gratis y ahorrarse dinero? La automatización ya está en camino de desplazar a millones de taxistas humanos, con-

[28] https://www.ampermusic.com/ Fecha de consulta 7 de febrero 2018

ductores de camiones y trabajadores de comida rápida. ¿Por qué los artistas y músicos deberían estar exentos de la misma economía?"[29]

Estos casos revelan el surgimiento de nuevas polémicas en torno a la idea de autoría, autenticidad y originalidad. La música se vende a los usuarios bajo la convención –aun imperante– de que se trata de una creación individual humana; sin embargo, detrás de esta estrategia publicitaria hay una transformación radical en los modos de producción musical. En términos técnicos no podría decirse claramente de quién es la autoría de una pieza, si de la enorme base de datos de canciones, del algoritmo que la produce, del grupo de técnicos que programaron el sistema, del usuario que elige los parámetros del sistema o del productor que da los toques finales a las piezas y la prepara para poder ser escuchada en los servicios en línea de las compañías. Lo que hay detrás del nombre ficticio de un artista es una cantidad enorme de personas y procesos automatizados basados en el análisis de miles de piezas existentes y la asimilación de múltiples estilos musicales. El resultado no es ni humano ni maquínico sino un arreglo emergente que trasciende la dicotomía aparente de este intercambio. La autoría individual deja de importar debido a estas nuevas formas de creación-colaboración híbridas. Por un lado, el concepto de compositor se difumina y comienza una nueva forma de entender la creatividad colectiva; por otro, miles de artistas que podrían ocupar estos espacios y recibir las regalías correspondientes por su trabajo, están siendo desplazados. Entonces ¿podría decirse que estas prácticas son las nuevas dinámicas que buscan monopolizar las prácticas

[29]Pogue, David. Is Art Created by AI Really Art? Scientific American. https://www.scientificamerican.com/article/is-art-created-by-ai-really-art/ Fecha de consulta 7 de febrero 2018

musicales, en las que todo el dinero generado se distribuye solamente a un grupo de personas? Como hemos visto a la largo de la historia, la tecnología ha tendido a suplantar los modelos de trabajo humanos; desde los mecánicos, y ahora también los modelos creativos, imaginativos, intelectuales y artísticos. Actualmente hay intereses económicos en el discurso de la "asistencia" al compositor, pero porqué afirmar que se trata de una herramienta de asistencia a la creatividad cuando más bien pareciera una sustitución del humano camuflada en "asistencia". En este sentido, surgen las siguientes preguntas que se quedarán como reflexiones para cerrar este pequeño apartado. ¿Cuáles son los límites de la asistencia en los procesos creativos (pensando también en la libre improvisación)? ¿Debería haber esos límites o algún tipo de regulación?

1.3.2. Algunas reflexiones sobre la escucha y el aprendizaje automático

Vivimos en una forma de determinismo tecnológico; éste es efecto y detonador de cambios estructurales en la cultura, la sociedad y la ecología, pues genera las formas "prediseñadas de interacción social y afectiva" actuales.[30] En este determinismo, los mecanismos automatizados desempeñan un papel muy importante ya que van tejiendo los puentes de una estructura interaccional y colaborativa entre las propias máquinas y diferentes seres vivos. El acrecentamiento de la domesticación tecnológica obliga a modificar los comportamientos de los usuarios, reforzando progresivamente conductas como la pérdida

[30]Juan Martín Prada. *Prácticas artísticas e Internet en la época de la redes sociales.* AKAL, 2015 p. 26.

de agenciación humana, o la distorsión/transformación de los procesos comunicativos entre los individuos. Siendo las máquinas una extensión de la condición humana que procura —a través de nosotros— imitar o incluso exceder nuestra propia capacidad ¿cómo nos relacionamos con ellas?, ¿implican un riesgo para nuestra existencia o la extienden y la transforman?, ¿qué modifican en nosotros, en nuestras interacciones con otros y con otras máquinas?

Estamos en un momento donde cualquier persona con una computadora e internet puede tener acceso a un poder de cómputo lo suficientemente alto para interpretar muchísima información. Asimismo, es posible acceder al estado actual del aprendizaje de máquinas e implementaciones tecnológicas (o productos culturales); bases de datos, código, artículos, publicaciones académicas, blogs, videos y libros (que es posible descargar desde páginas de internet de formas *clandestinas* y/o de acceso abierto). Esto habla por un lado, de una democratización del conocimiento enfocado en el aprendizaje de máquinas y la inteligencia artificial. Una forma de construcción del conocimiento colectiva y descentralizada (aunque también institucionalizada, véase ISMIR[31]), que interesa a las grandes corporaciones desarrollar lo más rápido y eficientemente posible. Por otro lado, es también una estrategia para la inmediata cooptación del conocimiento nuevo por las empresas y su abierta cartera de trabajo para futuras contrataciones en los grandes consorcios corporativos (Facebook, Amazon, Google, Apple, Microsoft, etc). Una vieja y nueva forma de control en la que el mismo capital va creciendo económica y tecnológicamente más rápido. Actualmente muchas empresas por esa razón abren cada vez más

[31] http://www.ismir.net/resources.html Fecha de consulta 19 de febrero 2017.

su código, desde concursos a lo largo del año[32] para el desarrollo de herramientas especificas, hasta congresos anuales, donde es posible conocer todo tipo de proyectos relacionados con la vigilancia, asistencia médica, etc.[33]

Es una estrategia bastante acertada ya que habla de una diversificación económica y de aprendizaje que va hacia todos lados y en todas direcciones. Entre más rápido avance el conocimiento en el desarrollo del aprendizaje de máquinas, más ingresos pueden percibir, más subjetividades pueden moldear y más inteligente se va haciendo el sistema. Si las corporaciones, agencias gubernamentales y privadas son capaces de desarrollar mayor poder de pensamiento y análisis de datos, pueden actuar con mayor eficacia a posibles nuevos cambios dentro del propio sistema, impulsando nuevas estrategias para llegar a más nichos de consumo y/o mantener en un estado de alienación eterno en la población.

Con las nuevas inserciones de dispositivos tecnológicos en nuestra vida, surgen nuevas formas de mirar, escuchar e interactuar, y también, otros mecanismos para concentrar nuestra atención así como para disiparla. Jonathan Crary los llama "recepción en estado de distracción". En este sentido, las diferentes formas actuales de escucha no necesariamente involucran la presencia sonora sino también otras formas de presencia que nos obligan a atender y entender la información que circula a través de los medios. Por ejemplo, Crawford señala que actualmente podemos pensar en una forma de escucha en línea, aquí usa la metáfora de escucha para dar cuenta del proceso imparable de la transmisión de información en red similar al que sucede

[32] https://youtu.be/AoRSIdLpFqU, https://www.kaggle.com/competitions Fecha de consulta 19 de febrero 2017.

[33] https://www.reddit.com/r/MachineLearning/

al escuchar ya que es imposible cerrar la escucha así como cerrar la transmisión de datos que se produce en la red debido a su carácter descentralizado.

> Escuchar en línea puede ser entendido en cualquier número de contextos: sean wikis, MUD's, blogs, listas de correo e incluso alimentadores RSS. [...] Escuchar no es una metáfora común para la actividad en línea. De hecho, la participación en línea ha tendido a confundirse con aportar una "voz". "Hablar" se ha convertido en la metáfora dominante para la participación en espacios en línea tales como blogs, wikis, sitios de noticias y listas de discusión.[34]

Poner atención *online* es también entendido como una práctica de escucha, de atender o entender algo. Pero, ¿quiénes son los que atienden y filtran esta información? ¿Para qué fines es empleada la información extraída? Kate Crawford destaca diferentes tipos de escucha que pueden actuar en la red: una persona, una compañía y un partido político. Estas dos últimas con la posibilidad de desarrollar una capacidad mayor para poder escuchar a más personas al mismo tiempo. Es cierto que actualmente mucha de la información, sonora, "vocal" y sensible que guardan de nosotros las compañías a las cuales hemos decidido regalársela, es vendida, traficada, comercializada a los grandes consorcios corporativos para ofrecer más productos y servicios que son completamente dinámicos y adaptables a necesidades y situaciones diversas. Cada vez es más común escuchar decir algo por la mañana y después de unas horas encontrar anuncios o información

[34]Kate Crawford. *Following You: Disciplines of Listening in Social Media.* The Sound Studies Reader. Taylor & Francis, 2012 p. 79.

relacionada con aquello que se había dicho en las cuentas de correo, teléfono o redes sociales. Es evidente que este sistema interconectado nos vigila y puede predecir cada uno de los pasos que damos, si entramos a un coche, si vamos a nuestra casa o al trabajo o si estamos interesados en cierto tema, etc. Estos procesos son mejor conocidos como minería de datos, *bigdata* o *machine learning* y están involucrados con las formas en las que diferentes compañías con intereses diversos ponen atención online para la generación de nuevas necesidades y productos a través de la vigilancia masiva.

Hay mucha información que fluye y prácticamente es asequible por cualquiera que esté interesado en tomarla y aprender de ella, evidentemente también mucha información que es inasequible o está encriptada, sin embargo, es posible generar una base de datos propia, comprarla u obtener alguna base de datos pública que esté disponible en línea. Prueba de ello son las siguientes plataformas que cuentan con una amplia diversidad de temas de interés para realizar análisis de datos con diversos fines.[35]

La extracción de información online es posible básicamente sobre cualquier página de internet; sabiendo utilizar esta tecnología es factible construir bases de datos, ya sean basadas en videos, imágenes, correos, tweets, mensajes de Facebook, etc. Un buen lugar para comenzar es la librería *Beautiful Soup* de Python.[36] Empresas como Youtube y Google han generado grandes bases de datos basadas en los contenidos que todos subimos a sus servidores. Estas pueden ser descargadas para su análisis, observación o implementación en proyectos

[35] https://www.kaggle.com/ Fecha de consulta 28 de enero 2018.
http://deeplearning.net/datasets/ Fecha de consulta 28 de enero 2018.
[36] https://do.co/2KzQNe8 Fecha de consulta 28 de enero 2018.

y aplicaciones.[37] Un ejemplo interesante que parte de estos métodos es el proyecto actualmente congelado de Daniel Jones y Peter Gregson *The Listening Machine* (la máquina que escucha). Ellos lo definen como un:

> sistema automatizado que genera una pieza de música continua basada en la actividad de 500 usuarios de Twitter alrededor del Reino Unido. Sus conversaciones, pensamientos y sentimientos son transferidos en patrones musicales en tiempo real, a los cuales te puedes sintonizar en cualquier momento a través de cualquier dispositivo conectado a la red.[38]

Este sistema basado en la observación de patrones de comportamiento para su transducción a piezas musicales es el claro ejemplo de la escucha *online* que menciona Crawford. Poner atención *online* para extraer información significativa y vaciarla en un objeto artístico sería lo más ingenuo que podemos hacer con estos datos, para los centros de poder ¿qué representan todas estas herramientas y nuevas posibilidades para observar y desentrañar nuestros patrones de comportamiento y geolocalizarnos físicamente en cualquier momento con un detalle y certeza brutal?

Las recientes aproximaciones al reconocimiento de señales de audio (Music Information Retrival, MIR) están mayoritariamente centradas en desarrollar algoritmos para el control automático de dispo-

[37] https://research.google.com/youtube8m/index.html Fecha de consulta 28 de enero 2018.

https://research.google.com/audioset/ Fecha de consulta 28 de enero 2018.

[38] http://www.thelisteningmachine.org/ Fecha de consulta 28 de junio 2017

sitivos para la guerra, robots de rescate, herramientas de vigilancia, el análisis clínico y el reconocimiento de géneros basados en un corpus musical con propósitos comerciales. Estos algoritmos identifican información relevante de audio, la cual extraen en grandes cantidades para almacenarla en listas de valores separados por comas (csv por sus siglas en inglés). Esta información tiene que ser analizada y organizada en clases o *clusters* —proceso conocido como entrenamiento o *training*— con programas enfocados en el aprendizaje de máquinas o *machine learning*. Entre los programas utilizados están Weka, R, CAFFE, la librería scikit-learn, Tensorflow, entre otros. Los asistentes personales inteligentes se han convertido en una parte fundamental para "comunicarnos con" la tecnología, enviando información en tiempo-real a las empresas que ofrecen estos servicios. Los sistemas de vigilancia que utilizan máquinas que escuchan han sido importantes para la detección de momentos relevantes con el fin de revelar información clave acerca del comportamiento de una persona a través de los micrófonos insertos en casi todos los dispositivos móviles o fijos que utilizamos actualmente.

Algunos países como Alemania[39] han tomado medidas contra el esparcimiento de dispositivos de vigilancia disfrazados de electrodomésticos o juguetes pero debido a la enorme cantidad de población y la desenfrenada velocidad de producción de consumibles no se han podido regular plenamente. Actualmente estamos en una época bastante característica, debido a la falta de regulaciones y políticas públicas por parte de los gobiernos y los sistemas de control para frenar el libre esparcimiento de la información así como de los desarrollos tecnológicos. Estos posibilitan la apertura y el acceso libre a muchas

[39] Ídem

tecnologías de punta relacionadas con la Inteligencia Artificial misma, que agencias gubernamentales así como privadas usan para llevar a cabo los desarrollos más sofisticados relacionados con asistencia y la autonomía policial y militar.[40]

La creación de armas autónomas es inevitable, así como el surgimiento de nuevas formas orden y control. El gobierno de EEUU ha mostrado abiertamente en un documento titulado Human Systems Roadmap Review[41] el uso de sistemas de aprendizaje automático para delegar decisiones a las máquinas cuando así se requiera, además de la colaboración humano-máquina y el empleo de armas autónomas que simulan a través de modelos computacionales las capacidades cognitivas, psicomotoras y perceptivas del humano. Estas armas pueden ser capaces de detectar a sus enemigos y tomar la decisión de disparar cuando considere conveniente. Todas estas estrategias están siendo destinadas para los ambientes de guerra cyber-electrónicas.[42] Para que decisiones de este tipo sean tomadas por un robot y los legisladores puedan realmente aceptar llevar a cabo este tipo de operaciones, es necesario que las máquinas puedan explicarse por sí mismas, es decir, justificar las razones por las cuales han decidido llevar a cabo cierta acción, especialmente en el aprendizaje profundo de máquinas que involucra procesos increíblemente complejos para resolver problemas.

> El aprendizaje profundo es especialmente críptico debido a su increíble complejidad. Está más o menos inspirado en

[40] https://bit.ly/2IcZxc1, https://bit.ly/2swd4Uh, https://bit.ly/2qJwYJO Fecha de consulta 28 de enero 2017.

[41] https://bit.ly/2HOhU7I Fecha de consulta 28 de enero 2017.

[42] https://engt.co/2rqtSb0, https://bit.ly/2HRxzD2 Fecha de consulta 28 de enero 2017.

> el proceso por el cual las neuronas en el cerebro aprenden en respuesta a una entrada. Muchas capas simuladas de neuronas y sinapsis son datos etiquetados y su comportamiento se ajusta hasta que aprenden a reconocer, por ejemplo, un gato en una fotografía. Pero el modelo aprendido por el sistema está codificado en el peso de muchos millones de neuronas y, por lo tanto, es muy difícil de examinar. Cuando una red de aprendizaje profundo reconoce a un gato, por ejemplo, no está claro si el sistema se enfoca en los bigotes, las orejas o incluso la [silueta] del gato en una imagen.[43]

Se ha vuelto común pensar a este tipo de sistemas equivalentes a cajas negras, donde no se sabe por qué ni cómo es que un determinado problema es resuelto por el algoritmo. Sin embargo, agencias como DARPA (Defense Advanced Research Projects Agency) actualmente financian proyectos destinados a la *autoexplicación* de sistemas basados en el aprendizaje de máquinas. Algunos avances en los proyectos por parte de las empresas financiadas como Charles River Analytics han logrado que el sistema de aprendizaje automático señale áreas de imágenes que son relevantes para la clasificación así como explicaciones basadas en lenguaje natural.[44]

[43] https://bit.ly/2lX0afj Fecha de consulta 28 de enero 2017.
[44] Ídem

Figura 1.5: StackGAN: texto a imágenes realísticas

Referente al uso del aprendizaje automático, una reciente propuesta es la experimentación con el algoritmo de aprendizaje profundo, GANs (Generative Adversarial Networks), una forma de aprendizaje profundo sin supervisión o semi-supervisado.[45] En recientes implementaciones artísticas el algoritmo ha demostrado que las máquinas pueden proponer comportamientos creativos bastante interesantes. Por ejemplo, el proyecto visual How computers are learning to be creative, caracteriza un continuo de transformaciones de una imagen inicial, simulando un "zoom alucinante" a imágenes que contienen a su vez otras imágenes, esto, a partir de lo que reconoce en su base de datos. Por ejemplo, si su base de datos fueran animales, crearía

[45]Esta forma de aprendizaje tiene la capacidad de corroborar la información mediante dos redes neuronales; la primera, evalúa si la información de entrada es correcta y la segunda, trata de engañar a la primera red para que esta se haga cada vez más inteligente y pueda determinar qué es un engaño y qué no. Asimismo la red neuronal que engaña, genera muestras modificadas o transformadas de las muestras originales, reforzando cada vez más su comportamiento. Estas dos redes entran en una continua retroalimentación de aprendizaje mutuo.

una sensación de metamorfosis pictórica fractálica en cada uno de los elementos que componen la imagen inicial.

Como hemos visto, el aprendizaje de máquinas es actualmente usado por muchos ámbitos y desde diversos intereses, además de que su uso se hace cada vez más cotidiano y necesario para acceder a la red masiva de información actualmente disponible.

Por otro lado, podríamos decir, respaldado de autores como Bostrom quien ha estado teorizando desde la filosofía sobre los posibles escenarios futuros del desarrollo de la inteligencia artificial, que en unos cuantos años sería posible vivir en un estado (político y territorial) dominado por las lógicas producto de la inteligencia de las máquinas. Si llegásemos a presenciar la llegada de la singularidad tecnológica, —es decir, el momento en el que las máquinas son capaces de desarrollar su propio software, generando una autopoiesis recursiva y por tanto un auto-mejoramiento constante y exponencial, ¿qué o quien podría detenerlas de su propia autoevolución desenfrenada?

> Si algún día construimos cerebros de máquinas que superen los cerebros humanos en inteligencia general, entonces esta nueva superinteligencia podría volverse muy poderosa. Y, como el destino de los gorilas ahora depende más de nosotros los humanos que de los gorilas mismos, entonces el destino de nuestra especie dependería de las acciones de la superinteligencia de la máquina.[46]

En un escenario como este, ni siquiera el conocimiento de toda la humanidad podría competir contra estas nuevas tecnologías, ya

[46] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014 p. 2

que la capacidad de entendimiento humana se vería constantemente sobrepasada por millones de redes inteligentes interconectadas capaces de actuar en tiempo-real; estaríamos presenciando el advenimiento de cambios sociales inimaginables que superan por mucho los actuales conceptos de dominación y control, volviéndose imposibles de prever o comprender por cualquier ser humano. Sin embargo, como señala Nick Bostrom:

> Tenemos una ventaja: podemos construir las cosas. En principio, podríamos construir una especie de superinteligencia que protegería los valores humanos. Ciertamente tendremos una razón fuerte para hacerlo. En la práctica, el problema del control —el problema de cómo controlar lo que haría la superinteligencia— parece bastante difícil. También parece que solo tendremos una oportunidad. Una vez que exista una superinteligencia antipática, nos impediría reemplazarla o cambiar sus preferencias. Nuestro destino estaría sellado. [...] Éste es posiblemente el desafío más importante y desalentador que la humanidad haya enfrentado. Y —ya sea que tengamos éxito o fracasamos— probablemente sea el último desafío al que nos enfrentaremos.[47]

En este sentido es vital el papel que juegan la fantasía, el arte y en general las humanidades en la construcción de los imaginarios colectivos y metáforas sobre el futuro, ya que de cierta forma son capaces de construir todo un conjunto de posibles realidades, o escenarios que

[47] Ídem p. 2

en el futuro —como hemos visto en esta época, muchos de los imaginarios actuales por más lejanos que estén de la realidad actual— posiblemente se lleguen a concretar. El género de la ciencia ficción podría ser considerado como una herramienta para prepararnos, un arma a favor de los vencedores para la generación de subjetividades que de múltiples formas nos van preparando el terreno del futuro, esos posibles escenarios que en unos cuantos años podríamos enfrentar. Una opinión personal y meramente didáctica sería imaginar cómo y dónde queremos estar en el futuro y hacer que nuestros más anhelados sueños se hagan realidad en favor de nuestro bien común.

La cuestión es si el uso de estas tecnologías en el arte podría aportar algo o más bien estaría legitimando las prácticas de la tecnociencia, sin siquiera cuestionarse qué es lo que, a una escala más amplia, soporta estos desarrollos tecnológicos. Pese a que cada vez se hace más común la utilización del aprendizaje de máquinas en las artes, aún hacen falta más iniciativas artísticas que cuestionen las implicaciones socio-políticas, lejos de las implementaciones tecnocráticas, y generar un conocimiento más allá de las clásicas aproximaciones científicas basadas no solo en el antropocentrísmo sino también en la dominación sobre el mismo ser humano, que por lo general atentan contra la propia vida sensible y la ecología del planeta. Atendiendo a una perspectiva que cuestione el trasfondo socio-cultural de las prácticas de *machine learning*, está la posibilidad de apropiarse de la tecnología para conocer su funcionamiento, límites y posibilidades, además de anticipar futuras formas de desarrollo y tomar las medidas necesarias para su transformación en favor de una ética adecuada para las futuras integraciones tecnológicas dentro de las sociedades, así como dentro del medio ambiente.

Una visión interesante podría ser la Friedjof Capra, aunque no lo dice desde la perspectiva del aprendizaje automático sino desde la necesidad urgente del surgimiento de una ética ecológica en relación al desarrollo de la tecnociencia, que contemple plenamente la vida y las interrelaciones de todos los organismos.

> Dicha ética profundamente ecológica se necesita urgentemente hoy en día y muy especialmente en la ciencia, puesto que mucho de lo que los científicos están haciendo no es constructivo y respetuoso con la vida [...]. Con físicos diseñando sistemas de armas capaces de borrar la vida de la faz de la tierra, con químicos contaminando el planeta, con biólogos soltando nuevos y desconocidos microorganismos sin conocer sus consecuencias, con psicólogos y otros científicos torturando animales en nombre del progreso científico, con todo ello en marcha, la introducción de unos estándares "ecoéticos" en el mundo [tecno]científico parece de la máxima urgencia. [48]

Para ir cerrando e intentar responder las preguntas bosquejadas en el texto, tales como ¿qué es lo que estos temas tienen que ver con el proyecto planteado y por qué me interesa mencionarlos? Además, ¿qué injerencia podría tener el arte dentro de la ciencia o el desarrollo tecnológico al usar tecnologías de aprendizaje de máquinas aplicada a la creación de obras o proyectos, sonoros, visuales, o hipermediales?

[48] Fritjof Capra. *La trama de la vida: Una nueva perspectiva de los sistemas vivos.* Colección compactos. Editorial Anagrama S.A., 2009 p.32

Por un lado, la tecnología empleada en muchos proyectos de índole diversa así como la metodología es finalmente la misma.[49] Me intriga pensar en que esta metodología que pareciera tan sencilla puede desembocar en herramientas que son usadas para el control, la manipulación, la vigilancia, la guerra y además para crear música y proyectos de arte. Por otro, considero de suma importancia lanzar estas reflexiones (aunque sean pocas) en torno a las tecnologías de aprendizaje automático debido al poco por no decir nulo criterio que hay en estos campos en relación con temas de orden político social y cultural, y, que se ve reflejado de forma general en las perspectivas actuales de la ciencia. A mi modo de ver estas procuran despolitizar y evadir cualquier cuestionamiento de carácter filosófico, ético, estético o epistémico en pos de la búsqueda por la eficiencia a través de una arborescencia súper especializante que con gran afán buscan los paradigmas epistémicos científicos actuales.

En mi caso pretendo apropiarme de algunas tecnologías de aprendizaje automático a través de su implementación aplicada al análisis de improvisaciones libres con el objetivo de generar modelos descriptivos que den cuenta de las formas en que algunos improvisadores libres nos acercamos a dicha práctica. Además el sistema propuesto en esta investigación difiere de los objetivos y usos comunes de las

[49]Esto es, seleccionar una base de datos para extraer información significativa, clasificarla mediante algún algoritmo de aprendizaje o segmentación, generar modelos descriptivos y comprobar que funcionen con información de entrada que no haya sido vista antes por el modelo,en este paso hay que corroborar si los resultados fueron o no satisfactorios, en caso negativo hay que regresar a la parte de extracción de datos y jugar con otros parámetros. Posteriormente, verificar con los nuevos datos extraídos si es posible obtener un modelo más fiel a lo que se pretende encontrar.

tecnologías de aprendizaje automático desde que no es un proyecto que se enfoca en analizar grupos de personas especificas con fines comerciales o de marketing, ni hacia la generación de herramientas musicales con intereses económicos sino más bien desde búsquedas estéticas experimentales que intenten cuestionar una práctica que de por si se considera liminal.

Más específicamente, el sistema de la máquina que escucha se caracteriza por realizar análisis dentro de una base de datos seleccionada por convicción propia con el objetivo de poder generar modelos de escucha que describan formas arquetípicas de aproximarse a la libre improvisación, en este caso con varias finalidades. Una de ellas sería generar modelos que den cuenta de como es la aproximación de un determinado improvisador libre a dicha práctica en cuanto a densidad sonora y timbre. Además, el modelo generado podría servir para entender como maneja los materiales ese improvisador, esto podría tener distintas utilidades y aplicaciones: para un análisis propio, si interesa al improvisador pensando en que este modelo sería como una huella a la cual es posible acceder en cualquier momento. Para que el modelo descriptivo al igual que las herramientas desarrolladas puedan ser empleadas, estudiadas, modificadas, transformadas, compartidas en otras improvisaciones por otros improvisadores o en otros contextos distintos a los de la improvisación. Para un futuro desarrollo de las herramientas donde un sistema sonoro pueda generar nueva música partiendo del modelo de improvisación generado. Finalmente para generar una reflexión a través del modelo generado que permita entender cómo los improvisadores se acercan a esta práctica sonora, revelar que características son las que definen estilistamente esta práctica y posiblemente para intentar modificarla, llegar a lugares nuevos, y se-

guirla transgrediendo a través de un conocimiento más profundo sobre de ella.

El aporte del arte es que puede fomentar esa generación de imaginarios posibles más allá de los que el cine comercial de ciencia ficción nos vende, más allá de eso generar imaginarios no solo sobre la destrucción del humano en manos de aquella singularidad de las máquinas, sino a través de imaginarios que alienten la posibilidad de redirigir esa tendencia desoladora y autodestructiva, en favor tendencias más sustentables con la vida misma en comunión con la tecnología, que abra cuestionamientos hacia los modos actuales en que nos hibridamos con la tecnología y cómo pasamos de ser seres conectados con la naturaleza a ser seres conectados por la tecnología a través alambres de cobre y frecuencias que dañan otras especies del planeta. Desde el arte y la filosofía podríamos preguntarnos ¿hay alguna idea de limite para la especie humana que vaya más allá de de ese catastrófico imaginario desarrollista? o tenemos que destruir a más especies para poder seguir sobreviviendo, al punto en el que creemos una especie que nos destruya a nosotros mimos y si tenemos suerte, que en el futuro se hable de la especie humana como la creadora de la especie de máquinas conscientes.

Además considero que como artistas y ciudadanos, es de suma importancia tomar el control sobre las tecnologías relacionadas con el aprendizaje automático de máquinas. Primordialmente no conformarse siendo usuarios consumidores de la misma tecnología sino apropiarnos de ella, saber hacerla, modificarla y crearla. Ya que si solo unos cuantos la conocen podrían hacer con esa tecnología lo que quisieran como esta ocurriendo y ha ocurrido siempre, no hay que esperar a que las legislaciones aparezcan en contra de nosotros sino nosotros

mismos hacerlas esas legislaciones en favor de un bien común. La era de la información en que vivimos es una puerta hacia un cambio paradigmático sin precedentes somos una sociedad interconectada una especie de sistema consciente que esta despertando de forma paralela e unificada.

Capítulo 2

Aproximaciones al aprendizaje y escucha automática en contextos de improvisación musical

Entre una enorme cantidad de proyectos, investigaciones e información relacionada con la generación musical a partir de máquinas que escuchan y aprendizaje de máquinas, pongo a discusión cinco proyectos que se relacionan de manera directa con los temas de escucha automática e improvisación. Elegí estos proyectos debido a la cercanía que guardan con el proyecto planteado en esta tesis en cuanto a sus aproximaciones metodológicas, prácticas e intereses estéticos y en la interactividad. Las metodologías de estos proyectos fueron útiles para conocer lo que se ha hecho en términos de experimentación creativa dentro de sistemas interactivos con el objetivo de modelar formas para improvisar. Además, todos los proyectos son iniciativas

recientes con una amplia documentación: artículos en blogs, artículos publicados en revistas, reseñas, entrevistas, muestras en festivales y grabaciones de conciertos. Sin embargo, encontré en ellos mucho de lo que sí ha funcionado en contraste con la escasez de los experimentos fallidos que considero también pueden aportar conocimiento para otras exploraciones.

2.1. Shimon

El caso de Shimon es excepcional; después de 12 años de trabajo, Gil Weinberg junto con su equipo de estudiantes y desarrolladores han creado un robot capaz de tocar la marimba e improvisar junto con humanos. [1]

Un antecedente de este proyecto es Haille,[2] un robot percusionista interactivo creado por los mismos investigadores en la Universidad de Geogia Tech en 2006. Este robot escucha, extrae, analiza y decide formas de comportamiento para ser ejecutadas de forma física en un tambor. Muchas de estas cualidades fueron heredadas a su sucesor Shimon. Además de improvisar de forma sonora, también lo hace a través de su presencia física (mientras que Haile solo mueve sus brazos al tocar) y el movimiento que establece con el espacio y con sus pares. A este respecto cabría cuestionarse cómo es la relación improvisatoria entre Shimon y el/los humano(s) con los que improvisa. Pensando en que la interacción va más allá del proceso de escucha

[1] https://qz.com/689827/moogfest-shimon-music-robot/ Fecha de consulta 25 de junio 2017.

[2] https://www.youtube.com/watch?v=veQS6tsogAA. Fecha de consulta 25 de junio 2017.

y acompañamiento, improvisar con un robot que supera físicamente las capacidades humanas implica inmediatamente un reto como improvisadores (y humanos) debido a su extensión corporal, destreza, precisión, energía inagotable, capacidad de anticipación y vigor (como será explicado más adelante). Además de estas características propias de Shimon, éste autoregula su comportamiento adaptándose al momento de la improvisación pudiendo asumir varios roles de interacción como se puede observar en los siguientes videos de la nota al pie.[3]

Algunos de estos roles de interacción incluyen el acompañamiento, seguimiento, generación de solos basados en una progresión armónica, aunque por lo visto en los videos no es capaz de proponer nuevas ideas musicales o crearlas desde su propio aprendizaje, ya que el proyecto está enfocado en generar una colaboración musical al momento con el humano.[4]

De acuerdo con Bretan y Weinberg la meta de esta colaboración musical no solamente se limita a que el robot imite o reemplace, sino que también busca enriquecer "la experiencia musical para los humanos."[5] La capacidad del aprendizaje de máquinas y su implementación en la dimensión cinética de Shimon, que le permite mover sus extremidades, cabeza y cuerpo, han ayudado a este proyecto a alcanzar un nivel de producción sonora y emocional mucho más rica que otros proyectos enfocados en el modelado físico, síntesis o reproducción sonora mediante bocinas, ya que Shimon retoma las formas de ejecución de

[3] https://www.youtube.com/watch?v=jqcoDECGde8
https://www.youtube.com/watch?v=l9OUbqWHOSk Fecha de consulta 25 de junio 2017.

[4] Mason Bretan and Gil Weinberg. A survey of robotic musicianship. *Commun. ACM*, 59(5):100–109, April 2016 p. 101

[5] Ídem

la marimba acústica. Además "los comportamientos físicos expresivos [de Shimon] que acompañan el sonido [generan] una aventura atractiva para muchos artistas"[6], ya que marcas visuales son ejecutadas por su cuerpo para ayudar a seguir el ritmo de la interpretación. De acuerdo con los autores, al integrar más sensores perceptivos en el robot sería posible tener una mejor habilidad para descifrar "eventos musicales relevantes" y así, un mayor acercamiento a la experiencia de una ejecución musical humana de la cual podría extraer nueva información generando otras formas de interacción. Varios de los objetivos de los autores apuntan a extender las relaciones entre músicos, su ejecución instrumental e incitan a ir más allá de los formatos tradicionales musicales y los modos de ejecución convencionales, empujando cada vez más nuestros límites perceptivos, motrices, interactivos e incluso creativos.[7]

> Los músicos robóticos llevan la promesa de crear música que los humanos nunca podrían crear por sí mismos e inspirar a los seres humanos a explorar nuevas y creativas experiencias musicales, inventar nuevos géneros, ampliar el virtuosismo y llevar la expresión musical y la creatividad a dominios desconocidos.[8]

Por otro lado, las funciones generativas de Shimon integran el modelado de ejecutantes humanos al extraer algunos componentes musicales como el tempo, ritmo, volumen, cromagrama (clases de altura) y altura de virtuosos músicos del jazz como Thelonious Monk y

[6] Ídem
[7] Ídem. p. 102
[8] Ídem. p. 102

John Coltrane.[9] A través del aprendizaje de frases melódicas predefinidas, Shimon escucha motivos específicos y dependiendo del modo puede imitar y sincronizar o tocar en canon creando nuevos tipos de armonías. En otro punto de la pieza Shimon comienza a cambiar los motivos utilizando procesos de decisión estocásticos y el intérprete humano tiene entonces la oportunidad de responder. Además, Shimon integra las capacidades de la visión computacional, la cual le permite a través de sensores infrarrojos predecir milésimas de segundo antes el momento en el que el baterista va a percutir un tambor, esto debido a que la punta de la baqueta del instrumentista tiene otro sensor infrarrojo el cual manda la señal al infrarrojo del robot. De este modo el sistema puede detectar el momento preciso, la velocidad, la intensidad y la localización con que la baqueta caerá sobre el tambor, datos que son usados para controlar los motivos, la dinámica y el tempo tocados por las extremidades de Shimon.

Según los resultados percibidos por los investigadores del proyecto, la presencia física de Shimon ayuda tanto al auditorio como a los músicos a entablar una relación más emotiva con el robot aportando cualidades expresivas y pistas visuales que indudablemente mejoran la comunicación con los humanos; en sus performances han encontrado mayor concentración, coordinación rítmica y sincronización por parte de los músicos. Algunas investigaciones que estudian la relación entre humano-máquina (aunque no precisamente hablen de procesos interactivos musicales) han demostrado que los efectos en la interacción mediante la presencia física del robot favorecen los procesos comunicativos, estableciendo mayores niveles de compromiso, confianza, cre-

[9]Mason Bretan and Gil Weinberg. A survey of robotic musicianship. *Commun. ACM*, 59(5):100–109, April 2016. p. 107

dibilidad e interacciones más amenas. Sin embargo Shimon no es el primer robot que integra estas características, ya desde mucho tiempo atrás los primeros autómatas generaron una ruptura con la cotidianidad al producir el encuentro en la realidad entre el humano y la máquina, como es el ejemplo del autómata capaz de soplar y digitar de forma mecánica y así, interpretar algunas melodías en la flauta, por cierto, también acústica.

La aproximación en Shimon a la improvisación es a partir de formas, instrumentos y estilos muy específicos, incluso piezas compuestas ex profeso para ser ejecutadas por ensambles junto con el robot donde todo ya está planeado para que suceda con ciertas variaciones, de manera que no hay mucho rango para la incertidumbre y lo impredecible, temas muchas veces centrales en procesos de improvisación libre. Es interesante destacar que si bien los autores mantienen un espíritu que busca expandir las posibilidades en la interacción entre humanos y entre máquina y humanos, así como expandir las formas, los géneros y los contenidos musicales, los resultados sonoros son demasiado convencionales, y están ajustados a una estética ya escuchada en múltiples ocasiones.

Si bien el gesto físico de Shimon aporta cierta información significativa para los músicos y ellos a su vez al robot dentro de un contexto musical claramente definido por un estilo, tonalidad, ritmo, etc., ¿qué resultados se obtendrían de un robot que integra estas capacidades en un contexto de improvisación libre? ¿Cómo se movería? ¿Cuales serían los movimientos que reconocería de un improvisador para extraer información significativa? ¿Qué aportaría el gesto físico del robot en la comunicación con los demás músicos en la libre improvisación? D onde frecuentemente se busca evitar un pulso común, una sincronía

en las voces, una tonalidad o hacer uso de un instrumento específico, la información visual—en muchas ocasiones—es igual de ambigua que la sonora y requeriríamos de altos niveles de abstracción para poder extraer datos significativos que realmente aporten al proceso comunicativo entre músicos y máquinas. A mi modo de ver valdría la pena hacer este esfuerzo en la improvisación libre solo si se buscara que la máquina tocara algún instrumento acústico, aunque como veremos en el siguiente caso de OMax y FILTER el software en abstracto puede emular y apropiarse de las cualidades acústicas de un instrumento sin necesariamente tener un cuerpo que le permita tocarlo. Aunque en estos dos casos los sistemas graban segmentos de improvisación que después reproducen a través de bocinas, sería difícil compararlos con el potencial que un instrumento acústico podría aportar al performance. Además, las exploraciones técnicas que devienen en el resultado sonoro muchas veces se alejan de la forma convencional de aproximarse a un instrumento. Debido a esto sería sumamente difícil desarrollar una técnica lo suficientemente versátil y adaptable como para poder interactuar con otros improvisadores de manera física.

Por otro lado, en algunos de los performances que pude ver sobre Shimon en un concierto de máquinas que escuchan, pareciera por las descripciones que manejan los autores de los performances que la idea no es comprobar el estado de avance científico-tecnológico para mostrar una máquina súper capaz que pueda interactuar con el humano, sino más bien señalar la tensión que existe entre lo que sucede realmente y lo que uno esperaría al interactuar con la máquina. Así mismo se cuestionan la idea de si es posible que la máquina pueda transmitir emociones a través de la inteligencia artificial.

> El concepto del performance es la relación amor-odio entre máquinas y humanos. Los robots también pueden mostrar emociones y el performance será sobre [...] la lucha y la frustración que enfrentan los humanos al lidiar con las máquinas. La pieza proporcionará fundamento para pensar si las máquinas pueden mostrar emociones usando inteligencia artificial, la interacción de las máquinas con los humanos generará relaciones interesantes."[10]

Algunas exploraciones más bien buscan entender y cuestionarse la relación humano/máquina, por ejemplo desde el reconocimiento del otro en el espacio.

> La tecnología genera tanto esperanzas como inquietudes, especialmente en esta época en que se vuelve cada vez más penetrante. *The Beginning of the End* explora este complejo fenómeno social que enfatiza las diferentes etapas de interacción entre los humanos y la máquina: encuentro inicial, curiosidad mutua, colaboración e independencia. Estas fases serán examinadas en esta actuación dramática a través del reconocimiento de gestos y la improvisación interactiva. La estrella de la obra es el jugador robótico de marimba, Shimon, que improvisará para su entretenimiento en un viaje que lo dejará sin aliento."[11]

La relación emocional que el humano puede generar con la máquina o la interfaz es una realidad inevitable ya que emocionalmente estos

[10] https://www.youtube.com/watch?v=ZXDpDmAF53A Fecha de consulta 25 de junio 2017.

[11] Ídem.

artefactos nos afectan de múltiples formas. En el caso de Shimon su interfaz corporal ayuda a los intérpretes a intuir comportamientos del robot para interactuar con él, y al público le puede dar pistas de cuál es el pulso que lleva a través del movimiento de su cabeza para generar cierta empatía visual con el robot. Sin embargo, el tema de la frustración y el tener que lidiar con la máquina para obtener una respuesta más o menos adecuada o incluso más humana sigue presente. Ya sea porque su respuesta es más lenta de lo que esperamos, o porque no se adecúa a lo que buscamos en un momento determinado. Comparando lo anterior con nuestras relaciones humanas, en ellas también ocurren estas tensiones. Imaginemos dos o más mentes que funcionan y piensan distinto e interactúan en un sistema complejo, donde lo que uno espera del otro humano muchas veces no sucede, o sucede de formas inesperadas y naturales o de formas intrincadas y forzadas. En este sentido la máquina podría asemejarse más de lo que pensamos a un humano si logra generar una frustración comunicativa entre ambas partes. La máquina muestra dichos comportamientos debido a su funcionamiento precario, limitado o poco responsivo, logrando generar una frustración en la comunicación entre las partes involucradas. Pero, si lograra romper esa barrera comunicativa y pudiera responder de manera oportuna u óptima, ¿nos confrontaría con nosotros mismos? ¿Intentaríamos superarla? ¿Cuestionaríamos nuestros propios hábitos comunicativos, para transformarlos? E incuso ¿Podríamos superar los vicios al comunicarnos entre humanos? Son algunas preguntas sin respuesta que necesariamente implican una negociación con una cosa que no es una persona pero que se acerca a ello a través de nuestras aspiraciones. Mientras tanto, es interesante pensar en estas máquinas como instrumentos ineficientes, ineficaces e improductivos, en el sen-

tido de contradecir la idea de eficacia que supuestamente prometen las máquinas, en favor de una desviación que pueda evocar experiencias estéticas otras, lejos de las elucubraciones sobre el progreso y la eficiencia tecno-científica.

2.2. OMax

OMax es un programa diseñado y desarrollado en el IRCAM (Instituto de Investigación y Coordinación Acústica/Musical, en francés Institut de Recherche et de Coordination Acoustique/Musique,) por Gérard Assayag y Shlomo Dubnov en 2004, el cual aprende al momento secuencias musicales de longitud variable a partir de escuchar improvisaciones libres ejecutadas al momento. De esta escucha extrae información para modelarla y crear una estructura formal compleja (frases musicales), después navega por esa estructura reorganizando los materiales sonoros que son grabados para crear cánones, variaciones y clones que interactúan con el improvisador. OMax integra el algoritmo Factor Oracle propuesto por Cyril Allauzen y Maxime Crochemore en 1999, este algoritmo es capaz de encontrar estructuras dentro de textos y en el ámbito musical puede generar una correspondencia bastante compleja de patrones musicales. OMax se caracteriza por la acumulación de gestos/frases musicales o sonoras y su identificación, delimitadas generalmente por los silencios en la ejecución del improvisador.

El sistema puede trabajar tanto con entradas de audio (versión de paga) como con entradas MIDI (versión gratuita) y puede generar ambas salidas respectivamente. OMax no utiliza una base de datos precargada, y tampoco deduce parámetros de la señal de audio; de

hecho, sus autores lo definen como un sistema agnóstico el cual puede adaptarse fácilmente a diferentes contextos musicales, debido a su programación interna .[12]

2.2.1. Experiencia con OMax

En su versión MIDI OMax presenta cuatro módulos generales: la entrada MIDI, el procesamiento de los datos, la salida y la visualización de los motivos ordenados por frases que van siendo grabados y después reproducidos. Las pruebas que realicé las hice con un piano eléctrico como entrada, mientras OMax mandaba directamente sus procesamientos de regreso al mismo piano, de manera que todo el audio saliera por la misma fuente.

Mi primera impresión fue encontrarme con un clon de lo que estaba tocando, ya que la interacción del sistema parte de la noción de repetición, por momentos literal, en otros, con ligeras variaciones pero finalmente bastante similar. Al seguir improvisando me encontré con que el sistema puede elegir de forma seudo aleatoria qué fragmentos reproducir, generando una suerte de memoria muy detallada de los gestos tocados que inmediatamente son almacenados en la memoria del sistema. Además, al sistema no le importa en qué estilo se esté tocando, simplemente graba e inmediatamente se adapta a eso, por lo que su capacidad de adaptación y versatilidad es muy veloz. Hablar de esto inevitablemente me remite nuevamente al tema de la inteligencia artificial como una construcción basada en la ilusión o el engaño, es como un acto mágico que simula la interacción y la respuesta para parecerse a las de un humano, es decir, que el sistema

[12]B. Levy. *OMax The Software Improviser*, 2004-2012

pareciera reaccionar de forma "inteligente" y pasa sin problemas la prueba de Turing; evidentemente, la pasa debido a que OMax reproduce y transforma al momento lo que yo había tocado antes. En este sentido es sumamente interesante que este tema de la ilusión siga estando presente en implementaciones de máquinas que improvisan en el siglo XXI al igual que en los autómatas del siglo XVI.

Siguiendo con la experiencia de probar OMax, en un principio encontré bastante retador interactuar con el sistema en su versión MIDI debido a esa suerte de memoria impecable capaz de reproducir mis propios gestos interpretados con anterioridad integrando ligeras variaciones. Estas van de modificaciones en alturas a cambios en la velocidad de reproducción y alteraciones rítmicas. Como ruidista y guitarrista, fue un reto para mí probar esta versión del programa, primero por los límites diatónicos que impone el teclado, del cual se extraen la altura y velocidad dinámica para ser usados como respuesta del sistema. Por otra parte, en la configuración que usé para tocar pareciera que uno mismo estuviera tocando esos gestos nuevamente. Es como poder acceder a una memoria personal privilegiada que recuerda todas y cada una de las teclas y gestos tocados. Además la similitud con que reproduce los fragmentos en el orden en que fueron tocados y la linealidad de la reproducción es tal que por momentos se crea una suerte de canon. Cuando introducía un nuevo material, el sistema regresaba a usar los primeros fragmentos con los que yo había comenzado a improvisar. Esto evidentemente va empujando poco a poco la improvisación a una interacción y forma estructural que puede ser predecible después de un tiempo y por momentos algo cansada

debido a la repetición de los gestos.[13] Además, el patch de Max/msp tiene la posibilidad de elegir entre tres modos distintos de improvisar, cada uno con distintas propiedades que pueden hacer un poco más rica la interacción si estos y sus diferentes parámetros son manipulados por un segundo ejecutante. Las posibilidades de asignar distintas velocidades de reproducción a cada uno de los modos de improvisación, de activarlos y desactivarlos, alterar las transposiciones al momento, cambiar aleatoriamente el lugar de reproducción o asignar las salidas a distintos timbres fueron elementos que indudablemente enriquecieron el panorama reactivo en la interacción del sistema.

Tuve la oportunidad de probar lo anterior con la pianista Rossana Lara Velazquez quien en un principio se mostró muy optimista con los resultados de la interacción, percibiendo un resultado "orgánico", dado que el programa captaba bien los gestos y se mantenía dentro del mismo estilo, acorde a lo que estaba tocando; no había, pues, una respuesta por parte del sistema que estuviera fuera de lugar. Una de las críticas que realizó fue que en algún momento ella no tenía ya qué tocar debido a que el sistema tiende a acumular gestos de forma lineal sin que el improvisador pueda controlar la acumulación y densidad de los materiales, de modo que después de unos seis minutos deja de existir el diálogo inicial, tornando el proceso en un monólogo de la máquina, que toma el control y se convierte en el protagonista de la improvisación. Incluso desde el sistema es imposible controlar el nivel de saturación. Por otra parte, no hay nada que el mismo sistema proponga por sí mismo respecto a lo cual uno pueda reaccionar. OMax tampoco puede regresar a las primeras partes de la improvisación pa-

[13] https://www.youtube.com/watch?v=4DlquqMNxTU, https://www.youtube.com/watch?v=zm5td1aSsug Fecha de consulta 20 de octubre 2017.

ra tocar materiales previos e intentar proponer nuevas tendencias en la improvisación, por lo que podría entenderse como un sistema lineal y cerrado. Lineal por su predictibilidad interactiva y cerrado porque no puede proveer o aportar conocimiento nuevo en términos audibles. La máquina no establece en ningún momento un rol de interacción y solo comienza a lanzar gestos automáticamente dejando de lado el intercambio que podría producirse en una improvisación y la escucha que emerge de las nuevas acciones e incuso de las suyas propias. El sistema finalmente impone su forma de operar, por tanto, la manera en la que uno podría interactuar está completamente mediada por su programación, herramientas interactivas y formas de reacción limitadas. Lo verdaderamente interesante en un sistema de improvisación de este tipo sería su capacidad de proposición y adaptación a distintas situaciones, procurando tomar un papel proactivo a través de la escucha consciente y no solo reactivo que supone una escucha basada en la ilusión. Con ello podría generarse una provocación que active ideas más allá de lo que un improvisador conozca o pueda proponer.

Además, como señalan Schankler, Smith, François y Chew en su artículo sobre *Factor Oracle*, el algoritmo de reconocimiento de frases musicales en OMax podría presentar características intrínsecas a su programación que predisponen al algoritmo a obtener resultados estructurales particulares en su análisis: independientemente de lo que toque el intérprete, el sistema termina por imponer sus propias formas de interpretar la información de entrada y por tanto sus formas de improvisación que la mayoría de las veces serán muy similares.[14]

[14]Isaac Schankler, Jordan B. L. Smith, Alexandre R. J. François, and Elaine Chew. *Emergent Formal Structures of Factor Oracle-Driven Musical Improvisations*, pages 241–254. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-21590-2

> Postulamos que estos patrones estructurales, si se observan consistentemente, podrían ser propiedades emergentes del comportamiento del factor Oracle [...], algunas de las estructuras formales observadas en las improvisaciones grabadas [...] surgen, al menos en parte, del comportamiento [...] inherente [del algoritmo]. [15]

Por otra parte, en *performances* que he encontrado con OMax en su versión de procesamiento de audio (en vez de la versión MIDI), encuentro varias ventajas al improvisar. El sistema muestra comportamientos más coherentes y una capacidad de adaptación mayor al contexto sonoro. Esto debido precisamente a la capacidad de grabación y reproducción del sistema y la interacción que se genera, en uno de los casos, entre los tres agentes que construyen la improvisación; improvisador (fagot), controlador humano (del sistema) y OMax.[16] Además la ligera alteración y transformación del sonido grabado permite conservar cierta flexibilidad en el sistema. De la reacción del sistema emergen tendencias sonoras y estilísticas que se "adaptan" de forma automáticamente inconsciente al proceso mismo de improvisación generado por el fagotista. El sistema no tiene consciencia alguna de lo que está ocurriendo, pero mágicamente la interacción global deja ver comportamientos bastante coherentes por parte de los tres agentes involucrados. Pese a sus limitantes, OMax privilegia ciertas características interactivas sobre de otras; por un lado la apertura técnica que podría tener hacia la libre improvisación, pero por otro lado, se olvida completamente de la estructura global de la misma y

[15]

[16] https://www.youtube.com/watch?v=pojhhJN1ySE. Fecha de consulta 20 de octubre 2017.

de la capacidad de escuchar activamente al otro para establecer un diálogo que le permita adoptar otros roles de interacción dentro de la improvisación.

2.3. GREIS

¿Qué ventajas tendría improvisar con un sistema que distingue tendencias sonoras durante el *performance* frente a un sistema previamente entrenado sobre un corpus musical o sonoro (como Shimon)? Partiendo de esta pregunta introduzco las ventajas que encontraron los creadores de GREIS (Sistema de Instrumentos Expandidos con Retroalimentación Granular), el cual aprende en vivo a través de una escucha humana prestada durante performances de improvisación libre.

> Si bien ciertamente consideramos que el pre-entrenamiento es una dirección importante e interesante a tomar [en la creación] de sistemas interactivos [...], hasta la fecha nos hemos centrado en sistemas que aprenden tendencias sonoras y estilísticas durante el performance para privilegiar la naturaleza abierta de la improvisación completamente libre, así como para ver hasta qué punto se puede llegar con este enfoque.[17]

GREIS fue generado de forma colaborativa por Pauline Oliveros, Jonas Braasch y Dough Van Nort en 2013. Cada uno aportó lo más

[17] Doug Van Nort, Pauline Oliveros, and Jonas Braasch. Electro/acoustic improvisation and deeply listening machines. *Journal of New Music Research*, 42 (4):303–324, 2013

significativo de su campo para crear un sistema que improvisara y pudiera adaptarse a través de una escucha profunda a diferentes contextos de la libre improvisación. El sistema no usa la tecnología de máquinas que aprenden, en cambio es una simbiosis entre sistema y performer, quien funge como "los oídos" de la máquina, éste tiene que detectar y grabar en vivo momentos que considera interesantes de la improvisación para que después sean almacenados en la memoria del sistema. Por tanto, la máquina no escucha, ni reconoce contextos, sino que requiere de un intérprete que escuche por ella y pueda anticipar ciertas acciones sonoras/musicales de los otros improvisadores y del mismo sistema para ayudar a conducir el performance hacia lugares interesantes. GREIS puede reaccionar y tomar decisiones en tiempo-real de performances de improvisadores libres partiendo del análisis y mapeo espacial de las muestras de audio almacenadas que posteriormente serán moduladas por la máquina y regresadas en forma sonora a los improvisadores.

Esta memoria capturada se puede mantener de forma indefinida o hasta que sea borrada de forma intencional, permitiendo moldear el contenido de audio almacenado también de forma manual por el intérprete. "El intérprete tiene un control refinado de las inflexiones gestuales pasadas, conduciendo [la improvisación] hacia un diálogo con intenciones gestuales pasadas y futuras, así como con variantes introducidas por la máquina [...] Esta propagación de la intención gestual es clave para que el sistema sea capaz de definir una estructura musical cohesiva que fluya orgánicamente [...]".[18] El sistema tiene una memoria episódica según sus autores, esta es capaz de registrar, además de audio, parámetros de control que después pueden usar-

[18] Ídem

se para representar y modular el contenido sonoro producido por el sistema, cambiando con esto las intenciones musicales y las modulaciones del intérprete hacia el sistema y viceversa. Además GREIS agrega al performance elementos indeterminados producto del ruido en la transmisión de la información generado por la captación, análisis, transformación y reproducción sonora, agregando a las improvisaciones, variantes completamente insospechadas. Además,

> Con el fin de mantener el carácter sorpresivo en la improvisación, el sistema de instrumento expandido [EIS, GREIS], cuenta con una capa de control de alto nivel de aleatoriedad dirigida, introduciendo ruido en los parámetros de procesamiento sonoro, así como la matriz de enrutamiento y filtrado por grano, mientras que el comportamiento de las líneas de retardo moduladas pueden controlarse de forma probabilística. [19] [20]

Como instrumento expandido GREIS tiene muchas posibilidades y ventajas que pueden ser usadas en el performance; sin embargo, como se ha mencionado, no cuenta con un sistema que le permita escuchar y adaptar su escucha de forma inteligente a los distintos escenarios de la improvisación. Justamente estas limitantes son las que posibilitaron crear a su sucesor FILTER, "el cual puede concebirse como una expansión del instrumento expandido en el territorio de la

[19] Ídem

[20] Al hablar de la matriz de enrutamiento los autores se refieren al espacio en donde están almacenadas las muestras de sonido, así mismo el contenido de audio es es dispuesto para ser procesado procesado por síntesis granular o filtrado por grano.

escucha de máquinas y una aproximación evolutiva hacia la creación de salidas musicales".[21]

2.4. FILTER

Para sus autores, FILTER se puede amalgamar de manera sorpresiva con las intenciones de los músicos debido al grado de interconectividad generada. De manera tal que por momentos resulta difícil saber quién es el que realmente está tocando, si la máquina o los músicos. Incluso en ensambles de libre improvisación de más de cuatro o cinco músicos, comienza a dificultarse reconocer quién está tocando qué sonido(s). Las casi infinitas posibilidades de combinatoria tímbricas entre un ensamble, además de las exploraciones técnicas usadas, vuelven sumamente difícil poder reconocer con claridad los materiales sonoros, e incluso los generados por los propios músicos. Al escuchar algunas improvisaciones realizadas con FILTER, por momentos uno puede dudar quién está produciendo un determinado ruido dentro del caótico mar sonoro que puede llegar a ocurrir en la improvisación.[22] Pauline Oliveros, Jonas Braasch y Dough Van Nort comentan:

> Como resultado del intercambio de señales sonoras, a veces, el resultado son tres gestos sonoros distintos con cualidades tímbricas únicas; otras veces estas formas gestuales pueden tener timbres muy similares. En otros momentos el trío se fusiona en una sola línea que se mueve en un movimiento coherente y da como resultado una textura sonora

[21] Ídem

[22] http://dvntsea.weebly.com/sound.html Fecha de consulta 15 de octubre 2017.

> que se sostiene sin dirección lineal. Las decisiones musicales improvisadas en el performance se guían puramente por la escucha focal y global en el momento, y ambas son conscientes del movimiento entre la dinámica gestual y el sostenimiento textural.[23]

Al igual que GREIS, FILTER graba materiales tocados por los improvisadores, lo cual resulta en una amalgama de sonidos inevitable e inherente a su programación. Determinar a partir de la escucha las acciones concretas de cada músico incluyendo las de los dos sistemas,[24] no sería una tarea fácil para una escucha normal, por más profunda que esta sea, y seguramente tuvieron que hacer varias modificaciones al sistema ya que en algunos momentos los autores buscaban tener claras las intenciones y acciones del humano, y diferenciar las de la máquina.

FILTER también registra la evolución temporal de los componentes sonoros analizados, con esto genera el nivel semántico y estructural de su memoria la cual contiene un conjunto de materiales sonoros dispuestos en frases, generalmente divididos por silencios o respiraciones. Estas características de aprendizaje e incluso la forma de interactuar con otros músicos están mediadas por la forma en que fue "enseñada" a escuchar.

Del análisis correspondiente de los componentes extraídos de cada gesto, FILTER aprende con un modelo de aprendizaje sin supervisión,

[23] Doug Van Nort, Pauline Oliveros, and Jonas Braasch. Electro/acoustic improvisation and deeply listening machines. *Journal of New Music Research*, 42 (4):303–324, 2013

[24] La formación del trío con la cual probaban, ensayaban y se presentaban en vivo incluía a Pauline Oliveros acordeón, Jonas Braasch saxofón, Dough Van Nort manipulando GREIS y FILTER funcionando de forma completamente autónoma.

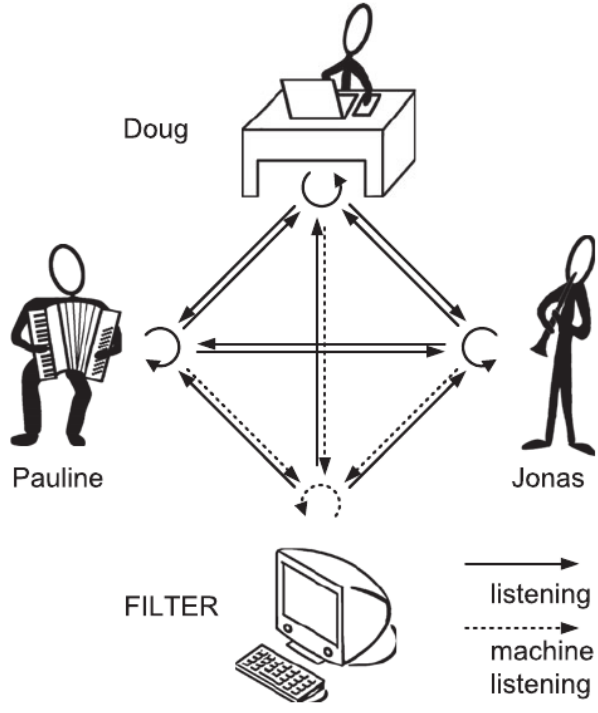

Figura 2.1: Sistema de improvisación basado en la escucha profunda

el cual agrupa de forma automática los diferentes sonidos introducidos en su sistema y los organiza por clases. Cuando un cambio de amplitud o frecuencia es detectado, sujeto a una compuerta temporal de ataques (onset threshold), se considera que un nuevo gesto o frase ha comenzado. Si el gesto detectado es diferente a los que están dentro del banco de gestos, será incluido como un nuevo miembro del banco sonoro. Este tipo de comportamiento es lo que los autores consideran como el nivel semántico de la memoria del sistema, capaz de representar objetos abstractos de la improvisación, con este nivel compara entre su banco de sonidos y todos los sonidos entrantes, su función específica es encontrar dichos gestos sonoros relevantes al momento de la improvisación. La "memoria semántica" de FILTER no tiene ningún tipo de ordenamiento temporal entre los gestos, es decir que estos se clasifican indistintamente del momento en el que suceden, a este respecto podría considerarse que la memoria semántica se encuentra fuera de tiempo. Esta podría ser una ventaja del algoritmo de aprendizaje sin supervisión que para fines de interacción en la improvisación libre podría resultar adecuado, si pensamos en conservar tipos de materiales similares pero que no necesariamente ocurren inmediatamente después de ser tocados sino en el pasado o en el futuro.

Además en una escala temporal más amplia, el sistema realiza cortes entre segmentos que tienen diferencias sustanciales texturales y tímbricas, de ellos extrae un promedio de los parámetros tocados al mismo tiempo por todos los intérpretes, este promedio será traducido como la textura general del ensamble y es entendida como la memoria estructural del sistema. Esta memoria será utilizada en momentos en donde el algoritmo de reconocimiento no alcance el puntaje requerido para detectar de forma certera gestos musicales individuales y

accionar sus modos de reacción. Sus reacciones son el resultado del análisis, de esa escucha profunda artificial conceptualizada, entrenada y programada por los autores.

Con las cualidades anteriores, y como se ha mencionado, FILTER compara los segmentos de entrada con los segmentos gestuales contenidos en su banco de datos, "proveyendo un vector de posibilidades para cada miembro [sonoro almacenado]. Mientras un sistema de clasificación estándar miraría hacia el miembro más parecido, "nosotros examinamos la probabilidad completa de los vectores y cómo cambia su forma dinámica en el tiempo."[25] De este modo podría decirse que FILTER cuenta con un sistema de atención dinámica, la cual identifica al momento de la improvisación ciertos arquetipos gestuales y texturales comparándolos con los almacenados en su memoria.

La metodología usada en la fase de escucha fue grabar gestos en dos escalas temporales una amplia que distingue la textura y estructura general de la improvisación y una escala temporal corta que distingue entre gestos musicales y sonoros, después se entrenó al sistema para reconocerlos al ser tocados nuevamente. El sistema especifica el grado de similaridad y certeza con que fue reconocido un gesto, posteriormente, ciertos rangos de certeza en el reconocimiento de las frases le permitirán elegir su respuesta de entre un banco de acciones posibles. Pero si el sistema no logra identificar con claridad los gestos y llegar al rango impuesto por la programación, el sistema focalizará su atención en una escucha de la textura global del ensamble. Esta misma forma de escucha textural también puede ser clasificada en es-

[25]Doug Van Nort, Pauline Oliveros, and Jonas Braasch. Electro/acoustic improvisation and deeply listening machines. *Journal of New Music Research*, 42 (4):303–324, 2013

cenas pre-programadas usando un modelado de sistemas dinámicos, implementado por Van Nort et al. en 2012. De manera que FILTER puede discernir entre distintos tipos de texturas musicales, y de igual manera asignar un grado de certeza al identificar texturas arquetípicas almacenadas en su memoria. Esta memoria considerada como episódica usa el procedimiento de Factor Oracle que anteriormente fue usado en OMax y posteriormente en otro sistema llamado MIMI.

FILTER reacciona partiendo de distinguir material muy similar o muy diferente de acuerdo con los materiales almacenados en su memoria, es decir, sus reacciones están reguladas por los gestos relativamente más sobresalientes del contexto, a lo cual correspondes acciones asignadas de acuerdo al modo de escucha del sistema. Asimismo, FILTER puede hacer estiramientos temporales y transposiciones, con lo cual puede tener un abanico muy amplio de recombinatoria para adaptarse a los diferentes momentos de la improvisación.

> En esta implementación, un conjunto de miembros de la población de N-dimensiones se mueven dentro de un complejo simple *(simplicial complex)*[26] donde cada nodo está asociado con un estado de comportamiento del sistema. El miembro de la población asociado con el estado gestual/textural más destacado, se utiliza para interpolar los nodos del complejo simple que lo rodea, determinando un estado de comportamiento de salida del sistema. Esto puede considerarse una nube de estados posibles que se mue-

[26]En matemáticas un complejo simple es un conjunto de lineas, segmentos, triángulos y su contraparte de N-dimensiones.

> ven con una naturaleza casi física tal como lo determina la salida del módulo de escucha.[27]

Las acciones que el sistema puede realizar dependiendo de lo que escucha son las siguientes:

- **Cualidad rítmica** mayor probabilidad de que los fragmentos reconocidos sean repetidos, así como variaciones sobre la repetición.
- **Cualidad salvaje** aumenta la probabilidad de que el sistema imite al improvisador, usando material reciente y pasado de forma caótica.
- **Estabilidad** mayor probabilidad de que el sistema cambie súbitamente sus estados de comportamiento.
- **Sostenimiento (sustain)** favorece los sonidos tenidos.
- **Densidad** dependiendo la densidad el sistema cambia entre duración y mayor o menor espaciamiento en las frases de salida.

Por último, cabe señalar que FILTER es un antecedente directo del proyecto planteado en esta tesis por varias razones: en términos sonoros hay una empatía estética debido a su búsqueda dentro de la música libremente improvisada, particularmente desde la noción de escucha profunda planteada por Oliveros. También es significativa la importancia que tiene FILTER en términos técnicos ya que su sistema de discernimiento gestual basado en el aprendizaje supervisado resultó útil para una de las aproximaciones que planteo para el

[27] Doug Van Nort, Pauline Oliveros, and Jonas Braasch. Electro/acoustic improvisation and deeply listening machines. *Journal of New Music Research*, 42 (4):303–324, 2013

reconocimiento de timbres en la libre improvisación. Sin embargo, resulta muy distinta a otra aproximación explorada en la tesis, basada en el aprendizaje sin supervisión usando la técnica de agupamiento (clustering) con el algoritmo K-Means, descrito en el apartado 3.2.2 y 4.2.7. Por sus características y además por los resultados sonoros escuchados, FILTER fue una gran motivación, ya que me permitió seguir trabajando con el sistema de escucha, perfeccionarlo y mantener abierta la posibilidad de llevar el sistema de escucha a la práctica performática a través de un sistema de reacciones.

2.5. Sonic-Mirror

Scott Toby artista y músico, programó en 2016 *SCML SuperCollider Machine Listening* que integra algoritmos de máquinas que aprenden directamente con la plataforma Wekinator. Scott explora las capacidades de Wekinator a partir del reconocimiento de vocales y silbidos los cuales son analizados por el descriptor de audio MFCC y el reconocimiento de altura ambos implementaciones de SuperCollider, estos datos son enviados via OSC a Wekinator donde el algoritmo KNN se encarga de clasificarlos. Supercollider luego recibe de vuelta esta información de Wekinator via OSC, la instrucción de silbido activa un procesador de audio que modula, al mismo tiempo que es producido, la altura del sonido emitido. Dependiendo de si reconoce un silbido o una vocal el sistema activa diferentes modos de procesamiento variando parámetros como la altura y la disposición de izquierda a derecha del sonido.

El mismo autor programó en el mismo año *Sonic Mirror* en la plataforma de SuperCollider. Ésta es una máquina que escucha (sin

hacer uso de algoritmos de máquinas que aprenden), graba y reproduce lo escuchado a través de procesamientos de sonido en vivo. Sonic Mirror esta basado en la activación de mecanismos aleatorios para lograr un auto-comportamiento y modular distintos parámetros de la grabación al momento de ser reproducida. Por momentos Sonic Mirror pareciera tener rasgos de inteligencia artificial en sus modos de operar ya que puede ser bastante asertivo y lograr una seudo-comunicación con el improvisador. Los temas del engaño, la ilusión, la simulación y la magia —bastante explorados en las primeras incursiones de la historia de los autómatas—, están presentes en este proyecto. La suerte o la impredecibilidad del comportamiento al momento del *performance* atribuyen al sistema dichas cualidades seudo-inteligentes, basadas en el audio grabado, modificado y transformado.

Resulta importante este tema ya que para hacer una máquina que improvise libremente y que por momentos decida hacer mención a algo ocurrido en el pasado o imitar ciertas frases o ideas sonoras, necesitaría jugar un poco con estos modos de comportamiento. En este caso los elementos de incertidumbre y azar presentes en la improvisación vuelven a ser elementos accesibles para el sistema, aunque de una forma controlada obviamente por la máquina. Pese a que sean generadores aleatorios, para nuestra percepción tal vez sea así, pero en realidad esos generadores computacionales tienen patrones periódicos que cada cierto tiempo se van repitiendo.

Hoy nos encontramos con múltiples iniciativas de proyectos de creación sonora, musical y generación automática de lenguaje que si bien no los podría considerar como trabajos artísticos ni "obras acabadas", son trabajos en desarrollo que están dispuestos para echar a andar proyectos más grandes o sencillamente para aprender la lógica

y las posibilidades del funcionamiento de las máquinas que aprenden. Este tipo de proyectos (algunos académicos, algunos desarrollados dentro de la comunidad digital, después apropiados por los grandes consorcios corporativos como o alguna compañía nueva de reciente formación) tienen el potencial de seguir creciendo debido a que mantienen un perfil de desarrollo abierto, de acceso libre y son potencialmente modificables a las necesidades de cada persona. Son abiertos porque comparten y explican de manera amplia (aunque algunos no tanto) lo que han desarrollado aunque no sea un proyecto terminado ni definitivo, además de que el software para desarrollar el código esta construido con software libre en su mayoría, generalmente basados en Python y sus respectivas librerías (muchas de ellas también libres y abiertas a la modificación). Algunos de estos proyectos son compartidos en plataformas e trabajo colaborativas y descentralizadas como GitHub a través de redes sociales tales como blogs, páginas de internet personales, artículos técnicos y/o videos en Youtube o Vimeo.

2.6. Comentarios finales del capítulo

Cómo podríamos evaluar sistemas como los descritos en este capítulo en un contexto como la improvisación libre, donde podría incluso no haber instrumentos, ni formas plenamente estructuradas para tocar, donde muchas veces las resultantes sonoras no son lo que esperamos y a una acción no corresponde necesariamente una sola reacción todas las veces, donde suele no existir reglas explícitas de cómo y cuándo interactuar, y si las llega a haber en algunos casos, éstas se traducen como límites que ayudan a focalizar la atención y la escucha. Haciendo a un lado los cuestionamientos de si el sistema funciona bien o mal

en un sentido estético del sonido, ¿cuáles serían los criterios tomados en cuenta para ir más allá y evaluar un sistema como FILTER o cualquier otro sistema interactivo? Además ¿qué es lo que realmente se está buscando al emular la realidad y a nosotros mismos?

Algunos de los criterios a considerar para evaluar un sistema interactivo musical podrían ser el grado de flexibilidad del sistema en diferentes contextos; es decir, cómo el sistema se podría *adaptar* a diversas situaciones de improvisaciones libres e incluso no libres. Aquí vale la pena señalar que la adaptación de un sistema interactivo al momento de la improvisación no implica necesariamente que este escuche, aunque evidentemente dicha aproximación tendrá sus limitantes. OMax o Sonic-Mirror son un ejemplo de ello, ya que el sistema no emplea algoritmos de aprendizaje automático sino que graba y reproduce una secuencia del músico o improvisador encadenada a otras que se van acumulando sin establecer un código dialógico activo que parta de la escucha con éste, sino más bien se genera una ilusión comunicativa. Romper con esa ilusión sería un paso importante para el desarrollo de este tipo de sistemas ya que en la libre improvisación se busca generar una apreciación estética de la escucha, hacer sonar el mundo a través de la escucha profunda que señala Pauline Oliveros. La generación de sonido al momento parte de una escucha atenta, sin ella no se abrirían los canales de diálogo con el otro ni hacia otros sonidos que circundan el espacio, que a su vez son el otro y constituyen los sonidos orgánicos de uno mismo.

Otro criterio podría ser la capacidad para producir espacios inmersivos que incluyan luz, movimiento, sonido, y que sean aptos para

despertar emociones en nosotros. Ejemplo de ello es el N-Polytope de Chris Salter.[28]

Otra característica podría ser la capacidad para cuestionar la práctica interactiva, creativa y los hábitos dentro de la improvisación libre a través de la maduración técnico-creativa de un sistema. Por ejemplo, mientras más información pueda almacenar y procesar la máquina, y mientras más desarrollados estén sus "sentidos" para analizar esa información al momento, más precisa y completa puede ser su apreciación del entorno. Sus instrumentos perceptivos serían útiles para reconocer diferentes porciones de realidad y con ello abrirían la posibilidad de conocer nuevas perspectivas de interacción y creación sonora. Además, por medio de la emergencia de una suerte de espontaneidad interactiva, el sistema sería capaz de cuestionar los hábitos empleados por improvisadores, el público o ambos.[29]

Si la máquina retuviera nuestros recuerdos «audio grabado», nuestras memorias «técnicas musicales», nuestra apreciación del mundo

[28]N-Polytope cuenta con 150 LEDs de 10 watts con sensores fotoeléctricos y 50 bocinas colocadas en cables de acero y aluminio de alrededor de 24 metros que forman una red compleja suspendida en el aire. Tanto las bocinas como los LEDs son controlados por un software en red basado en módulos y sensores inalámbricos (Xbee http://sensestage.eu). Mientras los LEDs crean un espacio dinámico de destellos luminosos, lasers de colores rojo y verde rebotan en superficies de espejos en fijos y en movimiento, éstos forman líneas y figuras que desaparecen y reaparecen en el espacio, creando una arquitectura completamente inestable y efímera.

[29]Esto podría ser contrario a lo que sucede con nosotros los humanos, ya que la sobre-estimulación informática resulta una limitante enorme para incentivar una concentración atenta y focalizada de la escucha. Esta puede estar bloqueada por el desbordamiento informático que vivimos actualmente, donde lo que se escucha ya no se entiende como un acto poético sino más bien como ese ruido que distrae, que enajena, que engaña y ensordece.

«escucha profunda», para generar una especie de clon sonoro de nosotros (aludo a la idea del clon gestual en OMax), sería posible encontrarse con el desdoblamiento de uno mismo, el de otros y a su vez ser encontrados por otros a través del sonido para producir nuevas experiencias y perspectivas de interacción y creación sonora. Sin embargo, la tecno-ciencia actual, en favor de una apropiación general del mundo y el control de sus condiciones naturales, incluido el cuerpo y la subjetividad humanas, encarna la idea de que la máquina es infalible y sus cualidades suficientes para desempeñar funciones similares a las características humanas e incluso sustituirlas; entablar una conversación o un intercambio sonoro o de cualquier otro tipo con una máquina de maneras más fluidas, precisas, acertadas podría ser posible en algunos años, y si esto ocurriese también podrían diluir y apropiarse de las posibilidades de azar e impredecibilidad de la vida. En este sentido ¿Cómo sería posible dar lugar a ese encuentro constructivo con la máquina y no al revés, la máquina al servicio de los poderes fácticos de dominación y control?

Finalmente, lograr que un sistema prime el diálogo interactivo de forma que sea imposible discernir si ella o el humano están improvisando, es una realidad como hemos visto en estos proyectos, los cuales son capaces de sobrepasar la prueba de Turing al mostrar comportamientos incluso creativos o simulaciones de una escucha inteligente. Sin embargo esto aún se encuentra en el terreno de la ilusión a través de la reproductibilidad de rasgos culturales, de nuestras formas de relacionarnos con otros, de comunicarnos con otros, de pensar y de escuchar el mundo. Aunque esa ilusión o ese engaño puede ser tan fiel como lo que entendemos por real, ya no hay una separación irreconciliable, más bien todo el tiempo se está conciliando esa relación con

las máquinas y los aparatos inteligentes que usamos día a día. En este momento generar un tipo de máquina consciente de sus actos y sus acciones sigue siendo una idea prometedora; más bien, esa máquina "consciente" es encarnada por los sistemas políticos dominantes encargados de mantener un estado de enajenación en muchos ámbitos sociales a través del uso de herramientas lo suficientemente poderosas para mediar relación humano-máquina.

Capítulo 3

Máquinas que escuchan y aprenden: Marco teórico-funcional

Los sistemas de reconocimiento de propiedades sonoras como detectores de ritmo, alturas, seguimiento de tonalidad, etc., comúnmente usados para analizar, clasificar y crear música basada en ritmos más o menos estables y con tonalidades específicas, son insuficientes para aproximarse al análisis de texturas sonoras más complejas tales como las generadas en la libre improvisación. Los principales parámetros que describen algunas de las aproximaciones a la libre improvisación son las cualidades espectrales del sonido; el timbre, la densidad sonora, la amplitud y por otro lado, las formas de interacción entre los integrantes que determinan la estructura de la improvisación. Estas formas de describir el audio (salvo la última) son empleadas generalmente para el reconocimiento de voz en aplicaciones de dispositivos

móviles y computadoras, así como en sistemas de detección de momentos sonoros importantes.

El objetivo de este capítulo es analizar brevemente los descriptores de audio utilizados en el sistema propuesto en la tesis. Estos fueron MFCCs, contraste espectral, centoide espectral y *onsets*. Estos descriptores son los que, según mi propuesta de escucha, arrojan mayor coherencia para clasificar y agrupar elementos sonoros de la improvisación libre. Es importante señalar que aunque hay muchos otros descriptores de audio,[1] [2] de acuerdo con los resultados obtenidos en el apartado 4.2.9 no aportan nueva información para la clasificación, y en la mayoría de los casos pueden producir confusión al momento de clasificar el audio. En este trabajo la utilidad de los descriptores elegidos radica en descifrar para posteriormente clasificar con aprendizaje de máquinas elementos sonoros similares que son empleados posteriormente para generar un método de análisis de la densidad y la estructura tímbrica con base en algunas improvisaciones libres (de lo cual se hablará en el capítulo 4). Por último, se discutirán y definirán los algoritmos de clasificación basados en el aprendizaje supervisado y no supervisado tales como Percepción multicapa o redes neuronales artificiales, aprendizaje profundo y K-Means.

[1] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley, 2012

[2] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 01 2004

3.1. Descriptores de audio para el reconocimiento de elementos sonoros en la libre improvisación

En los últimos años se han realizado múltiples investigaciones respecto a la forma en la que opera el sistema auditivo humano para reconocer y procesar diferentes sonidos complejos[3]. Los avances más significativos han estado enfocados en entender cómo percibimos la frecuencia. Estos sonidos complejos están caracterizados por poseer múltiples bandas de energía distribuidas en varias frecuencias codificadas por el filtrado coclear. Además "en el tronco encefálico auditivo, particularmente en el núcleo dorsal, hay mecanismos inhibidores responsables de mejorar los contrastes espectrales y temporales en sonidos complejos".[4] [5] A nivel cortical, es evidente de acuerdo con los estudios de Evans que los mecanismos internos son capaces de abstraer biológicamente las características de sonidos complejos. El autor señala que el oído realiza tres tareas para reconocer un sonido complejo: primero, la selectividad de frecuencias o análisis frecuencial, es la habilidad del oído para separar un sonido complejo de otros que no son relevantes y extraer los componentes individuales de frecuen-

[3]Un sonido complejo es entendido como la suma de varias sinusoidales puras, es decir, un sonido con diferentes armónicos que suenan a distintas intensidades, también podría entenderse como un ruido que se caracteriza por tener muchas frecuencias que suenan simultáneamente, a diferencia de un sonido simple el cual sería una sinusoidal pura

[4]E. F. Evans. Auditory processing of complex sounds: An overview. *Philosophical Transactions: Biological Sciences*, 336(1278):295–306, 1992 p. 295

[5]Es importante precisar que la investigación de Evans esta basada en el procesamiento monoaural de la escucha.

cia. Este análisis permite y es determinante para entender el timbre del sonido percibido. Además los momentos de inicio de un sonido también son importantes para determinar el timbre y la altura de un "sonido complejo (para el habla, la frecuencia laringal percibida), esta información es esencial para posibilitar que el sistema auditivo pueda diferenciar entre [diferentes] hablantes".[6] Segundo, mejorar los contrastes temporales y espectrales, para compensar la señal sonora que muchas veces viene cargada de ruido. Tercero, "extraer y abstraer las señales de comportamientos significativos de los resultados del análisis espectral periférico".[7] En el habla sería equivalente a determinar los espacios de los componentes frecuenciales y sus cambios en el tiempo.

Continuemos con un análisis general de los cinco componentes del sistema auditivo monoaural. El oído externo y medio condicionan lo que escuchamos ya que enfatizan las frecuencias que son más relevantes para cada especie, en los humanos serían las frecuencias que van de 1 a 3 kHz. La cóclea por su parte, es el analizador de las frecuencias. La membrana basilar y el órgano de Corti crean una especie de bancos de filtros distribuidos que posteriormente son enviados a las fibras nerviosas cocleares. Estas fibras nerviosas codifican las respuestas filtradas en términos de sus patrones temporales. Las respuestas filtradas corresponden a las frecuencias graves y agudas, que son enviadas en paralelo a los núcleos cocleares a través del tronco encefálico (por este circulan las vías sensoriales del gusto, el tacto y el oído). Finalmente, no queda claro hasta dónde se mantienen separados y llegan al cerebro en forma de impulsos eléctricos codificados por diferentes

[6] Ídem
[7] Ídem

tipos de neuronas con funciones muy específicas.[8] Para entender cómo es que el oído analiza sonidos complejos, hay que remitirse a la curva psicológica de afinación de cada fibra individual en el nervio coclear[9].

Si bien la investigación de Evans es muy importante para entender cómo es que funciona el sistema auditivo humano a detalle, seguir por este camino saldría de los límites de la presente investigación, se recomienda su lectura de cualquier forma, ya que hace un análisis más profundo sobre los tres componentes auditivos principales; el nervio coclear, el núcleo coclear y la corteza auditiva.

3.1.1. MFCC descriptor de audio para el reconocimiento de texturas sonoras complejas

Vayamos ahora al analizador de audio MFCC y entendamos para qué es útil en el análisis de señales complejas de audio. El MFCC (Mel Frequency Cepsral Coefficients) puede ser entendido como una descripción compacta de la forma de una envolvente espectral de una señal de audio. Ha sido ampliamente usado para el procesamiento de señales de voz desde que Steven Davis y Paul Mermelstein lo introdujeron en 1980 con el fin de generar un sistema de reconocimiento del habla basado en sílabas. También ha sido ampliamente usado en aplicaciones para procesar y analizar señales musicales.

El primer paso es dividir las señales de audio en cuadros temporales (típicamente 20 milisegundos) con el objetivo de crear múltiples filtros. Debido a que se asume que en pequeños períodos de tiempo sus características cambian poco, es posible realizar varios procesa-

[8] http://www.cochlea.eu/es/cerebro-auditivo/tronco-cerebral Fecha de consulta 20 abril 2018

[9] Ídem p. 296.

mientos para extraer características fijas a cada cuadro de la señal. Además, una ventana Hamming[10] es aplicada para remover los bordes del audio que pueden ser causantes de imprecisiones.

El segundo paso es aplicar la transformada discreta de Fourier a cada uno de los cuadros y obtener la amplitud del espectro de la señal de audio. La información respecto a la fase es descartada debido a que estudios sobre percepción muestran que la amplitud del espectro es mucho más relevante que la fase para identificar sonidos complejos. Además, se toma la amplitud logarítmica del espectro ya que algunos estudios han demostrado que la amplitud percibida de una señal por un humano es aproximadamente logarítmica, como se puede apreciar en la siguiente figura.

El tercer paso es tomar el logaritmo de la amplitud espectral. El cuarto paso es aplicar un banco de filtros correspondientes a la escala Mel y posteriormente suavizar el espectro enfatizando las frecuencias perceptibles más importantes. Por ejemplo se colectan 256 componentes espectrales dentro de 40 recipientes o bancos (mejor conocidos como bins) de frecuencias como se muestra en la figura 3.2.[11]

Para comprender mejor lo anterior, habría que hacer un paréntesis y entender cómo es que funciona la escala de Mel. Básicamente consiste en hacer un mapeo entre las frecuencias y las alturas percibi-

[10]las ventanas son funciones matemáticas cuya utilidad es seguir una continuidad al principio y final de los bloques de audio analizados, la señal de audio se multiplica por la función de la ventana de la cual se obtiene una señal truncada, en este caso la ventana hamming tiene la forma de una campana cuya amplitud comienza arriba de 0 a diferencia de la ventana Hann la cual comienza desde 0, también con la forma de una campana. Utilizar estas ventanas para truncar señales de audio cambia el espectro de frecuencias en la señal original

[11]Beth Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

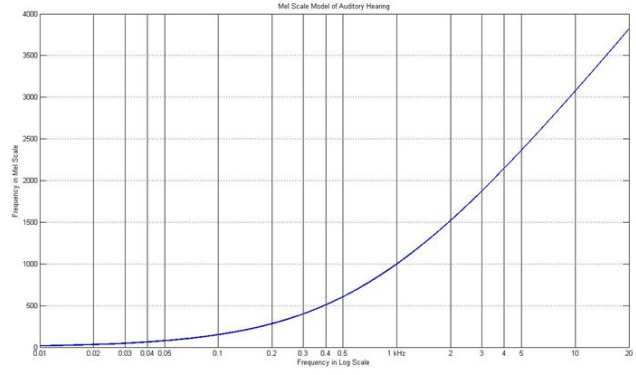

Figura 3.1: Representación logarítmica de la percepción humana

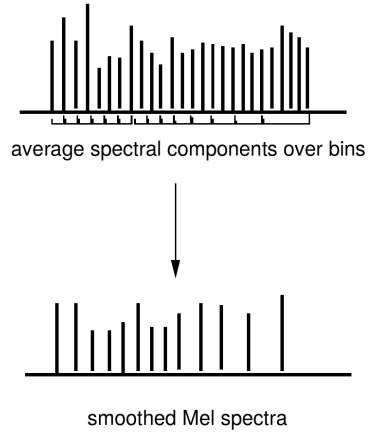

Figura 3.2: Suavizado en la escala de frecuencias de Mel sobre la amplitud espectral. Los componentes espectrales son promediados sobre los bins para obtener un espectro suavizado.

das de acuerdo al sistema auditivo humano que no percibe de manera lineal sino aparentemente logarítmica. Debido a las variaciones del ancho de banda crítica con respecto a la percepción frecuencial humana, la curva es espaciada linealmente a intervalos fijos en las frecuencias graves debajo de los 1000 Hz y de forma logarítmica en las frecuencias agudas arriba de los 1000 Hz. El nombre de Mel fue otorgado por Stevens, Volkmann y Newman en 1937 (S. Smith, J. Volkmann, 1937), derivado de la palabra melodía, y sirve para indicar los rangos melódicos en los que el oído puede identificar de manera más certera los intervalos musicales. La escala de Mel parte de los experimentos perceptuales aplicados a diferentes escuchas que juzgan subjetivamente las distancias interválicas, y difiere de las escalas musicales así como de la escala de frecuencias, las cuales no son subjetivas (Volkmann, Newmann, 1937).

Para reducir el número de parámetros obtenidos en el análisis el último y quinto paso es aplicar una transformada de coseno discreta (DCT) a los vectores espectrales de la escala de Mel obtenidos. Lo anterior debido a que "los componentes calculados del espectro Mel para cada uno de los cuadros están íntimamente relacionados y las características del habla son típicamente modeladas por mezclas de densidades Gaussianas".[12] Al usar esta transformación se obtienen 13 valores por cada cuadro analizado. Por último, debe destacarse que al aplicar la DCT sobre los vectores obtenidos, se realiza el proceso inverso al aplicado durante la transformación inicial (DTF), y en teoría se puede pensar que los valores obtenidos serían los valores de la misma señal, pero con modificaciones para que se parezca a

[12]Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

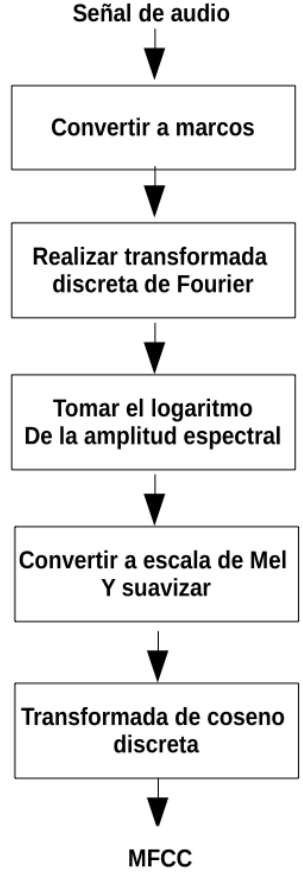

Figura 3.3: Flujo de procesos para obtener el MFCC

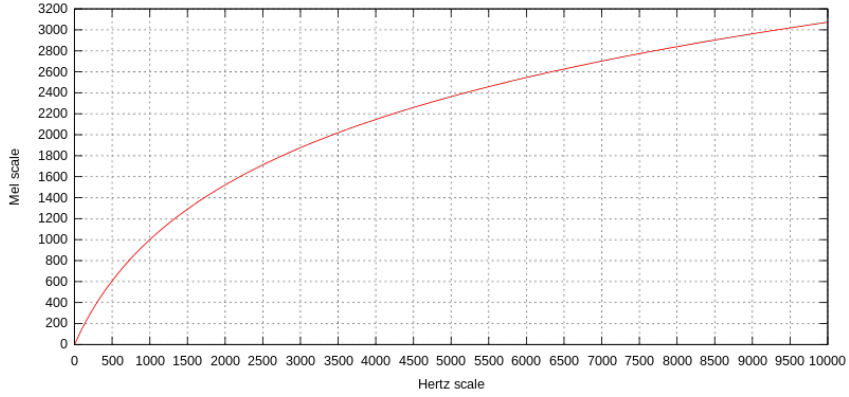

Figura 3.4: Representación de la escala Mel. By Krishna Vedala - Own work. This image was created with gnuplot. https://commons.wikimedia.org/w/index.php?curid=3775197

lo que escuchamos los humanos. "El cálculo [de MFCC] está íntimamente relacionado con el cálculo del *ceptrum* ya que transforma una representación espectral logarítmica. La principal diferencia es que el ceptrum estándar usa una escala de frecuencias no-lineal (escala Mel) para modelar la percepción de alturas no-lineal humana y el uso de la DCT en vez de la DFT".[13] El flujo completo de todos los procesos para la obtención de MFCCs puede ser observado en la figura 3.3.

Cabe destacar que diferentes estudios han demostrado que el uso de MFCC para el análisis espectral de voz y música puede ser suficiente para obtener información relevante respecto a la señal de audio

[13] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley, 2012.

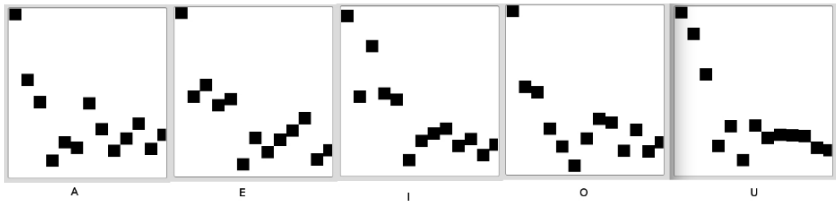

Figura 3.5: Análisis MFCCs de la emisión de las vocales representada con 13 vectores en el eje de las abscisas. En las ordenadas, se representan de forma numérica las características tímbricas de las vocales.

analizada[14]. En la clasificación de señales de audio se ha demostrado que solo al utilizar una pequeña cantidad de vectores MFCC (de 4 a 20 vectores por cuadro de audio analizado) es posible revelar la información más importante para describir el audio introducido[15]. A continuación se muestra el análisis de la emisión de las vocales en lengua hispana sin cambiar de tono, donde se ejemplifica que con los primeros 4 vectores MFCCs es posible diferenciar cada una de éstas. Incluso al introducir otro tipo de vocales como las alemanas o al emitir sonidos extraños el algoritmo sigue reconociendo en los primeros cuatro vectores la información necesaria para ser diferenciadas.

3.1.1.1. Aplicaciones

Debido a su precisión y eficiencia en el análisis de señales de audio el MFCC ha sido empleado en muchas áreas de la vida cotidiana. Ya

[14]Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

[15]Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley, 2012.

sea en sistemas de reconocimiento de información específica introducida por voz para la automatización de aplicaciones, como en sistemas de búsqueda, control y vigilancia. Algunas de estas aplicaciones son Shazam, Google, Echo (Amazon), Cortana (Microsoft), Siri (Apple), Nest, etc. Incluso el Instituto de Telecomunicaciones Estandarizadas Europeo ha normalizado el uso del algoritmo para ser implementado en todos los teléfonos móviles inteligentes.[16] Cada vez más se desarrollan otras aplicaciones desde el campo de la recuperación de información musical (Music Information Retrieval por sus siglas en ingles) y otras instancias como por ejemplo la clasificación de géneros, medidas de similitud de audio, aplicaciones médicas e investigaciones de diversa índole.[17]

Otra aportación importante actualmente dentro del campo de la identificación de señales de audio son las huellas dactilares de audio o fingerprinting. Su objetivo consiste en identificar una grabación de audio específica comparándola con otras. Las huellas de audio pueden ser obtenidas de distintas formas pero una de las más comunes es a través del análisis vectorial que MFCC proporciona. Estas huellas de audio son obtenidas al hacer el análisis de pequeñas secuencias audio de un corpus musical almacenado en una base de datos que después son comparadas con la muestra de audio introducida.

Hasta aquí hemos visto cómo el algoritmo MFCC puede ser bastante útil para describir una de las características que considero fundamentales para el análisis de segmentos en la improvisación libre

[16] European Telecommunications Standards Institute (2003), Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.

[17] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 2007.

como son las cualidades espectrales y tímbricas del sonido. El MFCC nos ofrece la posibilidad de obtener los datos que describen el espectro de una señal de audio compleja, facilitando su identificación y clasificación a partir de las magnitudes que despliega el algoritmo. Asimismo me gustaría resaltar la íntima relación que establece el proceso de extracción de los MFCC con el funcionamiento del sistema auditivo humano, desde la selectividad de frecuencias y su extracción en componentes individuales, pasando por los mecanismos inhibidores que mejoran los contrastes espectrales y temporales de los sonidos hasta la forma en la que las frecuencias graves, medias y agudas son enviadas a los diferentes subsistemas a través del núcleo encefálico, un diseño evidentemente prototípico.

Por último, cabe señalar que actualmente hay varias aplicaciones, librerías, programas, e implementaciones que distribuyen bajo licencias de software libre o de código abierto el algoritmo de MFCC junto con otros descriptores de audio, ejemplos de ello son la libreria Librosa,[18] Marsyas,[19] Vamp Plugins,[20] etc. Esta apertura ha sido de gran utilidad para que múltiples investigaciones y proyectos musicales, sonoros y artísticos se apropien e implementen sistemas basados en la escucha de máquinas particularmente usando MFCC. Sobra decir que debido a su adaptabilidad en múltiples situaciones el MFCC ha resultado uno de los descriptores más robustos que existen actualmente.

[18] https://librosa.github.io/
[19] http://marsyas.info/
[20] http://www.vamp-plugins.org/

3.1.2. Spectral Centroid

El centroide o centro de gravedad se encuentra en la posición media entre todos los puntos de cualquier objeto de n-dimensiones. El Spectral Centroid se calcula sacando la media ponderada[21] de las frecuencias presentes en la señal del espectro, determinadas mediante una transformada de Fourier, dividida por la suma no ponderada[22] del espectro de las frecuencias presentes.

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

Aquí, x(n) representa el valor de las frecuencias ponderadas del número de bins y f(n) representa el centro de frecuencias de esos bins.[23]

En el análisis de audio digital (Spectral Centroid) sería el centro de un conjunto de valores que conforman los datos de un sonido, determinados mediante una transformada de Fourier. Este centro ha sido utilizado para describir el punto de mayor energía del espectro sonoro. Para la percepción auditiva, esta concentración de energía puede estar relacionada con la dimensión tímbrica o el brillo de un sonido.

[21] La media ponderada es una medida de tendencia central, que es apropiada cuando en un conjunto de datos cada uno de ellos tiene una importancia relativa (o peso) respecto de los demás datos.

[22] suma ponderada: sirve para sustentar la toma de la mejor decisión al tener varias alternativas de importancia relativa y poder definir el valor de los criterios tomados para cada una de ellas.

[23] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics.* Wiley, 2012

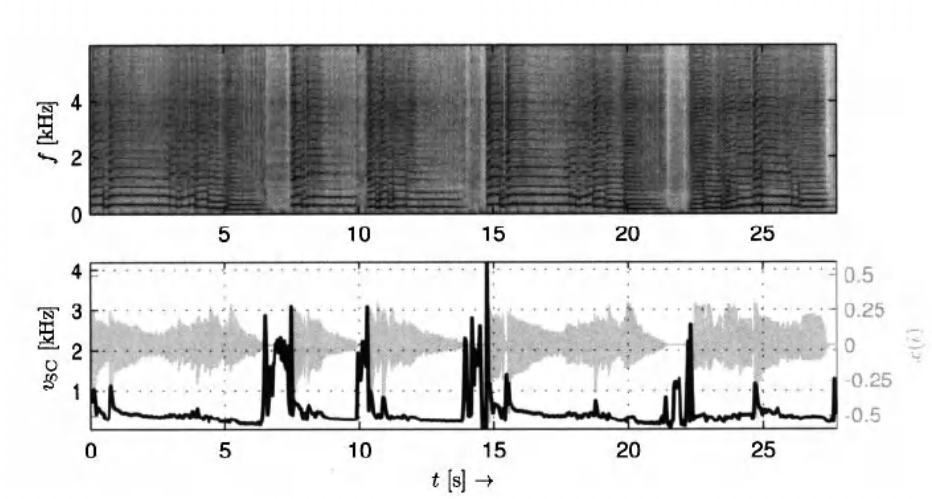

Figura 3.6: Espectrograma de frecuencias graves altamente brillantes

En la figura 3.7 podemos observar que las frecuencias graves contienen mayor energía (estos serían sus componentes más significativos) y los componentes en las frecuencias altas son la cantidad de brillo en esas frecuencias graves. Asimismo en las pausas visualizadas en la forma de onda, podemos observar un alto nivel de ruido visualizado en el espectrograma de la figura 4.1, éstas pausas requieren una consideración especial ya que no son un simple silencio sino más bien parecen ruido. Los inicios y finales de frase están marcados por un cambio significativo que salta súbitamente.

Implementación sencilla de Spectral Centroid

```
from __future__ import division
from numpy import abs, sum, linspace
from numpy.fft import rfft
```

```
spectrum = abs(rfft(signal))
normalized_spectrum = spectrum / sum(spectrum)
# Like a probability mass function
normalized_frequencies = linspace(0, 1, len(spectrum))
spectral_centroid = sum(normalized_frequencies * normalized_spectrum)
```

3.1.3. Onset: detector de momentos de inicio sonoro

La percepción humana es bastante precisa para identificar los comienzos espaciados entre diferentes elementos sonoros, su precisión es tal que ha sido una fuente de inspiración para modelar diversos sistemas de detección de inicio de momentos sonoros. Durante el proceso de percepción auditiva humana, la transmisión sonora es segmentada en una serie de eventos contiguos, varios estudios han demostrado que la velocidad con la que podemos discriminar temporalmente dos o más eventos sonoros es de apenas dos milisegundos entre los inicios de dichos eventos. En 1959 el psicólogo Ira Hirsh encontró que la discriminación temporal de inicios entre dos sonidos es posible para los humanos si la diferencia de tiempo de inicio es tan pequeña como 2 milisegundos, de este modo es posible detectar dos en vez de un solo sonido en una pequeñísima fracción de tiempo. Sin embargo, para determinar cual de los dos estímulos precede al otro se requieren de 15 a 20 milisegundos.[24] Asimismo el resultado de estas mediciones está sujeto a los sonidos sintéticos, mismos con los que se llevó a cabo el

[24]Ira J. Hirsh. Auditory perception of temporal order. *The Journal of the Acoustical Society of America*, 31(6):759–767, 1959

experimento, y es de esperarse que estos resultados varíen hasta diez veces más si son usados otros instrumentos en la medición.

> Se puede concluir que la precisión de la percepción humana de *onsets* [inicios] depende de los datos de prueba y que las desviaciones provocadas por las habilidades motoras parecen estar en el mismo rango. Los resultados presentados implican que un sistema de detección de inicio automático que apunta a la precisión de detección humana (o que se evalúa con datos de prueba anotados por humanos) tendrá un error absoluto medio mínimo en el rango de 5-10 ms; se puede esperar que el error sea tan alto como 10 veces más para instrumentos específicos y lanzamientos con tiempos de subida largos.[25]

La definición sencilla del onset o inicio de un sonido es el lugar en donde un evento sonoro es apenas percibido por un humano. Para la escucha de máquinas los *onsets* son de suma importancia ya que permiten realizar seguimiento de pulsos, estimación del tempo, transcripción automática, segmentación para clasificación de audio, etc. Asimismo algunos algoritmos de detección de momentos de inicio pueden ser bastante sencillos de implementar lo que posibilita su aplicación para realizar detecciones en tiempo-real, a diferencia de otros algoritmos que requieren "mirar hacia el futuro"(look-ahead) y una gran cantidad de recursos de cómputo para realizar sus cálculos.[26]

[25] Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley, 2012. p. 121.

[26] Dan Stowell and Mark Plumbley. Adaptive whitening for improved real-time audio onset detection. *International Computer Music Conference (ICMC)*, 2007.

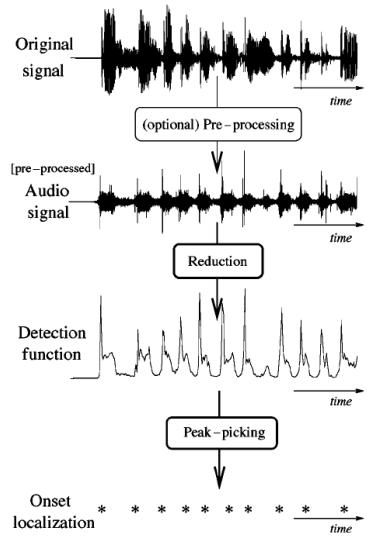

Figura 3.7: Tres pasos para la obtención del onset.

El procedimiento para obtener los *onsets* típicamente abarca tres pasos: el pre-procesamiento de la señal de audio que involucra la transformación de la señal para acentuar algunos rasgos dependiendo de la tarea a realizar, esto puede ser normalizarla o aplicarle alguna compuerta de ruido; la reducción, incluye convertir la señal de audio a una velocidad de muestreo más baja, ayudando a detectar los momentos más significativos en cambios de amplitud, y finalmente la función de detección de momentos de inicio (ODF onset detection function) en la cual se localizan los *onsets*.

El onset está directamente relacionado con el concepto de *transient* (momento transitorio o fugaz de máxima amplitud que transcurre en un evento sonoro), aunque no deben ser confundidos ya que el último

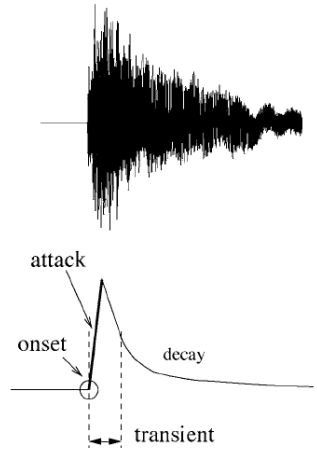

Figura 3.8: *Onset*, Ataque, *transient* y decaimiento

no está necesariamente presente en todos los sonidos.[27] El *transient* se encuentra en el momento donde el audio llega al pico máximo de amplitud y una de sus características es que dura una fracción de segundo (¿=2 milisegundos), además de que sus componentes de frecuencia por lo general son más altos de la altura original producida, esto es debido al estiramiento del cuerpo elástico excitado, por ejemplo, la cuerda de guitarra.

Hay tres momentos importantes distinguibles que pueden estar presentes en los inicios de eventos sonoros[28]:

[27] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, Sept 2005.

[28] Bruno H. Repp. Patterns of note onset asynchronies in expressive piano performance. *The Journal of the Acoustical Society of America*, 100(6):3917–3932, 1996

- **Tiempo de inicio de la nota:** es el momento en el cual un instrumento es tocado para producir un sonido. Por ejemplo el movimiento del martillo del piano, antes de tener contacto con la cuerda.

- **Tiempo de inicio acústico:** el momento en que el sonido esta presente físicamente y puede ser medido. (Éste y el punto anterior no implican necesariamente que el sonido sea audible para la percepción).

- **Tiempo de inicio perceptivo:** el momento en que el sonido es percibido por el oído como un ataque.

En la detección de *onsets* es posible enfocarse en dominios específicos para obtener una detección más adecuada para diferentes contextos musicales o sonoros. Algunos de estos dominios están involucrados con la optimización para detectar frecuencias polifónicas o monofónicas, detectar ataques como los producidos por un instrumento de percusión, cambios en la distribución energética espectral y reconocimiento de patrones espectrales extraídos con técnicas de aprendizaje de máquinas y redes neuronales.

La detección de momentos de inicio es actualmente una área de investigación bastante activa desde 2005, estas investigaciones se enmarcan en el ámbito de la Recuperación de Información Musical, Evaluación e Intercambio MIREX aunado a un proyecto más grande conocido como ISMIR (Sociedad Internacional de Recuperación de Información Musical).[29] Estas organizaciones año con año buscan perfeccionar las diferentes aproximaciones y técnicas para la detección de

[29] http://www.ismir.net/

momentos importantes de audio en distintos campos de conocimiento y aplicación.

3.1.4. Contraste espectral

Sus aplicaciones principales son los implantes de filtros cocleares, ya que en ciertos sistemas auditivos humanos es posible encontrar que los filtros encargados de decodificar el sonido del habla están más abiertos de lo normal "y en algunos casos anormalmente asimétricos",[30] lo cual genera una especie de anublamiento espectral que impide escuchar de forma correcta."El procesamiento a través de estos filtros anormales puede producir una mancha de detalle espectral en la representación interna de estímulos acústicos".[31] A partir de este tipo de casos y otras cualidades auditivas es que se han generado distintos algoritmos de contraste espectral que pueden ser capaces de funcionar en tiempo-real y ayudar a tener una mejor audición. Respecto a esto se han sumado varios experimentos en el ámbito de la Recuperación de Información Musical,[32][33] que han demostrado que el contraste espectral basado en octavas y el basado en formas de bandas (agregando información acerca de la forma de la banda del espectro) pueden alcanzar mayores niveles de precisión al intentar clasificar di-

[30] Jun Yang, Fa-Long Luo, and Arye Nehorai. Spectral contrast enhancement: Algorithms and comparisons. *Speech Communication*, 39(1):33–46, January 2003.

[31] Ídem

[32] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, 2002.

[33] Vincent Akkermans, Joan Serrà, and Perfecto Herrera. Shape-based spectral contrast descriptor. In *Sound and Music Computing Conference*, pages 143–148, Porto, Portugal., 25/07/2009 2009.

ferentes estilos musicales que otros descriptores incluyendo el MFCC. Estas dos aproximaciones pueden ser más precisas al momento de discriminar entre los diferentes estilos musicales.

El contraste espectral se puede definir como la diferencia de decibeles entre picos y valles del espectro de una señal de audio. Para extraerlo, primero se obtiene el espectro calculando la FFT sobre las muestras digitales, después se buscan las diferencias promedio de decibeles entre picos y valles. Dentro de estas diferencias se busca obtener por separado cada sub-banda espectral considerando las divisiones por bandas en octavas que van de 0, 200Hz, 400Hz, 800Hz, 1.6kHz, 3.2kHz, y 8kHz. Su objetivo es tratar de representar de forma resumida las características espectrales de una señal de audio. De este modo pueden ser obtenidas las distribuciones de los componentes armónicos y no armónicos.

> El contraste espectral basado en octavas considera el pico espectral, el valle espectral y su diferencia en cada sub-banda. Para la mayoría de la música, los fuertes picos espectrales se corresponden aproximadamente con los componentes armónicos; mientras que los componentes no armónicos o ruidos a menudo aparecen en los valles espectrales. Por lo tanto, la característica de contraste espectral podría reflejar aproximadamente la distribución relativa de los componentes armónicos y no armónicos en el espectro.[34]

[34]Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, 2002

La principal diferencia de este algoritmo con el MFCC es que para representar el audio analizado el MFCC utiliza una envolvente espectral promedio de la distribución espectral en la escala de Mel, de manera que no podría representar las características espectrales relativas de cada sub-banda, que según Jiang et al. parecen ser más importantes para discriminar los diferentes tipos de música. "La distribución espectral promedio no es suficiente para representar las características espectrales de la música. Sin embargo, el contraste espectral conserva más información y puede tener una mejor discriminación en la clasificación del tipo de música."[35]

3.2. Aprendizaje de máquinas

Actualmente, el aprendizaje de máquinas esta inserto en muchas de las aplicaciones que usamos día con día; programas de recomendaciones de productos tales como los usados en Amazon, MercadoLibre, Ebay, etc, son prácticamente la norma para cualquier empresa que quiera ofrecer sus productos y servicios en línea. Asimismo, los sistemas de reconocimiento de voz capaces de comprender lo que decimos para automatizar tareas, ya sea desde llamar a alguien, tocar una canción, programar un calendario de actividades o disparar un arma. Las redes sociales y buscadores presentan en casi todas sus aplicaciones (como los sistemas de traducción, corrección automática, despliegue de información, etc.) esta tecnología en su punto más álgido de desarrollo. Con otras aplicaciones es posible conversar de forma natural

[35]Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, 2002

con un asistente automático que si tuviera una voz humana podría pasar la prueba de Turing fácilmente.[36]Incluso hay varios foros de discusión de usuarios que han sido sorprendidos por su programa (Echo: Alexa[37]) debido a la supuesta inteligencia que pueden llegar a desarrollar mediante la convivencia y el intercambio generados con dichos sistemas. Como mencionan algunos autores el aprendizaje automático es el principio de la inteligencia artificial, que subyace en mayor o menor medida en nuestras vidas, e involucra herramientas (o armas) que conoceremos y empezaremos a usar y/o a enfrentarnos de forma cotidiana en unos pocos años.[38]

> [El aprendizaje automático es] la base del nuevo enfoque en informática en el que no escribimos programas sino recopilamos datos; la idea es aprender los algoritmos para realizar tareas automáticamente a partir de los datos. A medida que los dispositivos informáticos se vuelven más omnipresentes, una parte más grande de nuestras vidas y trabajo se registra digitalmente y, a medida que aumenta la Big Data y la teoría del aprendizaje automático, la base

[36]Actualmente hay un proyecto titulado WaveNet de la empresa DeepMind encargado de este respecto el cual usa redes convolucionales neuronales (convolutional neural network) de aprendizaje profundo los cuales están llegando a resultados bastante aceptables al respecto de la generación automática de voces humanas.https://deepmind.com/blog/wavenet-generative-model-raw-audio/

[37]http://adage.com/article/digitalnext/amazon-alexa-dangerous/307328/

[38]Por ejemplo algunos ya afirman que en 2020, se espera que diez millones de automóviles autónomos circulen por las calles del primer mundo. http://adage.com/article/digitalnext/autonomous-cars-pave-road-advertisers/307103/

> de los esfuerzos para procesar esos datos [y convertirlos en] conocimiento también ha avanzado.[39]

De forma práctica el aprendizaje automático es la capacidad de una máquina para aprender sin necesariamente estar programada para tales fines, más bien su aprendizaje es el resultado emergente del funcionamiento de sus algoritmos. En la programación se pueden dejar ciertos valores sin definir y dejar que el sistema determine cuales son los mejores valores para la resolución de un problema concreto. Inclusive, se ha llegado a considerar el enfoque del aprendizaje automático como una suerte de "afinamiento de parámetros de una caja negra hasta que produce resultados satisfactorios"[40]

El aprendizaje de máquinas es empleado para obtener o detectar patrones y llegar a conclusiones o hacer ciertas predicciones acerca de una base de datos (*dataset*) determinada con la cual es abastecido el sistema. Las máquinas pueden aprender de tres formas: de manera supervisada, sin supervisión o aprendizaje reforzado. En el siguiente apartado hablaremos a detalle sobre las dos primeras. De manera general los pasos involucrados en el aprendizaje automático se pueden dividir en 5 secciones:

Selección de los datos

Actualmente es posible encontrar, comprar o extraer múltiples y colosales bases de datos de diversa índole con las cuales se pueden realizar experimentos casuales y/o sociales, tesis de maestría, publicaciones en revistas acádemicas, investigaciones de clase mundial o

[39]Ethem Alpaydin. *Machine Learning: The New AI.* The MIT Press Essential Knowledge series. MIT Press, 2016

[40]Nishant Shukla. *Machine Learning with TensorFlow.* Manning Publications, 2017.

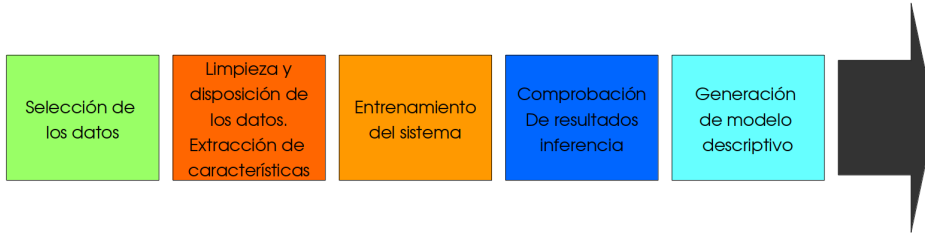

Figura 3.9: Mapa de aprendizaje automático

aplicaciones de control al servicio del capitalismo. La cantidad de información que circula hoy en internet es brutal y sin embargo, hay muchas iniciativas tanto privativas como colectivas que buscan mantenerla organizada, accesible, e incluso (algunas) de distribución bajo licencias libres. Aunque muchos otros optan por crearla, ya sea por la falta de diversidad de bases de datos o por la especialización de datos que un tema específico pueda requerir. Para esta investigación se propuso la creación de una base de datos enfocada solamente en la libre improvisación.

Extracción de características

Los datos tienen una estructura que presenta información redundante aunque muchas veces esta no sea evidente, para poder analizarla, las propiedades de los datos deben estar dispuestas en forma numérica, comenzando por extraer ciertos valores que la describan de manera abstracta y simplificada. Estas representaciones pueden ser vectores, matrices o gráficos.[41] En el análisis sonoro por ejemplo,

[41] Nishant Shukla. *Machine Learning with TensorFlow.* Manning Publications, 2017. p. 7

se extraen descriptores que definan ciertas características del audio; frecuencias, timbre, densidad, altura, volumen, de acuerdo al descriptor empleado. En la presente investigación se extrajeron varios de estos descriptores aunque el que generó mejores resultados fueron los vectores MFCC que describen de forma abstracta el timbre y que comprenden el promedio de un segmento de un archivo de audio.

> Este tremendo esfuerzo para seleccionar tanto el número de mediciones como las medidas para comparar se llama ingeniería de características. Dependiendo de las características que examine, el rendimiento del sistema puede fluctuar drásticamente. Seleccionar las características correctas para rastrear puede compensar a un algoritmo de aprendizaje débil.[42]

Entrenamiento del sistema

Es el momento en el que el algoritmo de clasificación se encargará de trabajar de forma autónoma con la información presentada. Una característica fundamental de este proceso son las iteraciones, es decir, las veces que el sistema corroborará la información para aprender de ella, ordinariamente por cada iteración el sistema estaría aprendiendo más y de manera más rápida y eficiente de la información presentada para generar un modelo que permita deducir e inferir resultados y llegar a conclusiones. El programador puede dejar algunas variables

[42] Idem.

con valores abiertos y dejar que el sistema de aprendizaje automático determine los mejores valores por sí mismo durante cada iteración.

Comprobación de resultados

En este paso los resultados tienen que ser analizados por un humano para verificar que los valores solicitados sean correctos y que el sistema de clasificación esta funcionando de forma correcta. Aunque esta comprobación también puede ser realizada por un algoritmo como es el caso de las redes adversas generativas (Generative Adversarial Networks, GAN).[43] En esta investigación una forma de comprobación fue realizada partiendo de la escucha atenta de las muestras de audio agrupadas por clases, otra estuvo basada en la generación de una tabla en la cual se evidenciaron las combinaciones entre algoritmos de clasificación y descriptores con un mayor índice de certeza para discernir entre distintos tipos de información descrita en el apartado 4.2.2.

Generación de un modelo descriptivo

Implica la generación de un modelo que nos permita visualizar, comprender o almacenar la información de forma resumida, generalmente esto se hace partiendo de la visualización de un gráfico, además este modelo puede ser guardado y empleado en futuras pruebas con el

[43]Estas tienen la capacidad de corroborar la información mediante dos redes; la primera, evalúa si la información de entrada es correcta y la segunda, trata de engañar a la primer red para que esta se haga cada vez más inteligente y pueda determinar qué es un engaño y qué no. Asimismo la red neuronal que engaña, genera muestras modificadas o transformadas de las muestras originales, reforzando así más su comportamiento. Estas dos redes entran en una continua retroalimentación de aprendizaje mutuo.

sistema. Por ejemplo, al aportar nueva información al modelo, podría identificar de que clase de información se trata, idealmente desplegando un índice de certeza que indique la similitud de esa nueva muestra respecto a las muestras almacenadas anteriormente por el modelo. En el caso de esta investigación los modelos generados pueden ser usados para futuros desarrollos.

3.3. Aprendizaje supervisado

El aprendizaje supervisado requiere un cierto ordenamiento y una clasificación anotada de forma manual de la base de datos. Es recomendable dividir en secciones la base de datos inicial para que la primera parte sea utilizada en el momento de entrenamiento del sistema y la segunda parte sirva para corroborar si el modelo de clasificación generado por el algoritmo es lo suficientemente robusto como para obtener los resultados que esperamos. Un escenario ideal para el aprendizaje supervisado sería que el algoritmo fuera lo suficientemente robusto para determinar de forma correcta instancias nunca antes vistas pero que son similares al modelo construido en el proceso de entrenamiento. Es común que los resultados obtenidos no sean del todo favorables, muchas veces debido al ruido o a las equivocaciones que puedan surgir en las anotaciones manuales de los elementos de la base de datos. Otro tipo de problema puede surgir si la misma cantidad de elementos no es lo suficientemente grande para obtener un buen modelo de clasificación. Por el lado contrario si la información extraída de esos elementos es demasiada, puede sobrealimentar (*overfeed*) al sistema ocasionando una repetición de la misma información más que obtener predicciones posibles con nuevos datos de entrada.

Asimismo, una clasificación con demasiada información puede no ser relevante para el análisis, ya que puede generar ruido y ocasionar que la predicción del algoritmo sea difusa o no tenga lógica alguna para nuestra percepción.[44]

Los algoritmos de aprendizaje supervisado tienden a fijar límites entre los elementos de la base de datos que son similares. Una forma muy concurrida de asignar estos limites o fronteras entre las regiones del espacio de datos, una vez que han sido agrupados por su similitud, consiste en encontrar los lugares en donde hay menor concentración de datos, idealmente se buscaría una linea divisoria que creará un espacio vacío que sea lo más amplio posible para distinguir claramente entre las diferentes categorías. Al entrar nuevos ejemplos de información el algoritmo debería identificar cuales de esas nuevas instancias pertenece a alguna de las categorías con cierto índice de precisión.

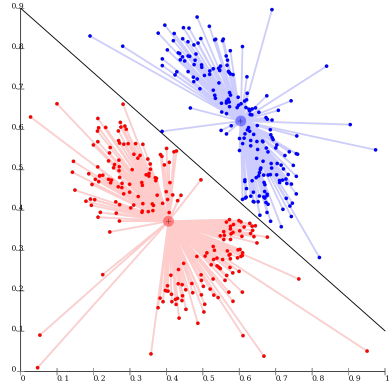

Figura 3.10: Linea de limites entre clases

[44]Nishant Shukla. *Machine Learning with TensorFlow*. Manning Publications, 2017. p. 13.

3.3.1. Aprendizaje profundo

El aprendizaje profundo integra múltiples capas de redes neuronales que extraen la información de forma no-lineal, es decir, cada capa sucesiva de información procesada usa la salida de la capa anterior como una nueva entrada a procesar, de manera tal que las salidas obtenidas en la última capa de las redes involucra la información de todas las anteriores, generando resultados de clasificación mucho más complejos. Usa varios tipos de algoritmos de segmentación los cuales pueden actuar de forma simultanea para combinar los valores de capas anteriores y aprender funciones más complejas de la entrada principal. Al final lo que se obtiene de las últimas capas son salidas con descripciones y conceptos más abstractos. En el siguiente enlace puede apreciarse y experimentar en linea con diferentes configuraciones de las redes neuronales de aprendizaje profundo para la solución de diferentes problemas de clasificación.[45]

3.3.2. Percepción multicapa o redes neuronales artificiales

Es una herramienta inspirada en la biología y particularmente en la forma en la que se cree que funciona el cerebro humano, que posibilita a las computadoras a aprender de la observación de múltiples datos. Actualmente las redes neuronales junto con el aprendizaje profundo (que involucra los mismos principios aplicando otras técnicas), son las mejores soluciones a muchos problemas surgidos en el campo del aprendizaje automático. Primero se crea un conjunto de neuronas que se conectan entre si para intercambiarse información, después se

[45] https://bit.ly/2KDvusu, https://bit.ly/2yasHTp

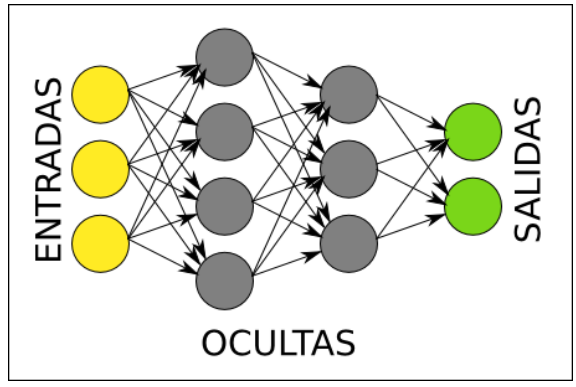

Figura 3.11: Red neuronal profunda

le pide al sistema que resuelva un problema asignado tratando de resolverlo una y otra vez, recordando y reforzando las conexiones que generan mejores resultados y disminuyendo su atención a aquellas que producen resultados insatisfactorios. Una aproximación para que esto suceda sería otorgar pesos sinápticos en la estructura de conectividad, prestando mayor atención a aquellas neuronas que generan mejores resultados es posible obtener resultados más adecuados a la hora de clasificar.

3.4. Aprendizaje sin supervisión

El aprendizaje sin supervisión es mucho más flexible y no depende de un supervisor ni requiere una base de datos anotados manualmente. Dicha situación ayuda a evitar posibles errores causados por el ruido o por equivocaciones que pudieran ocurrir en las anotaciones manuales. Además de que no habría una subjetividad particular ge-

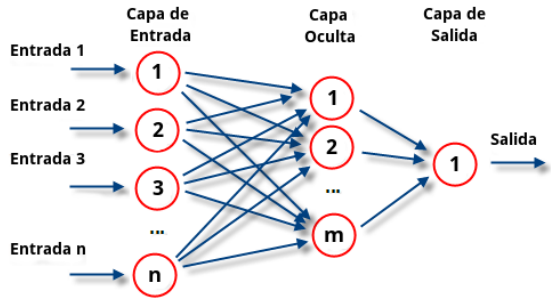

Figura 3.12: Red neuronal profunda

nerada por el apego que podemos tener como humanos al respecto de un objeto específico y definirlo no por cualidades intrínsecas sino por una apreciación personal, cuando para otro(s) el mismo objeto podría pertenecer indudablemente a otro grupo de objetos. La meta del aprendizaje sin supervisión es encontrar regularidades en los elementos de una base de datos, para inferir como estos se interrelacionan la mayor cantidad de las veces. En el campo de la estadística este procedimiento es conocido como modelo de mezcla, el cual es un sistema probabilístico que ayuda a identificar la presencia de subpoblaciones dentro de una población más grande, mediante la distribución probabilística de observaciones sobre todos los elementos que conforman una población.

La reducción de dimensionalidad se trata de manipular los datos para verlos bajo una perspectiva mucho más simple. Por ejemplo, al eliminar funciones redundantes, podemos explicar los mismos datos en un espacio de dimensiones más bajas y ver qué características realmente importan. Esta simplificación también ayuda en la visualización de datos o el preprocesamiento para la eficiencia del rendimiento.

Si bien hay varias aproximaciones para el aprendizaje sin supervisión, nos centraremos en la aproximación presentada en este trabajo: K-Means es uno de los métodos de agrupación de datos (clustering) más antiguo, propuesto por Stuart Lloyd en 1957. Este algoritmo particiona un conjunto de datos para ordenarlos en grupos de similitud de acuerdo con ciertos criterios. El algoritmo puede intentar organizar los elementos de una base de datos partiendo de distancias de similitud y calculando la media más cercana en la cual agrupa los elementos que más se parecen entre si. La textitk en este algoritmo, representa una variable y en este caso será aplicada al número de grupos tímbricos que se desea obtener. Cabe mencionar que esta variable será seleccionada por el usuario manualmente y no por el programa. El algoritmo consiste básicamente en asignar los vectores —generados en la descripción numérica, en este caso de los archivos de audio— a diferentes grupos y re-centrar los puntos medios o centroide mediante múltiples iteraciones del proceso donde se van a agrupar los vectores. A cada iteración los conjuntos de vectores extraídos se van acercando cada vez más con sus semejantes. Si dos conjuntos de vectores están muy juntos, eso significa que sus características tímbricas son similares. "Queremos descubrir qué archivos de audio pertenecen al mismo vecindario, porque esos clústeres probablemente serán una buena forma de organizar nuestros archivos de música".[46]

El algoritmo de K-Means fue bastante empleado en esta investigación debido a su implementación sencilla con Python y TensorFlow y por los resultados obtenidos en las clasificaciones, ya que éstas resultaron adecuadas en la tarea de asignar fragmentos similares dentro de diversas improvisaciones libres.

[46] Ídem, p. 14.

3.5. Corolario

En este capítulo fueron definidos y mencionados algunos conceptos, métodos y herramientas empleados en los campos del aprendizaje y la escucha automática desde una perspectiva más técnica para abordar el siguiente capítulo, especialmente a lo largo del apartado 4.2 sobre la máquina que escucha, en el cual se asumen varias de las definiciones y conceptos expuestos para posibilitar una mayor fluidez en aspectos metodológicos, técnicos y perceptivos del sistema propuesto vinculados con aspectos de carácter estético, gestual y formal de la improvisación libre.

Capítulo 4

¿Una máquina que escucha libre improvisación?

4.1. Discusiones sobre la libre improvisación

> *Es creación; los creadores musicales nos la entregan en el acto, en el concierto, creándola con y delante de nosotros. No nos están entregando un producto, una interpretación cuidadosamente ensayada (a veces) de una obra musical preconcebida por otra persona; nos están invitando a participar en un proceso, a vivir la creación musical in situ.*
>
> – Wade Matthews, *Y la libre improvisación, ¿qué tiene de improvisada?*

> *La improvisación libre no tiene un sonido idiomático preescrito. Las características de la música libremente impro-*

> *visada son establecidas solo por la identidad musical-sonora de la persona o personas que estén tocándola.*
>
> – Derek, Bailey*Improvisation: Its Nature And Practice In Music*

> *Contrarespuesta al freejazz, más que un énfasis en la técnica éste se posiciona en la escucha y en los modelos de la relación interpersonal. Se busca una horizontalidad, en la que no haya una pretensión de autoridad en el dominio técnico o más aún respecto a un lenguaje establecido. Se encuentra abierta a la emergencia y contingencia de lo que se produce en colectivo, no hay una búsqueda por expresar las personalidades de cada quién, sino un sonido más homogéneo, construido de forma grupal.*
>
> – Rossana Lara, Conversación personal

En la improvisación, donde pareciera no haber una estructura ni una forma muy claras, donde las interacciones parecen evaporarse y las características tímbricas responden a una necesidad que va más allá de los timbres tradicionales de una orquesta o donde las formas de organización jerárquicas de otras prácticas musicales se difuminan, pareciera que surgen otras formas de interacción, producto de resultados emergentes no predecibles por ninguno de los integrantes. Así, la improvisación libre puede ser pensada desde sus interacciones y dinámica cambiante, como un sistema[1] social que es influenciado y

[1]La palabra sistema proviene del griego *systema* que a su vez se deriva de *synistemi* que significa: conjuntar, mezclar, organizar. Un sistema es un conjunto de componentes donde cada elemento está relacionado al menos con algún otro y la manera cómo un elemento afecta el todo depende de al menos algún otro elemento.

retroalimentado no tanto por individuos específicos sino por la misma comunicación y los medios que los conectan. En esta práctica "todas las fronteras internas [son] disputadas y todas las solidaridades cambian. Todos los límites internos dependen de la autoorganización de los subsistemas y no más de un «origen» histórico o de la naturaleza o lógica del sistema abarcador".[2] Estos subsistemas los entiendo precisamente como los agentes que al regular su dinámica producen la autoorganización del sistema y por consiguiente sus procesos emergentes respectivos. En este sentido, el resultado sonoro en un sistema como la improvisación libre (creada entre varios músicos) no responde necesariamente a su origen ni a su historia, sino más bien a esas fronteras que van siendo mediadas, negociadas o traspasadas todo el tiempo por los medios que vinculan a sus participantes.

Es importante aclarar que no es fortuito que al mismo tiempo que hago improvisación libre, esté tratando de emular un sistema que pueda improvisar libremente. La raíz del asunto —en mi caso— parte de crear algo que sea capaz de transgredir, resquebrajar y violentar mis propios hábitos dentro de la práctica de la improvisación para transformarla. Considero de suma importancia dicha transformación ya que la misma práctica de la improvisación libre desde sus inicios, buscó desenmascararse y mantenerse en un estado de liminalidad en la práctica sobre los ya conocidos géneros musicales y el estatus quo del arte de ese momento.

Muchas de estas expresiones artísticas son hijas del periodo de entre guerras y posguerras, el arte en general estaba tendiendo hacia lo abstracto, ganando terreno hacia la percepción del mundo desde

[2]Eguzki Urteaga. La teoría de sistemas de niklas luhmann. *Univerisdad del País Vasco*, 2009.

modos de hacer colectivos. La música no podría quedarse atrás; como menciona Joe Harriot: "si puede haber pintura abstracta, ¿por qué no música abstracta?"[3] La libre improvisación comienza a surgir en esa búsqueda por ir más allá de las formas tradicionales como las del jazz —por mencionar una—, en donde hay una clara delimitación de todos los aspectos que ocurrirán a lo largo de una pieza, o como en las dinámicas producidas en la música académica occidental, primadas en su mayoría por la relación jerárquica que se establece entre compositor, director e intérprete(s).[4] En la improvisación libre pasa lo contrario, se comienzan a buscar otras formas de aproximarse a la materia sonora para crear. Además, no debe valorarse solamente por los momentos de comunicación lograda o satisfactoria entre todos los integrantes del colectivo, sino también por los momentos donde la unidad y cohesión sonora son más débiles. En la improvisación no se priorizan las formas ni las estructuras predefinidas, sino más bien, la identidad sonora que va sugiriendo al momento producto de las interacciones y subjetividades sonoras de cada músico. Se busca llevar al límite los instrumentos o incluso crearlos, y encontrar desde ellos una forma de libertad dentro de la opresión simbólica de las instituciones del arte. Al respecto Coleridge Goode, contrabajista de la escena del jazz en Inglaterra junto con Joe Harriot a inicios de 1960, comenta: "[...] era posible jugar sin esas armonías predeterminadas y, de hecho, sin ningún tipo de preconceptos sobre lo que se iba a tocar y cómo. La

[3]David Toop. *Into the Maelstrom: Music, Improvisation and the Dream of Freedom Before 1970*. 2016. p. 148

[4]No quiero que esta afirmación parezca que es así en toda la música académica ya que hay muchos casos en donde el compositor justamente dota de libertad creativa a los músicos para que construyan desde de una serie de posibilidades delimitadas, parte o todo el discurso de la pieza.

idea sería confiar en la espontaneidad completa de todos los músicos: permitir que la interacción libre del grupo cree armonías y relaciones musicales [...] que no se habían planificado de antemano."[5]

4.1.1. Lo libre de la libre improvisación

Aunque discursos puede haber muchos acerca de la libertad, hablar de libertad puede significar muchas cosas dependiendo el contexto desde donde se enuncie. Por ejemplo, en el quinteto de Joe Harriot hay claramente una permanente influencia jazzística, hay una relación sumamente estrecha con la construcción tradicional de temas y motivos del jazz que parten también de pulsos estables y compases binarios. Sin embargo, en sus composiciones es muy notorio que están buscando una singularidad dentro del estilo del jazz. Lo que escucho en su música es una extensión de los límites impuestos por las convenciones musicales y críticas de aquella época. Harriot inicia al tocar una de sus composiciones musicales diciendo: "Me gustaría tocar por primera vez aquí una de nuestras composiciones abstractas, de esta forma musical no estamos usando ningún set [instrumentación], armonía o estructura particular, estamos intentando pintar sonidos, colores y efectos con esta música vieja, espero que les guste, *Coda*".[6] Resulta extraño que al escuchar esta composición, desafortunada o afortunadamente para mis oídos y perspectiva contemporánea, no podría ser considerada como improvisación libre; ni si quiera una improvisación libre temprana —como lo enuncia David Toop–. Sería como pensar que solamente por agregar elementos nuevos a cualquier música actual se estuviera haciendo improvisación libre. Sin embargo, lo que Toop

[5] Ídem, p. 150.

[6] https://bit.ly/2rjbCT4 Fecha de consulta 4 de mayo de 2018.

quiere señalar es que: "Al igual que con otros ejemplos de improvisación libre temprana, escuchamos con los oídos equivocados, incapaces de comprender su impacto contemporáneo, solo capaces de escuchar a través del filtro de las innovaciones posteriores".[7] Esto podría ser discutible en el sentido que no todos vamos a dejar de apreciar esa música "innovadora" y comprender su impacto inmediato social y en el entorno, pero entiendo que, a lo que Toop hace referencia es a una escucha que es conducida por los gustos de la industria musical, la crítica y los estándares comerciales. Entonces esos oídos equivocados más bien son oídos alienados. Sin embargo, es posible que esas innovaciones posteriores sean las que me obligan a escuchar a Joe Harriot como un jazz formal y estructurado como cualquier otro sistema de estándares estructurados por progresiones de acordes. Entonces, ¿a qué hace referencia la palabra *libre* en la improvisación, cuando la misma práctica reafirma una técnica y modos de proceder específicos? o ¿cuándo la improvisación libre ya está enmarcada como un género reconocible por los mismos que la ejecutan así como por los que la escuchan?

Derek Bailey, incluso comenta en los noventa:

> Hay innovaciones hechas, como uno esperaría, a través de la improvisación, pero el deseo de mantenerse por delante del campo no es común entre los improvisadores. Y en lo que respecta al método, el improvisador emplea el más antiguo en la creación musical.[8]

[7] Ídem, p. 150.

[8] Derek Bailey. *Improvisation. Its Nature and Practice in Music.* Moorland, 1980. p. 83

¿Podría ser que agregar la palabra "libre" es una estrategia retórica para convencer al otro de algo que merece la pena ser atendido o escuchado? o por el contrario, ¿se nombra para repeler a las audiencias que creen que al hablar de libertad se está vinculando la práctica a un discurso politizado —que hoy día muchos músicos evitan, con argumentos como: "hay que ser diplomáticamente correctos"— y más bien termina por llegar a audiencias más exclusivas/conocedoras? ¿o a quién se está apelando con esta idea de lo libre; es que se trata solo de una diferenciación de otros, que le permite a los músicos, como en el caso de Harriot, posicionarse y afirmar: mi acción o mi obra artística es relevante? No lo creo. Debe haber relaciones más profundas que nos ayuden a comprender su significado.

Al respecto Chefa Alonso, saxofonista, improvisadora, compositora y educadora española plantea:

> [...] la improvisación ha sido una constante en la historia de la música, a pesar de su mayor o menor reconocimiento. Lo que distingue a la improvisación libre contemporánea de cualquier otra es la ausencia de un marco normativo. La improvisación en la música barroca, en las músicas folklóricas y populares, en el jazz, juega siempre con la presencia de una gramática, de un contexto regulado que puede obedecer a reglas de contrapunto, a acuerdos melódicos o rítmicos [...]. Cualquiera de estos referentes facilitan el flujo creativo de los improvisadores ofreciéndoles un marco, unos límites, unas reglas. La improvisación libre carece de una gramática referencial. Es una música

> que nace desde el deseo de crear en el momento y colectivamente una música nueva.[9]

Es precisamente ese marco normativo que menciona la improvisadora el que se negocia todo el tiempo y se adapta de forma dinámica a cada época, lugar y contexto. Hoy día, y desde hace unos 40 años, ese marco en la improvisación libre está compuesto por toda una tradición, valores, juicios estéticos, decisiones y principalmente cambios de paradigma en cuanto al papel que juega el improvisador como creador e intérprete de lo que crea. Como afirma el compositor e improvisador Wade Matthews:

> La libre improvisación europea es heredera de los inmensos cambios producidos en todos los aspectos del arte entre el final del siglo XIX y nuestros días. En la música, estos incluyen: la revaloración del timbre, el reconocimiento del pulso regular como una convención de la que se puede prescindir o no, la aceptación de la materia sonora como algo expresivo de por si, la incorporación de sonidos otros que los doce tonos temperados, la explosión de nuevas técnicas instrumentales, la incorporación de nuevas tecnologías, el abandono de las formas preestablecidas en favor de formas orgánicas y/o abiertas, una explotación de las múltiples relaciones posibles entre la intencionalidad y el azar, etc. Esta música de creación en tiempo real propone una alternativa a una de las pocas estructuras que había salido

[9] Chefa Alonso. *Improvisación libre la composición en movimiento.* 2007

> más o menos intacta de un siglo de revolución continua: la tradicional jerarquía de compositor-intérprete.[10]

Asimismo, la improvisación libre se diferencia por sus dinámicas sociales —en su mayoría autogestivas— e interacciones que genera el acto de la improvisación; ya sean redes de colaboración para generar sus propios espacios para tocar, organización y gestión de recursos para financiar festivales, generar sellos discográficos, producciones o invitar a improvisadores de otras partes del mundo con los que nunca se había tocado antes. El improvisador comprometido con su práctica se convierte en una suerte de *todólogo*: es autogestor, creador, ejecutante, productor y cuando no esta tocando también es público.

> Los improvisadores se reúnen con otros músicos con los que nunca han tocado, músicos de otros países, de otras lenguas, de otras culturas, y el milagro puede suceder. Partituras invisibles vuelan sobre sus cabezas. Pero la libertad es exigente y tiene sus reglas: hay que oír al otro, hay que sentirlo.[11]

Tocar con otros es un factor imprescindible en las dinámicas de la improvisación libre, aunque no completamente necesario. Agrega un ingrediente más contundente de azar que el que ya existe al improvisar solo, o siempre con los mismos improvisadores. Asimismo, un elemento nuevo —sea una persona, un instrumento, un lugar, un público— provocará el desajuste y desequilibrio de los materiales o recursos de un improvisador, agregando nuevas posibilidades al tocar. Al entrar en

[10] Wade Matthews. ¡escucha! claves para entender la libre improvisación. 2001. URL www.wadematthews.info/ Fecha de consulta 15 de Mayo 2018.

[11] Chefa Alonso. *Improvisación libre la composición en movimiento.* 2007

contacto con otras subjetividades, realidades y posibilidades sonoras, el rumbo de la improvisación cambia inevitablemente, dando lugar a nuevas significaciones en los roles de la improvisación libre.

> Se ha dicho muchas veces que en la libre improvisación, cualquier instrumento puede asumir un papel melódico, armónico o rítmico, solista o de acompañamiento, etc. Suponemos que esto se considera digno de comentario porque la estructuración por papeles es tan rígida en algunas otras músicas como para que esta libertad llame la atención. Pero en la libre improvisación la realidad es aún más abierta, ya que algunos de estos papeles pueden ser completamente abandonados. Basta con que se construya en base a timbres y duraciones, variaciones de nivel dinámico, un juicioso empleo del silencio, etc.[12]

La improvisación es una creación que parte de la negociación con el otro —resultado de múltiples convergencias entre imaginación, escucha, subjetividad, percepción y acción— para aportar a la materia sonora elementos que le permitan trabajar con el tiempo y modelar el espacio. Si bien, no se puede ir para atrás en el tiempo, se puede ir atrás en la memoria, y es allí donde toman lugar las ideas que modificarán el presente de una improvisación. Desde estas lógicas va surgiendo una estructura al momento a la cual se puede regresar o referir si se quiere. El mismo escucha tiene esta posibilidad. En la improvisación colectiva se busca una co-independencia; existe un alto grado de autonomía pero esta depende de los demás, pues la autonomía de uno modifica la del otro y viceversa. Se trata de establecer

[12]Wade Matthews. Y la libre improvisación, qué tiene de improvisada. 2002. URL www.wadematthews.info/ Fecha de consulta 15 de Mayo 2018.

todo tipo de relaciones, no ponerse de acuerdo, no saber de antemano qué papel jugar, en qué momento regresar a un material previo, si seguir adelante proponiendo nuevos materiales o dejar de tocar y mejor escuchar.

> El improvisador, como hemos visto, tiene plena libertad. En primer lugar, está haciendo su pieza en el acto, con plena consciencia de las características del lugar donde la realiza. A su diálogo con los demás músicos, se añade su diálogo con el entorno: la acústica del lugar, los ruidos ambientales, lo atento o no que está el público. El improvisador es consciente de todos los sonidos que emiten los demás músicos, de cómo estos afectan lo que está tocando él, y de cómo quiere proceder en consecuencia. No es de extrañar que tenga esa misma sensibilidad y capacidad de reacción ante sonidos que no proceden de la intencionalidad de otro músico, pero que no por ello estén menos presentes en el espacio sonoro.[13]

Para Wade Matthews, esta es una música que requiere de una escucha atenta hacia los otros y al entorno sonoro ya que puede ser un factor que entre en juego y al mismo tiempo en resonancia con las interacciones sonoras de los improvisadores. Pero esto no es tarea fácil y se requiere de cierto tipo de entrenamiento y predisposición. Asimismo, comenta Matthews que para poder escuchar plenamente la música libremente improvisada sería necesario hacer a un lado otros criterios que se tengan sobre la escucha de otras músicas, y al mismo tiempo,

[13] Wade Matthews. ¡escucha! claves para entender la libre improvisación. 2001. URL www.wadematthews.info/ Fecha de consulta 15 de Mayo 2018.

comenzar a generar nuevos constructos que nos permitan acceder a la escucha atenta del mundo sonoro y de la improvisación libre, debido a que"nuestra experiencia de obras anteriores no solo nos enseña, sino que incluso impone una forma de acercarnos a otras de manera que nuestras propias expectativas nos esconden lo que realmente hay".[14] A este respecto, pienso que esta postura estaría imponiendo también una forma idealizada de escucha e incluso una postura que obliga al público a guardar silencio, al igual que en cualquier concierto de música académica. Además, en la actualidad sería difícil encontrar un público gustoso de deshacerse por completo de sus referentes musicales y hábitos de escucha para acceder a un encuentro que de por sí impone un silenciamiento de la voz. Por otra parte, es cierto que en algunos eventos de improvisación libre se puede transitar libremente por el espacio y se puede fumar o tomar, pero siempre en silencio. Entonces dónde queda la libertad del público para ser parte de ese proceso libertario de improvisación comprometida, tal vez, la forma para ser partícipes de esta música libremente improvisada debería permitir un involucramiento activo o no, del público a través de su escucha, su voz y sus ruidos, no debería exigirlo ni reglamentarlo.

Al escuchar improvisaciones en vivo y grabaciones propias o de otros músicos, he detectado pautas y tendencias estilísticas que enmarcan la forma en la que se hace improvisación libre y lo que se espera escuchar en ella, contrario a lo que menciona Chefa Alonso refiriéndose a que no hay normas en la improvisación libre. Dentro de esas pautas parece frecuente la evasión o rechazo hacia referentes provenientes de tradiciones musicales canónicas de occidente, orientadas por un pensamiento tonal, como son escalas, arpegios, melodías,

[14]Idem

secuencias de acordes, inflexiones o modulaciones, desarrollo temático y ritmos estables. Este último elemento puede estar presente en la improvisación libre, si bien, no como ritmos de subdivisiones binarias o ternarias, sí como pulsos semi estables o quebrados.

Entre las tendencias estilísticas que encuentro, enlisto las siguientes:

Modos de accionar objetos o instrumentos: raspar, frotar, tirar, rasgar, apretar, estirar, aflojar, quebrar, alterar, moldear.

Formas de distribuir el sonido entre los músicos: olas que van y vienen, interacciones casi aleatorias o secuenciadas.

Velocidad de interacciones o acciones sonoras en el tiempo: de lo más lento que pueda tocar a lo rápido.

Secuencias de armónicos: persistentes, ritmos estables o inestables.

Sonidos sostenidos/pedales/drones: fijos, largos, cortos, estables, cambiantes, distorsionados, metamorfoseados, camuflados

Tipos de ataques: constantes, intermitentes, estables, aleatorios, estocásticos, secos, resonantes.

Tipos de fraseo: bien delimitado, difuso, a través de cambios de volumen crescendos, drecrescendos.

Secuancias: secuencias de un solo material tocado de forma intermitente o varios materiales superpuestos tocados de forma intermitente

Cambios de timbre súbitos o transitorios: agregaciones o substracciones tímbricas.

Activación sonora del espacio u objetos que se encuentren en él: estos pueden incluir actos gestuales o movimientos corporales.

Transiciones a partir de cambios de materiales: modulantes, súbitas, de riesgo[15] o de inflexión.[16]

Estas son solo algunas de las posibilidades que he encontrado en mi experiencia como improvisador y escucha, solamente con éstas ya tenemos un gran abanico de múltiples opciones para improvisar libremente y más si se interacciona con un grupo de improvisadores. Es importante decir que algunas de estas tendencias son las que han permitido pensar en un sistema de clasificación de materiales de la improvisación libre como veremos en el siguiente apartado donde se explicará a detalle los elementos de la máquina que escucha y clasifica materiales de improvisaciones libres.

Antes de entrar al apartado siguiente me gustaría hacer la aclaración de que, debido a la complejidad del tema, y el trabajo que requiere subdividir distintos instrumentos en una grabación estéreo, en la fase actual del proyecto solo se ha trabajado con instrumentos solistas. Sin embargo, como hemos visto en en las anteriores discusiones, la improvisación libre cobra un mayor sentido cuando se hace en grupo, aunque no se limita a esta forma de interacción. ¿Qué pasa cuando se aproxima a la improvisación libre desde la práctica solista?

[15]El riesgo sería como estar por unos segundos en un estado de incertidumbre, incapaces de pronosticar exactamente que es lo que ocurrirá en el siguiente momento. Este riesgo puede ser favorable y conducir las reacciones de otros músicos a un lugar determinado —tal vez de mayor energía—, o desfavorable cuando esto lleva a la improvisación a un estado de estancamiento o naufragio, donde ya no se sabe como salir.

[16]Cambiar a un sonido sumamente diferente sin regresar a el, puede o no tomar un lugar importante en el discurso, lo colorea o lo modifica por un instante. Estos sonidos también pueden venir de la interacción con el entorno sonoro, y es más común que se den en la improvisación colectiva donde el azar alcanza una mayor presencia

Estaríamos hablando de una creación individual en tiempo real, en donde puede no haber un diálogo con otros músicos pero sí una estrategia de producción individual que ayude a formalizar un modelo para improvisar. La limitante en este modo de improvisación sería que esta se ve atravesada por la subjetividad y memoria individual del improvisador. El azar tiene un papel mucho menor, debido a que el improvisador tiene mayor control y la última palabra de lo que quiere sonar —o por lo menos lo que él produzca. Si trabaja con sistemas dinámicos, caóticos, algorítmicos computacionales o sistemas de instrumentos expandidos, podría de algún modo regresar el factor de azar e incertidumbre a la improvisación. Además, el factor del ambiente sonoro inevitablemente también estará presente pero depende del improvisador si genera un dialogo con este o lo deja pasar.

Algo que considero importante problematizar y poner a discusión es qué diferencia habría en la escucha de un neófito hacia la improvisación libre y una máquina que aprende a escuchar. En el caso de la máquina no hay expectativas sobre obras anteriores, ni una preconcepción de cómo debería escucharse la música o cómo tendría que ser interpretada. Dubnov y Assayag mencionan en su proyecto OMax, que más bien es un sistema agnóstico, que no cree en nada, ni tiene juicios de valor. La máquina discrimina entre géneros y, para el caso de este proyecto, entre materiales tímbricos de la improvisación, pero en su forma de discriminar no establece a priori juicios de valor. Por otro lado, un escucha neófito de la improvisación libre con todo su bagaje y hábitos hacia la escucha musical podría o no sentir inmediatamente un desagrado, un extrañamiento debido a sus referentes o gustos construidos, o vincularla con la música destinada a la ambientación de películas de terror o suspenso. A este respecto, Wade Matthews

propone un cambio de paradigma en la escucha para poder acceder a los significados que la libre improvisación esconde. Este cambio radica en la complejidad de la música misma y sus prácticas; en sus formas de producción colectiva y sus dinámicas cambiantes. Además Matthews propone tres tipos de escucha: la escucha focalizada en un gesto o instrumento particular, la escucha panorámica de la improvisación desde adentro o desde afuera de la práctica y la escucha del espacio acústico que envuelve a la agrupación.

Independientemente de que algunos discursos sobre la improvisación libre digan que es necesario hacer a un lado preconcepciones de escucha hacia otras músicas, la escucha en un humano nunca parte de cero, siempre está mediada primero por la escucha de productos sonoros culturales y después envuelta por fenómenos sonoros naturales o urbanos. Contrario a esto, la máquina sí podría tener la posibilidad de aprender relativamente desde cero, ya que puede aprender a distinguir solamente un único género o un grupo de géneros específicos de música, y, de ahí, generar su propia "inclinación" hacia el fenómeno musical, aunque al final sería realmente la inclinación de programador quien le da las instrucciones de aprendizaje a la máquina. La escucha humana es holística, interconectada, dinámica y adaptativa, puede abstraer la totalidad de un paisaje sonoro y puede focalizar para acceder o bloquear las micropartes que conforman ese paisaje. En cierta forma, la escucha sí está basada en el reconocimiento de sonidos a partir de fragmentos similares que ya han sido escuchados. Y aquí es donde esta el punto en común con la máquina y las motivaciones y acercamientos a la simulación de esa escucha.

Aún así, la escala y criterios de segmentación de materiales para entrenar a la máquina no son los mismos que los que los requiere un

humano para generar una forma de escucha, por ejemplo, aprender a distinguir la estructura de una canción. Las conexiones que puede hacer la escucha humana frente a la música son múltiples y complejas, al contrario de la máquina la cual aprenderá una forma o dos formas (o incluso más, si le son programadas) para distinguir y asociar materiales sonoros, dependiendo de cuál sea su entretamiento, pero lo más importante es que estas conexiones serían lineales, unívocas (a una determinada entrada corresponderá una determinada salida) a menos que, en el mejor de los casos, sea una máquina autónoma, adaptativa y abierta que pueda modificar sus formas de interpretar al momento. Si una máquina pudiera abstraer más sonidos, conocer más músicas y entenderlas podríamos interactuar con "máquinas músicos que han escuchado más música que nosotros".[17]

4.2. La máquina que escucha

Durante nuestra actividad como improvisadores en el colectivo Ruido 13, al cual pertenezco desde 2012, hemos detectado varios modos de interacción que mantienen un equilibrio y dan coherencia a las interacciones generadas entre varios músicos. Estos modos son: escucha, imitación, proposición (de una nueva idea musical), acompañamiento, ruptura y solo. De estos estados en la tesis de maestría me concentraré sólo en el primero: la escucha. Debido a la complejidad del tema, los otros modos de interacción quedan pendientes para un futuro desarrollo, el cual pretende usar los modelos de improvi-

[17]Nick Collins. Towards machine musicians who have listened to more music than us : audio database-led algorithmic criticism for automatic composition and live concert systems. *Computers in entertainment.*, 14(3):2, December 2016.

sación (descritos más adelante) obtenidos por el sistema de escucha automática, generados en esta fase de la investigación.

¿Porqué comenzar con el modo de escucha? El modo de escucha es el primer paso para la generación del sistema de improvisación automática por varias razones. La escucha es el acto de atender a uno o varios objetos sonoros específicos. Desde una escucha atenta se puede conocer y comprender el mundo musical y sonoro para poder transformarlo. Además, involucra su relación con el presente, activa la memoria de las acciones sonoras anteriores y ambas contribuyen a la realización al momento de una conciencia activa de escucha; en algunos casos, involucra una planeación subjetiva difusa o clara de lo que podría suceder en el futuro. La escucha atenta de la música libremente improvisada requiere, en los músicos, de un conocimiento y bagaje profundo de muchas otras músicas para encontrar puntos de quiebre, inflexión o equilibrio entre ellas que permitan crear algo nuevo, aunque esto no es del todo necesario. Un niño, una persona sin formación musical, ni conocimiento profundo sobre otras músicas, podría improvisar libremente. Desde mi experiencia al trabajar con niños o adultos lo único que se requiere es tener la disposición de escuchar al otro y estar abierto a la escucha del los sonidos circundantes del mundo. De manera similar, en el desarrollo de la máquina que escucha que aquí se plantea, se busca aprender a escuchar desde la misma música improvisada sin la necesidad de tener que conocer profundamente otras músicas. Para el caso de este sistema todo parte de la base de datos o corpus musical/sonoro con el que la máquina que escucha aprende a reconocer elementos del mismo corpus.

A continuación voy a hacer una relatoría de todos los momentos transcurridos a lo largo de esta investigación sin adentrarme profun-

damente en detalles técnicos, sino en las exploraciones y rasgos perceptivos, así como decisiones que fueron tomadas para el análisis de los sonidos complejos producidos en libres improvisaciones. Además explicaré cómo a través de la experimentación (prueba y error) y mi propia retroalimentación con la máquina que escucha, fueron paulatinamente perfeccionándose los procesos de clasificación, hasta llegar a un resultado más satisfactorio.

4.2.1. Identificación y segmentación de improvisaciones libres

En los primeros experimentos de la máquina que escucha, el desarrollo se focalizó principalmente en explorar de qué forma el timbre de una improvisación libre podría ser interpretado de manera certera por la máquina. Se seleccionaron diferentes ejemplos de grabaciones solistas de improvisadores libres como Okkyung Lee (chelo), Clare Cooper (arpa), Eddie Prévost (gong y objetos), Derek Bailey (exploraciones con feedback en guitarra eléctrica) y Akiyama (guitarra eléctrica). Estas grabaciones fueron segmentadas manualmente en pequeños pedazos, distinguiendo de cada uno de ellos frases o gestos.

Al intentar segmentar por gestos las improvisaciones libres, encontré que es una tarea sumamente difícil ya que de mi parte, hay un sesgo y subjetividad implícita de lo que considero un gesto o frase sonora, que evidentemente para otro puede no ser igual. Por mi parte yo entiendo al gesto como una unidad sonora que por sí misma o aislada de otras con un sentido o carácter que la hacen reconocible sin necesidad de recurrir a un contexto más amplio, sin embargo, este contexto en muchos eventos sonoros o musicales, debido a su carácter evolutivo,

resulta necesario para el entendimiento de una frase o gesto sonoro. Por ejemplo, en un sonido continuo que se transforma paulatinamente en otro sonido, resulta más complicado decidir dónde comienza y termina cada uno de los gestos.

En la improvisación de Okkyung Lee titulada *Stricly Vertical* es realmente difícil hablar de gestualidad ya que toda la improvisación está construida por un continuo fluir de ideas que se van transformando gradualmente como puede apreciarse en el espectrograma de la figura 4.1. De ahí que, en vez de segmentar en gestos o frases, comencé a segmentar de acuerdo con otros criterios como son los cambios tímbricos y cambios en los materiales usados por la improvisadora en su continuo fluir sonoro. A lo largo de esta improvisación fui encontrando diversas estrategias que me permitieron realizar la segmentación de la improvisación, estas fueron: cambios abruptos en la densidad sonora empleada, *agregaciones dinámicas*, —es decir, cuando a un primer material se agrega otro nuevo, de manera que se genera un contrapunto complejo de materiales—, cambios súbitos de velocidad, un gesto que conduce rápidamente a otro gesto (como una inercia por seguir tocando), ataques quebrados que poco a poco adquieren un ritmo estable. Por otro lado, hacia el final de la improvisación la idea de gesto vuelve a cobrar sentido nuevamente, ya que encuentro algunos gestos resaltados a través de pequeños cortes abruptos de fraseo, después de lo cual la improvisadora agrega materiales con otros espectros tímbricos, generando un contrapunto complejo que oscila entre la altura grave inicial y una segunda en el registro agudo. Al finalizar obtuve un total de 21 archivos de audio que corresponden a los gestos detectados.

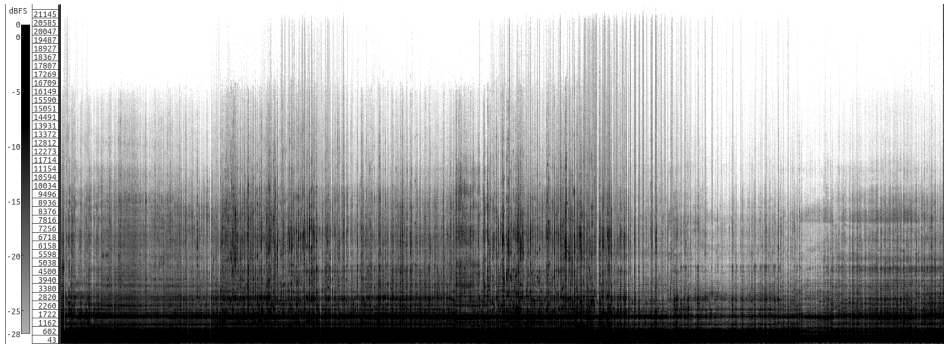

Figura 4.1: Espectrograma Strictly Vertical. Nota: todos los espectrogramas en este documento tienen la misma duración que la forma de onda que lo acompaña ya que están sincronizadas.

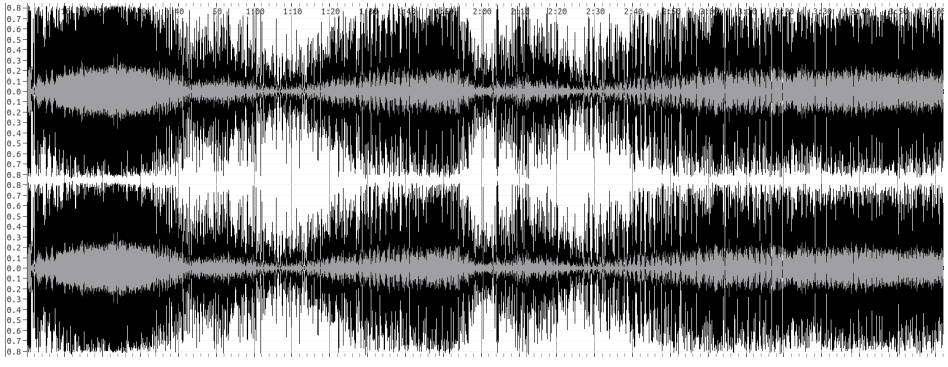

Figura 4.2: Forma de onda Strictly Vertical

Por otro lado, un ejemplo sencillo donde considero que habría un consenso entre diferentes personas al realizar la tarea de segmentar una libre improvisación por gestos, sería una interacción sonora dividida por silencios cortos y prolongados. Esta forma de interacción

puede ser bastante recurrente, un ejemplo de ello lo encontré en *Sink Into, Return* para arpa de Clare Cooper.

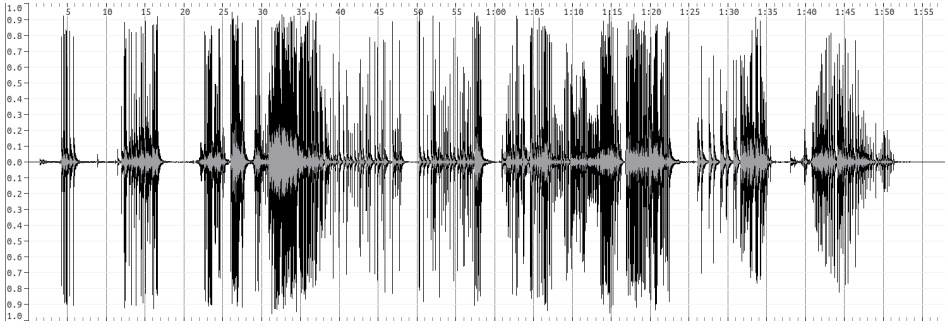

Figura 4.3: Forma de onda Sink Into, Return

Como puede apreciarse en la forma de onda de esta improvisación, Cooper es muy clara en sus inicios y finales de frases, además encuentro su improvisación muy acotada debido a que solamente usa cuatro materiales que permanecen constantes a lo largo de un minuto y cincuenta y siete segundos: 1) un sonido que me recuerda a una tarjeta de plástico que golpea las cuerdas al imprimir presión sobre ellas, 2) un juego entre tensión y distensión producida tal vez al estirar y aflojar la afinación de las cuerdas generando glisandos irregulares, 3) sonidos aislados percutidos con una amplitud baja y 4) el uso deliberado del silencio. Esto no quiere decir que no escuche las alturas, pero son tan complejas (debido a su cualidad ruidosa) y tan variadas que mi percepción tiende a enfocarse en los materiales generados por las técnicas empleadas. Finalmente me quedo con la percepción de un registro muy amplio que abarca todo el arpa. Mi escucha responde a la dinámica de interacciones, que encuentro siempre constantes, entre los cambios abruptos de silencio a sonido y las técnicas empleadas

por la improvisadora, de lo cual emerge la densidad sonora. Además, esta segmentación también se basa por momentos en la apreciación de cambios abruptos en el registro, la amplitud y el contenido armónico, aunque en otros casos es muy claro que precisamente esos cambios son parte del mismo gesto. El criterio que se tomó en cuenta en casos como el último es la secuencialidad energética que permanece constante. Al finalizar se obtuvieron 23 fragmentos de audio entre 0.35 segundos y casi 14 segundos.

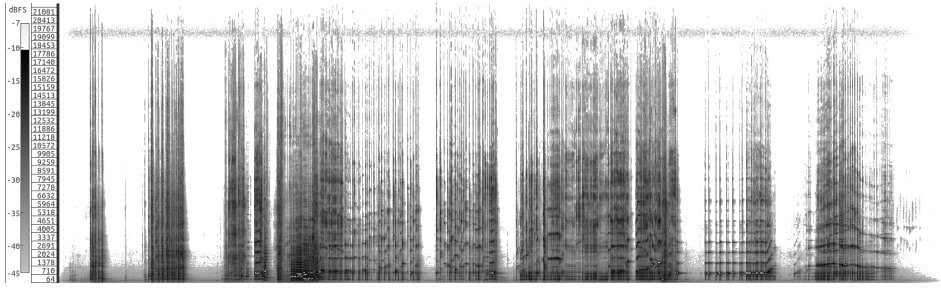

Figura 4.4: Espectrograma de Sink Into, Return

La siguiente pieza que segmenté fue *Scraped* de Eddie Prevost para tam-tam y objetos. Al escuchar por primera vez de manera atenta esta improvisación me queda la sensación de que el gesto en muchas improvisaciones libres se diluye, y más bien encuentro un continuo fluir caracterizado por pequeñas transiciones tímbricas. Aún así, encuentro todavía formas de segmentar la improvisación, aunque por momentos de forma más difusa e intrincada que en la improvisación de Clare Cooper. En algunos momentos Prevost usa el silencio para

conectar sus gestos, pero la mayoría están basados en estas continuas transformaciones tímbricas.

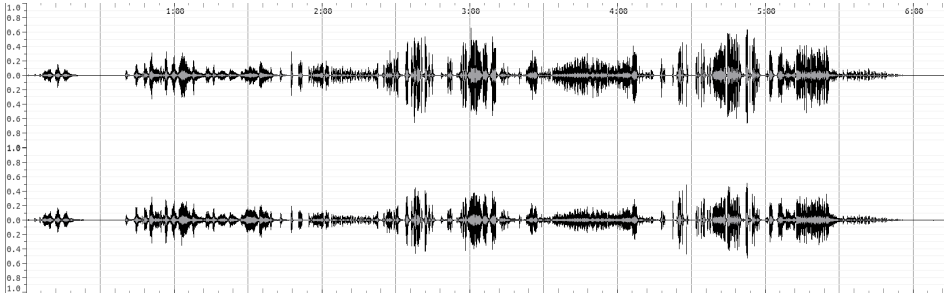

Figura 4.5: Forma de onda Scraped

Una pista realmente particular que me permitió segmentar esta improvisación fueron las alturas. Prevost usa dos armónicos intercalados casi todo el tiempo, que son contrastados con dos nuevas alturas. Estos cambios son controlados perfectamente, lo que ayuda a la detección de gestos diferentes en la pieza. Por momentos esta interacción sonora tiene un mayor dinamismo que en otros, lo que también ayuda a segmentar las ideas del improvisador. La exploración que hace del registro es también una buena guía ya que comienza con el registro medio, asciende a lo más agudo de su instrumento y comienza toda una nueva sección donde genera un contrapunto grave y oscuro, producido por un sonido chirriante, tal vez generado con una cuerda metálica, y la fricción del superball sobre el tam. Finalmente, me quedo con la sensación de una gran oleada que alberga otras oleadas más pequeñas, donde encuentro una elasticidad sonora orgánica y cohe-

rente. Siguiendo estos criterios generé un total de 39 gestos a lo largo de 6 minutos y 15 segundos.

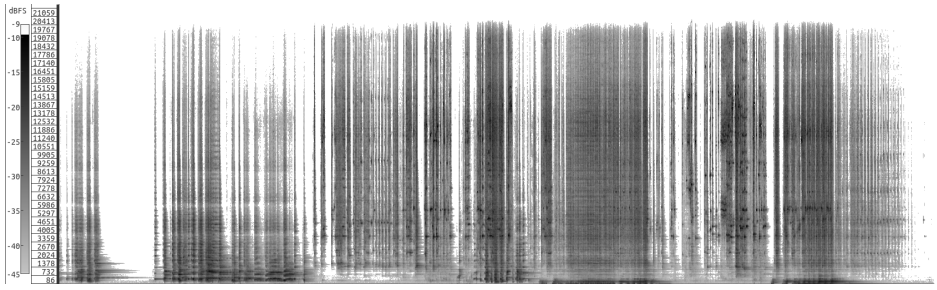

Figura 4.6: Espectrograma de Scraped

La improvisación de Derek Bailey titulada *f b (i) electric* es un feedback continuo que cambia progresivamente a través de la emergencia producida por la interacción de acercar y alejar la guitarra del amplificador. En ella la pauta que seguí para segmentar por gestos fue muy similar a *Strictly Vertical* de Okkyung Lee, es decir, identificar los momentos en que las alturas cambian o donde hay ligeras alteraciones al feedback producto de pequeños golpes que hace al puente, diapasón o a las cuerdas, así como reconocer los cambios de energía empleados por el improvisador.

Si bien esta improvisación podría ser considerada como un gran gesto de cuatro minutos, es necesario segmentarlo para que el algoritmo de clasificación tenga mayor información sobre la improvisación con la cual trabajar, de lo contrario, terminaría con un solo vector de 12 MFCCs derivado de la media de los vectores analizados en cada archivo de audio, lo cual sería muy poca información para el algoritmo de clasificación. Al proporcionar mayor información subdividida es posible generar un análisis más adecuado.

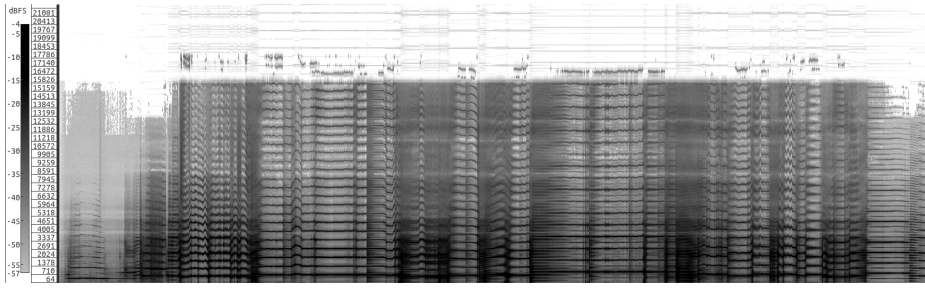

Figura 4.7: Espectrograma de f b (i) electric

El solo de guitarra eléctrica de Akiyama se caracteriza por presentar una idea del gesto bastante clara respecto a lo que entiendo como gesto, ya que opta por construir su improvisación partiendo de acciones sonoras divididas casi todo el tiempo por silencio, que oscila entre ataques sumamente agresivos e intervenciones muy sutiles que se llegan a fundir nuevamente con el silencio. De esta segmentación obtuve un total de 23 archivos que corresponde a los gestos sonoros detectados y 49 archivos que corresponden a los diferentes silencios producidos a lo largo de la improvisación. Debido a las características de esta improvisación, en la cual impera el uso del silencio, hice una clara distinción entre gestos sonoros y silencio, aunque éste no es absoluto debido al *hum* proveniente del amplificador del improvisador.

Debido a que en las improvisaciones es posible encontrar diferentes tipos de silencio, se generó otro grupo de archivos que responde a los distintos tipos de silencio producidos en las grabaciones de las improvisaciones libres, y que por las circunstancias de grabación pueden resultar en diferentes características acústicas.

Después de segmentar por gestos las improvisaciones solistas, se nombró cada archivo de audio, éstos se almacenaron en distintas car-

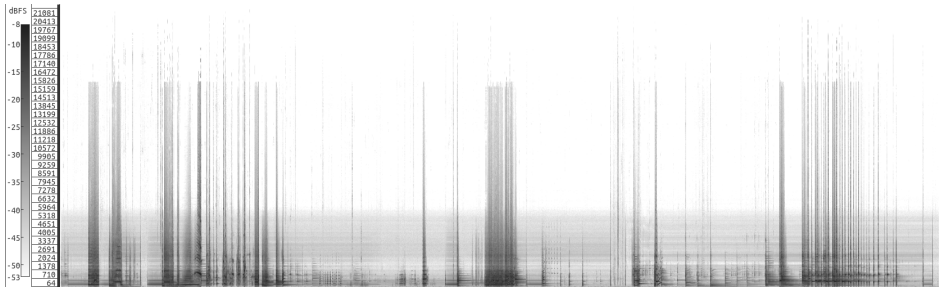

Figura 4.8: Espectrograma del solo de Akiyama

petas priorizando su división por tímbricas particulares que corresponden a los instrumentos usados por cada improvisador. Cabe mencionar que el objetivo de este experimento era encontrar la mejor combinación entre descriptores de audio y algoritmos de clasificación aplicados a la predicción tímbrica más adecuada de nuevas instancias.

Las herramientas que sirvieron para extraer la información de cada uno de los archivos de audio en esta aproximación fueron; la librería de SuperCollider creada por Nick Collins SCMIR (SuperCollider Music Information Retrival)[18] así como varios descriptores de audio, por ejemplo: MFCC, Loudness, SpecCentroid, SpecPcile, SpecFlatness, FFTCrest, FFTSpread, FFTSlope, SensoryDissonance.

El código utilizado para realizar la extracción de datos en los audios es derivado de uno de los ejemplos de Nick Collins que acompaña su librería, la cual puede ser descargada de la siguiente página.[19] Este código está diseñado para crear un archivo .arff derivado de los des-

[18] Nick Collins. SCMIR: A SuperCollider music information retrieval library. In *Proceedings of the International Computer Music Conference 2011*, pages 499–502, 2011.

[19] https://composerprogrammer.com/code.html Fecha de consulta 18 de Marzo 2018.

criptores extraídos de cada uno de los archivos de audio, los cuales fueron dispuestos en una estructura que alberga múltiples estructuras de datos (*array of arrays*) y que en este caso equivalen al número de clases en las cuales el sistema de clasificación aprenderá de la información dada. Así, se agregaron al código 5 carpetas enteras con un total de 179 archivos de audio segmentados por gestos, y una carpeta con 38 archivos de los tipos de silencios detectados (por mi escucha), listos para ser procesados por algunos algoritmos de clasificación de Weka. Cabe destacar que el código de SuperCollider obtiene el promedio de cada uno de los descriptores de los audios analizados, con lo cual tenemos la misma cantidad de datos (por ejemplo, 13 MFCCs) de todos los archivos de audio independientemente de su duración. Esta disposición de los datos extraídos es imprescindible para llevar a cabo el análisis debido a que los algoritmos de clasificación de Weka requieren que el tamaño de los vectores sea el mismo, de lo contrario, daría error al cargar el archivo.

4.2.2. Pruebas de clasificación con Weka

En esta aproximación se utilizó el aprendizaje supervisado descrito en el apartado 3.2.1 que consiste en dos fases: la fase de entrenamiento y la fase de prueba. Para realizar lo anterior, Weka cuenta con la opción que le permite dividir de forma aleatoria la base de datos, de manera que el 66 % de la información se destinó para entrenar al sistema y el 33 % se reservó para realizar las pruebas con el modelo generado. Con el fin de corroborar los porcentajes de certeza de predicción entregados, los algoritmos de clasificación fueron ejecutados un total de 10 veces variando el punto de inicio en cada corrida, de manera que Weka usa un generador aleatorio para revolver la infor-

mación en cada corrida y finalmente obtiene un promedio de las 10 ejecuciones del algoritmo.

La siguiente tabla se realizó partiendo de diferentes combinaciones entre los descriptores arriba señalados y varios algoritmos de clasificación. En la primer fila se muestran los descriptores de audio, los algoritmos utilizados, así como el porcentaje de certeza para clasificar los archivos utilizados, entre mayor sea el índice de certeza del algoritmo significa que hay una mejor clasificación sobre los diferentes tipos de improvisaciones. En las columnas se muestran los descriptores usados (marcados como 1, o 0 en caso de no haberse usado) así como distintas configuraciones de los mismos (marcados como más de 1). Estas configuraciones incluyen en el caso de *Spectral Percentile* diferentes magnitudes y en el caso de *FFTCrest* varias subdivisiones por bandas en el rango audible. Cada 3 algoritmos (MultilayerPerceptron (Percepción multicapa), DL4MlpClassifier (Aprendizaje Profundo) y K-Means) representan una corrida sobre toda la base de datos segmentada manualmente, manteniendo las mismas configuraciones entre los descriptores.

Es relevante realizar las pruebas con cada una de las configuraciones ya que evidencia la idea de la caja negra y el comportamiento de difícil predicción de los algoritmos de aprendizaje automático, por ejemplo, si se usan más de 3 MFCCs el índice de certeza baja, si se usan menos descriptores (por ejemplo menos de 3 SpecPcile) también baja. Asimismo, si se usan más épocas (la cantidad de veces que se va a iterar la información para entrenar al sistema, todos estos experimentos se realizaron con 20 épocas) o si el índice de aprendizaje varía en los algoritmos de clasificación (en los experimentos se uso 0.3), entonces los resultados pueden variar de forma significativa.

MFCC	SPECTRAL CENTROID	LOUDNESS	SPECTRAL PERCENTILE	SPECTRAL FLATNESS	FFT CREST	FFT SPREAD	FFT SLOPE	SENSORY DISSONANCE	ALGORITMO	PORCENTAJE
2	1	1	3	1	4	1	1	1	MultilayerPerceptron	91.75%
2	1	1	3	1	4	1	1	1	DL4MlpClassifier	89.17%
2	1	1	3	1	4	1	1	1	k-Means	59.53%
2	1	1	3	1	4	1	1	1	MultilayerPerceptron	77.20%
2	1	1	3	1	4	1	1	1	DL4MlpClassifier	67.25%
2	1	1	3	1	4	1	1	1	k-Means	48.57%
4	1	1	0	0	0	0	0	0	MultilayerPerceptron	60.20%
4	1	1	0	0	0	0	0	0	DL4MlpClassifier	55.10%
4	1	1	0	0	0	0	0	0	k-Means	36.60%
4	1	1	2	1	1	0	0	0	MultilayerPerceptron	71.88%
4	1	1	2	1	1	0	0	0	DL4MlpClassifier	65.97%
4	1	1	2	1	1	0	0	0	k-Means	37.30%
4	1	1	3	1	4	0	0	0	MultilayerPerceptron	71.90%
4	1	1	3	1	4	0	0	0	DL4MlpClassifier	66.26%
4	1	1	3	1	4	0	0	0	k-Means	40.52%
4	1	1	3	1	4	1	1	1	MultilayerPerceptron	70.90%
4	1	1	3	1	4	1	1	1	DL4MlpClassifier	74.22%
4	1	1	3	1	4	1	1	1	k-Means	42.70%
12	1	1	1	1	1	1	1	1	MultilayerPerceptron	70.90%
12	1	1	1	1	1	1	1	1	DL4MlpClassifier	71.40%
12	1	1	1	1	1	1	1	1	k-Means	48.00%
12	1	1	2	1	3	0	0	0	MultilayerPerceptron	71.60%
12	1	1	2	1	3	0	0	0	DL4MlpClassifier	63.00%
12	0	0	0	0	0	0	0	0	MultilayerPerceptron	38.00%
12	0	0	0	0	0	0	0	0	DL4MlpClassifier	36.00%
12	0	0	0	0	0	0	0	0	k-Means	37.80%
12	0	1	0	0	0	0	0	0	MultilayerPerceptron	39.40%
12	0	1	0	0	0	0	0	0	DL4MlpClassifier	38.70%
12	0	1	0	0	0	0	0	0	k-Means	30.30%
12	1	0	0	0	0	0	0	0	MultilayerPerceptron	53.00%
12	1	0	0	0	0	0	0	0	DL4MlpClassifier	50.00%
12	1	0	0	0	0	0	0	0	k-Means	40.60%
12	1	1	0	0	0	0	0	0	MultilayerPerceptron	51.20%
12	1	1	0	0	0	0	0	0	DL4MlpClassifier	48.60%
12	1	1	0	0	0	0	0	0	k-Means	36.00%
12	1	0	1	1	4	0	0	0	DL4MlpClassifier	59.42%
24	1	1	0	0	0	0	0	0	MultilayerPerceptron	47.20%
24	1	1	0	0	0	0	0	0	DL4MlpClassifier	42.30%
24	1	1	0	0	0	0	0	0	k-Means	28.00%

Figura 4.9: Combinaciones de descriptores y algoritmos para determinar la configuración más adecuada para realizar predicciones con nueva información

Como se puede observar en la tabla existe un punto de equilibrio entre el tipo, número y configuración de descriptores usados y de esto depende el resultado en la clasificación. De ahí que, si las configuraciones del algoritmo de clasificación y el número de descriptores de audio son los adecuados, es posible generar altos índices de predicción para nuevas instancias, y llegar a resultados satisfactorios en la clasificación del sistema como puede verse en las tres primeras filas de la tabla, esta configuración permitió obtener un porcentaje de certeza relativamente alto (91.75 % (Percepción multicapa), 89.17 % (Aprendizaje Profundo) y 59.53 % K-Means). Las configuraciones usadas en los distintos descriptores para obtener estos resultados fueron: [[MFCC, 2],[Loudness],[SpecCentroid],[SpecPcile, 0.90],[SpecPcile, 0.85],[SpecPcile, 0.75],[SpecFlatness],[FFTCrest],[FFTCrest, 0, 2000],[FFTCrest, 2000, 10000],[FFTCrest, 10000, 20000],[FFTSpread], [FFTSlope],[SensoryDissonance]].

Independientemente de los resultados obtenidos, esta aproximación puede ser limitada ya que no es posible obtener una retroalimentación más allá del porcentaje de certeza del algoritmo que dé cuenta de forma audible cómo el sistema clasifica los diferentes gestos de las improvisaciones. Sin embargo, las configuraciones empleadas y los resultados obtenidos pueden ser implementados en otro sistema que permita evaluar la clasificación a través del sonido.

Otra prueba realizada fue guardar el modelo generado por Weka, lo que permite abrirlo en otras computadoras y usarlo sin volver a realizar el análisis de clasificación. El modelo guardado fue llamado para probarse con una nueva base de datos usando otras improvisaciones de los mismos autores y del mismo disco. De esta forma el modelo generado podría ser evaluado con todo un arsenal de nueva información

proveniente de los mismos discos de los artistas antes mencionados. Al realizar esta prueba, el índice de certeza del modelo bajó considerablemente a 48.66 % por tres razones: primero, por la poca cantidad de archivos de audio con los que fue entrenado el modelo de clasificación; segundo, la clasificación personalizada de cada uno de los gestos puede contener errores, omisiones o malinterpretaciones, ya que las grabaciones pasaron por el filtro de mi oído, de modo que si había alguna equivocación mía al clasificar los ejemplos, el sistema sería susceptible a identificar erróneamente el material nuevo presentado. Tercero, porque la variabilidad tímbrica que puede tener un instrumento entre improvisaciones puede ser vastísima, esto no solo por la enorme cantidad de posibilidades técnicas que un solo instrumentista pueda tener sino también por el tipo de filtros y efectos usados en las grabaciones que pueden presentar cambios extremos, independientemente de que sea el mismo disco. Esto representa un serio problema ya que cada instrumento requeriría un análisis más que exhaustivo para tener un buen modelo que pueda describirlo. Los resultados completos de la evaluación pueden ser analizados en los anexos.

Esta aproximación solo podría ser mucho más acertada si se anotan y segmentan manualmente los gestos encontrados en diferentes improvisaciones libres. Lo que implica la construcción de una base de datos que incluya muchas (sino todas) las posibilidades que cada instrumento puede tener al ser tocado, y que permita definir modelos de improvisación libre basados en los diferentes componentes tímbricos. Para poder realizar lo anterior parece necesaria una metodología que implique el aprendizaje supervisado y sin supervisión.

Como hemos visto antes, las estrategias y los criterios que se tomaron en cuenta para segmentar por gestos una improvisación libre

implica un trabajo de programación sumamente complejo que excedería los límites de esta investigación. En su lugar se propone realizar una segmentación basada en la detección de ataques (*onsets*). Esta segmentación basada en pequeños fragmentos de audio puede proporcionar una mayor precisión al algoritmo de clasificación debido al grado de detalle que pudiera extraer de los diferentes fragmentos, focalizando y definiendo de forma microscópica la información presentada, generando así una mayor resolución para analizar y relacionar la información. De esta manera sería posible segmentar de forma automática improvisaciones de discos completos, usar aprendizaje sin supervisión (agrupamiento por K-means) para identificar los momentos de silencio, anotar manualmente los archivos generados y finalmente usar aprendizaje supervisado para correr nuevamente el experimento.

4.2.3. Pruebas con Python, Librosa y K-Means

Partiendo de las limitantes que tuvo la aproximación anterior para comprobar de forma audible los resultados obtenidos por el clasificador se construyó un sistema de reconocimiento de diferentes timbres y un programa de análisis de densidad sonora para la identificación de perfiles arquetípicos de la improvisación libre, basado en momentos de inicio (onsets) a través de la escucha y el aprendizaje automático.[20] Este sistema cuenta con con variables de entrada que serían las improvisaciones libres, propiedades específicas del sistema, tales como configuraciones de onsets, número de descriptores y de clases, y varia-

[20] Estos dos campos, si bien no dependientes uno del otro, son usualmente empleados a la par, aunque puede haber máquinas que escuchen sin necesariamente tener que aprender y puede haber aprendizaje sin máquinas que escuchen.

bles de salida que en este caso serían los resultados sonoros obtenidos, además hay un bucle de interacción entre el sistema y mi escucha al momento de supervisarlo ya que a partir de los resultados audibles, fui calibrando o modificando las distintas propiedades del sistema con el objetivo de intentar modelar mi escucha a través de la máquina.

Para la programación del sistema de la máquina que escucha se empleó el lenguaje de programación *Python*, la librería *Librosa* que posee varias herramientas de análisis de señales digitales de audio, y la libreria Tensorflow para clasificar los datos extraídos. Uno de los criterios para continuar con el desarrollo del proyecto fue usar el algoritmo de clasificación sin supervisión K-Means que permitió introducir indiscriminadamente al sistema cualquier cantidad de archivos de audio sin la necesidad de hacer previamente una clasificación manual de cada uno de los fragmentos. De este modo fue posible realizar pruebas desde un archivo de audio hasta un corpus musical mucho más grande y corroborar los resultados obtenidos respecto a las pruebas anteriores, donde el modelo de clasificación generado por Weka no fue lo suficientemente útil debido a la falta de retroalimentación sonora, ésta es muy importante ya que esclarece los criterios de clasificación del algoritmo de forma audible y ayuda a seleccionar las propiedades de los descriptores para intentar modelar a través de la máquina lo que percibo en la libre improvisación. Entonces se propuso generar un modelo descriptivo que permite visualizar cómo son organizadas las clases de timbres empleados en una improvisación libre a lo largo del tiempo y además incluir un modelo auditivo que sea de mayor utilidad para identificar si el sistema tiene o no un funcionamiento adecuado de acuerdo con mi escucha, de manera que la retroalimentación generada con estos dos modelos sea útil para optimizar o calibrar a la máqui-

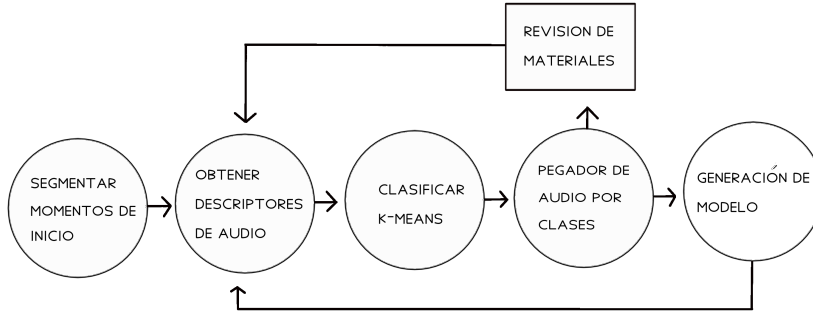

Figura 4.10: Fases del sistema de análisis de momentos sonoros de la improvisación libre: Mapa de proceso para la extracción y clasificación de audio

na que escucha. Se optó por realizar varios códigos por separado para mantener una conceptualización sencilla en cuanto a la programación, de ello se derivaron varias herramientas para el proceso de análisis de improvisaciones libres las cuales pueden ser descargadas directamente de GitHub para su consulta, estudio, modificación o aportaciones.[21] En la figura 4.10 se muestra el mapa del proceso seguido.

4.2.4. Segmentación

La parte de segmentación de audio consiste en la detección de todos los momentos de inicio (*onsets*) de los sonidos de cada audio. Para su detección se usó el objeto onset.detect de la librería Librosa. Este objeto devuelve una lista de todos los momentos de onsets detectados en el audio. Posteriormente, Python abre un archivo de texto el cual

[21] https://github.com/Atsintli/La-m-quina-que-escucha

será llenado por la lista de todos los segmentos detectados indicando el momento de inicio y final de cada onset (un cuadro —muestras por segundo— antes de que inicie el siguiente onset). Finalmente se exportaron todos los segmentos detectados a un archivo de texto y en una carpeta se exportaron los segmentos de audio del archivo original en formato wav. Este programa soporta audios tanto en formatos mp3 así como wav.

```
#Segmentación

import librosa
import sys
import os
import numpy as np
import matplotlib.pyplot as plt
import librosa.display

fileName = sys.argv[1]
fileJustName = fileName.split('.')[-2]
if not os.path.exists(fileJustName):
                os.makedirs(fileJustName)

y, sr = librosa.load(fileName)
onset_env = librosa.onset.onset_strength(y=y, sr=sr,
                                         hop_length=512,
                                          )
peaks = librosa.util.peak_pick(onset_env, 3, 10, 7, 7, 0.20, 0.01)
times = librosa.frames_to_samples(peaks)

file = open(fileJustName + '_durs.txt', 'w')
amountOfSegments = times.size
for cont in range(amountOfSegments - 1):
        posFrameInit = times[cont]
        posFrameEnd = times[cont + 1]
        duracionSeg = posFrameEnd - posFrameInit
        print("duracion de segmento = " + str(duracionSeg))
        file.write(str(duracionSeg) + "\n")
```

```
        librosa.output.write_wav(fileJustName + '/' + fileJustName + '_'
                + '{:05d}'.format(cont) + ".wav", y[posFrameInit:posFrameEnd], sr)
file.close()
posFrameInit = times[amountOfSegments-1]
posFrameEnd = y.size
librosa.output.write_wav(fileJustName + '/' + fileJustName + '_'
        + '{:05d}'.format(amountOfSegments-1) + ".wav", y[posFrameInit:posFrameEnd], sr)

times = librosa.samples_to_time(times)
amountOfSegments = times.size

file = open(fileJustName + '_times.txt', 'w')
for cont in range(amountOfSegments -1):
        posFrameInit = times[cont]
        file.write(str(posFrameInit) + "\n")
file.close()

print ("numero de segmentos = " + str(amountOfSegments))
#Plot
timess = librosa.frames_to_time(np.arange(len(onset_env)),
                                sr=sr, hop_length=512)
plt.figure()
ax = plt.subplot(2, 1, 2)
D = librosa.stft(y)
librosa.display.specshow(librosa.amplitude_to_db(D, ref=np.max),
                         y_axis='log', x_axis='time')
plt.subplot(2, 1, 1, sharex=ax)
plt.plot(timess, onset_env, alpha=0.8, label='Onset strength')
plt.vlines(timess[peaks], 0,
           onset_env.max(), color='r', alpha=0.8,
           label='Selected peaks')
plt.legend(frameon=True, framealpha=0.8)
plt.axis('tight')
plt.tight_layout()
plt.show()
```

4.2.5. Extracción

En la siguiente sección se extrajeron las características que describen de forma tímbrica cada uno de los de los fragmentos de audio generados en el paso anterior. En este proceso se hicieron pruebas con descriptores de audio como MFCC, espectrograma en la escala Mel, centroide espectral, contraste espectral y contraste de ancho de banda, para determinar cuál combinación de descriptores generaba resultados más coherentes al momento de la clasificación. Estos pueden ser representados como vectores de dimensiones variables, es decir que pueden contener mayor o menor información. El algoritmo que se propuso fue el siguiente: 1) leer todos los archivos de audio generados en el proceso anterior; 2) extraer los valores del espectro normalizado en valores absolutos con la transformada rápida de Fourier (esta función solo será aplicada en el contraste espectral); 3) extraer la media de cada uno de los vectores de los descriptores usados;[22] 4) concatenar cada uno de los vectores en una lista y finalmente escribir la lista de cada archivo de audio en un archivo de texto que pueda ser leído posteriormente por el clasificador K-Means.

```
#Extracción

import librosa
```

[22]Obtener la media de cada uno de los descriptores garantiza tener vectores de las mismas dimensiones para diferentes tamaños de archivos de audio e información. Por ejemplo, si en el caso del MFCC se indicó al programa usar 12 vectores, el contenido de estos puede variar dependiendo de la longitud del archivo y la cantidad de información que contenga, al sacar la media terminaríamos con 12 vectores cada uno con un valor que representa la cantidad de energía encontrada en cada uno de los cuadros de la muestra de audio.

```
import glob
import sys
import numpy as np

def extract(fileName):
        fileJustName = fileName.split('.')[-2]
        print(fileJustName + ' Proceced')
        y, sr = librosa.load(fileName)
        stft = np.abs(librosa.stft(y))
        mfccs = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=60).T,axis=0)
        #chroma = np.mean(librosa.feature.chroma_stft(S=stft, sr=sr).T,axis=0)
        #mel = np.mean(librosa.feature.melspectrogram(y, sr=sr).T,axis=0)
        #contrast = np.mean(librosa.feature.spectral_contrast(S=stft, sr=sr).T,axis=0)
        #rms = np.mean(librosa.feature.rmse(y=y,).T,axis=0)
        #cent = np.mean(librosa.feature.spectral_centroid(y=y, sr=sr).T, axis=0)
        #allFeature = np.concatenate((mfccs,chroma,mel,contrast,rms,cent))
        allFeature = mfccs
        allFeature = map(str, allFeature)
        file = open(fileJustName + '.txt', 'w')
        file.write(', '.join(allFeature))
        file.close()
folder = sys.argv[1]
for file in glob.glob(folder + '/*.wav'):
        extract(file)
```

Es importante en este paso mantener un uso acotado de los descriptores con la finalidad de que la cantidad de información extraída sea la mínima necesaria para obtener resultados coherentes en el proceso de clasificación y no generar ruido al incluir más descriptores ya que pueden confundir al algoritmo de clasificación. A continuación un ejemplo de los descriptores MFCCs extraídos de Sink into Return:

```
-171.0666688, 53.2165495172, 0.00200234677783, 0.00191597549245, 0.00214453310339,
0.00527150875621, 0.0197066184522, 0.00601303218952, 0.00386559367895, 0.009616449500,
0.0171364816208, 0.0332278038208, 0.151078531753, 0.605798112197, 1.75344814496,
1.33227143601, 1.08364838024, 0.153504411889, 0.202100707189, 0.135147203363,
```

```
8.2861798818e-05, 19.8567688021, 18.5945683267, 14.9633150468, 22.1317683406,
22.8998839088, 16.4432639688, 29.872423609, 2437.85475999, 3045.48644169
```

4.2.6. Clasificación

El siguiente paso involucró la programación del algoritmo para clasificar los datos extraídos de los fragmentos de audio con el clasificador K-Means. El código de clasificación comprende: 1) seleccionar el número de clases que se quieren obtener de la totalidad del audio; 2) seleccionar el número de iteraciones a realizar, es decir el número de veces que se entrenará al sistema para clasificar los datos; 3) leer la información extraída por los descriptores de audio a partir de todos los archivos de texto, generados en el paso anterior; 4) disponer los datos en una donde se compara el nombre de cada archivo con los datos de los vectores obtenidos; 5) seleccionar un centro inicial a partir de los datos extraídos para agrupar los demás valores; 6) agrupar con sus vecinos más cercanos/parecidos cada uno de los vectores de datos analizados; 7) actualizar el centro de los grupos hasta encontrar su punto medio o masa central; 8) iterar el número de veces seleccionado hasta agrupar todos los vectores en el número de clases deseadas.

```
#Clasificación

import tensorflow as tf
import numpy as np
import sys
import glob

k = 3
max_iterations = 100
folder = sys.argv[1]
```

```
def loadData(xs, names):
    for fileName in glob.glob(folder + '/*.txt'):
        file = open(fileName, 'r')
        print(fileName)
        names.append(fileName)
        content = file.readlines()
        data = content[0].split(',')
        data = map(float, data)
        print(data)
        xs.append(data)

def get_dataset():
    xs = list()
    names = list()
    loadData(xs, names)
    xs = np.asmatrix(xs)
    return xs, names

def initial_cluster_centroids(X, k):
    return X[0:k, :]

def assign_cluster(X, centroids):
    expanded_vectors = tf.expand_dims(X, 0)
    expanded_centroids = tf.expand_dims(centroids, 1)
    distances = tf.reduce_sum(tf.square(tf.subtract(expanded_vectors,
        expanded_centroids)), 2)
    mins = tf.argmin(distances, 0)
    return mins

def recompute_centroids(X, Y):
    sums = tf.unsorted_segment_sum(X, Y, k)
    counts = tf.unsorted_segment_sum(tf.ones_like(X), Y, k)
    return sums / counts

with tf.Session() as sess:
    sess.run(tf.local_variables_initializer())
    X, names = get_dataset()
    centroids = initial_cluster_centroids(X, k)
    i, converged = 0, False
```

```
    while not converged and i < max_iterations:
        i += 1
        Y = assign_cluster(X, centroids)
        centroids = sess.run(recompute_centroids(X, Y))
    results = zip(sess.run(Y), names)
    results.sort(key=lambda tup: tup[1])
    file = open(folder + "_clases.txt", "w")
    for res in results:
        print("clase " + str(res[0]) + " segmento " + res[1])
        segmento = res[1].split("_")[1].split(".")[0]
        file.write(str(res[0]) + " " + segmento + "\n")
    file.close()
```

4.2.7. Generación de modelo conceptual

El siguiente paso fue corroborar los resultados obtenidos por el clasificador. Esta tarea se realizó en dos fases; primero, los datos obtenidos del clasificador fueron depositados en un gráfico en donde es posible visualizar el número de materiales (o clases, para propósitos del análisis) y cómo fueron usados a través del tiempo, generando una estructura tímbrica temporal dividida por clases. Al mismo tiempo se generó una gráfica de pastel que muestra el porcentaje de apariciones de cada clase; esta información fue útil para establecer arquetipos que denotan la forma de tocar de un intérprete respecto a los materiales empleados.

```
#Estructura

import sys
import numpy as np
import matplotlib.pyplot as plt
from pylab import *
```

```
FACTOR_DE_REDUCCION = 256
audioname = sys.argv[1]
clases = open(audioname + "_clases.txt")
durs = open(audioname + "_durs.txt")
clasescontenttmp = clases.readlines()
durscontenttmp = durs.readlines()
clasescontent = []
durscontent = []
for item in clasescontenttmp:
        clasescontent.append(int(item.split(" ")[0]))
for item in durscontenttmp:
        durscontent.append(int(item) / FACTOR_DE_REDUCCION)

cantClases = max(clasescontent) + 1
reparticionDeClases = np.zeros(cantClases)
nombreDeClases = []

estructura = []
for clase, dur in zip(clasescontent, durscontent):
        reparticionDeClases[clase] += dur
        estructura.append([clase] * dur)
estructura = [item for sublist in estructura for item in sublist]

for n in range(cantClases):
        nombreDeClases.append("Clase " + str(n))

reparticionDeClases = reparticionDeClases / sum(reparticionDeClases)

eventData = [[] for i in range(cantClases)]
posEnSamplesCFR = 0
for clase, dur in zip(clasescontent, durscontent):
        for i in range(dur):
                eventData[clase].append(posEnSamplesCFR + i)
        posEnSamplesCFR += dur

cantClasessec = []
for i in range(cantClases):
        cantClasessec.append((cantClases - cantClases) + i)

#Plot
```

```
plt.figure(1)
plt.subplot(121)
plt.xlabel('Segundos')
plt.ylabel('Clases')
samplesReales = posEnSamplesCFR * FACTOR_DE_REDUCCION
segundosINT = (int)(samplesReales / 22050.0)
puntos = range(segundosINT)[0::10]
posiciones = np.array(puntos) * 22050 / FACTOR_DE_REDUCCION
plt.xticks(posiciones,puntos)
plt.autoscale(enable=True, axis='x', tight=True)
plt.autoscale(enable=True, axis='y', tight=True)
plt.gca().yaxis.set_major_locator(MaxNLocator(integer=True))
plt.yticks(cantClasessec, cantClasessec)
plt.eventplot(eventData, linelengths=1, lineoffsets=1)
limon = plt.subplot(122)
plt.pie(reparticionDeClases, labels=nombreDeClases, autopct='%1.1f%%')
plt.axis('equal')
plt.show()
```

Segundo, se realizó un código que pega o unifica todos los segmentos de los archivos de audio, generados en la primera fase, en la cantidad de clases elegida, agrupando los archivos similares en un solo archivo de audio. Es decir, si se obtuvieron mil archivos de audio y se seleccionó clasificarlos en 18 clases, se obtienen 18 archivos de audio, donde cada uno de los archivos corresponden timbricamente a una clase específica. Esto sirvió para generar un modelo de comprobación auditiva que permitió determinar si era preciso o no el algoritmo de clasificación, ya que al agrupar los segmentos de audio tímbricamente parecidos entre sí resultó más fácil identificar si los descriptores de audio seleccionados eran los adecuados.

```
#Unificador
```

```
import librosa
import sys
import numpy as np

NUM_MAX_CLASES = 8
audioname = sys.argv[1]
clases = open(audioname + "_clases.txt")
clasescontenttmp = clases.readlines()
clasescontent = []
for item in clasescontenttmp:
        clasescontent.append(int(item.split(" ")[0]))
print(clasescontent)
for clase in range(NUM_MAX_CLASES):
        print("iterando sobre " + str(clase))
        ele = np.where(np.array(clasescontent)==clase)[0]
        print("indices de clase " + str(clase) + " son ")
        print(ele)
        audiototal = np.array([])
        for elements in ele:
                conStr = '{:05d}'.format(elements)
                nomArchivo = audioname + "/" + audioname + "_" + conStr + ".wav"
                print("leyendo " + nomArchivo)
                y, sr = librosa.load(nomArchivo)
                audiototal = np.append(audiototal,y)
        librosa.output.write_wav(audioname + "/" + audioname + "_CLASE_"
                + str(clase) + ".wav", audiototal,sr)
```

Asimismo, se obtuvo obtuvo una matriz de correlación disponible en la librería *Librosa* en la cual se puede observar de manera concreta cómo algunos de los elementos de la improvisación se están repitiendo o no a lo largo del tiempo, como puede observarse en la figura 4.11. Si bien esta no es una aproximación basada en el algoritmo de clasificación, puede ser una estrategia para seleccionar "k" o corroborar cuántas clases hay en un archivo de audio o en la totalidad del corpus con el que se esté trabajando.

```
#Matriz de correlación

import librosa
import librosa.display
import matplotlib.pyplot as plt
import sys

audioname = sys.argv[1]

y, sr = librosa.load(audioname)
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=12)

# Find nearest neighbors in MFCC space
R1 = librosa.segment.recurrence_matrix(mfcc)

# Or fix the number of nearest neighbors to 5
R2 = librosa.segment.recurrence_matrix(mfcc, k=5)

# Suppress neighbors within +- 7 samples
R3 = librosa.segment.recurrence_matrix(mfcc, width=7)

# Use cosine similarity instead of Euclidean distance
R4 = librosa.segment.recurrence_matrix(mfcc, metric='cosine')

# Require mutual nearest neighbors
R5 = librosa.segment.recurrence_matrix(mfcc, sym=True)

# Use an affinity matrix instead of binary connectivity
R6 = librosa.segment.recurrence_matrix(mfcc, mode='affinity')

# Plot the feature and recurrence matrices
plt.figure(figsize=(8, 4))
plt.subplot(2, 3, 1)
librosa.display.specshow(R1, x_axis='time', y_axis='time')
plt.title('Binary recurrence (symmetric)')

plt.subplot(2, 3, 2)
librosa.display.specshow(R2, x_axis='time', y_axis='time')
plt.title('Binary recurrence (symmetric)')

plt.subplot(2, 3, 3)
librosa.display.specshow(R3, x_axis='time', y_axis='time')
plt.title('Binary recurrence (symmetric) DIF')
```

```
plt.subplot(2, 3, 4)
librosa.display.specshow(R4, x_axis='time', y_axis='time')
plt.title('Binary recurrence (symmetric)')

plt.subplot(2, 3, 5)
librosa.display.specshow(R5, x_axis='time', y_axis='time')
plt.title('Binary recurrence (symmetric)')

plt.subplot(2, 3, 6)
librosa.display.specshow(R6, x_axis='time', y_axis='time', cmap='magma_r')
plt.title('Affinity recurrence')

plt.tight_layout()
plt.show()
```

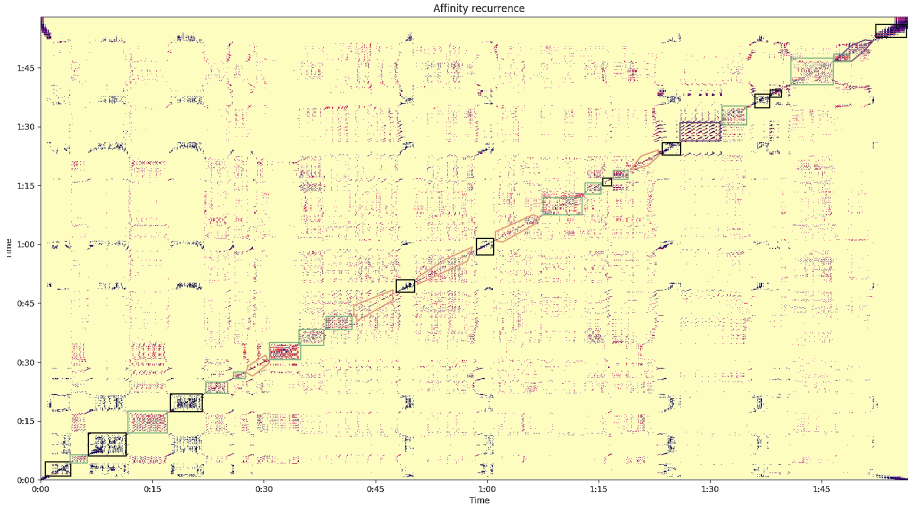

Figura 4.11: Ejemplo de matriz de correlación donde pueden identificarse 4 clases diferentes

4.2.8. Análisis de la densidad sonora

Finalmente, se generó un pequeño programa para hacer un análisis de la densidad sonora entre dos improvisaciones. Se optó por una solución bastante primitiva pero que sirve a corto plazo para determinar el grado de densidad que hay en momentos específicos de la improvisación y, de igual forma que el paso anterior, generar arquetipos de la improvisación basados en el análisis de la densidad usada por distintos improvisadores.

El algoritmo llama los archivos de audio originales (sin segmentar) y calcula cuántos ataques hay cada cinco segundos (aunque esta regla puede variar, siendo el análisis de ataques cada n segundos). A partir de este cálculo genera una lista que representa la cantidad de ataques que hubo por cada segmento. Después la lista es comparada con todas las listas generadas de otros ejemplos de audio de un mismo improvisador. Para hacer la comparación de ejemplos de audio con distintos tamaños se optó por llenar de ceros los lugares en donde ya no había información sonora en los audios más cortos, de manera que al hacer la comprobación con varios archivos de audio, pudiera haber coherencia en los tamaños de los archivos al compararlos mutuamente. La comparación se realizó con una simple resta y después una suma de todos los valores resultantes, de manera que se obtiene un promedio de similaridad entre las distintas aproximaciones al parámetro de densidad en un improvisador así como entre distintos improvisadores.

```
# Analisis de densidad

import librosa
import sys
```

```
import os
import numpy as np
import glob
import math

fileName1 = sys.argv[1]
fileName2 = sys.argv[2]

#slice time are
FramesSec = 22050.0
SecSlice = 10
SecSliceFrames = SecSlice * FramesSec

def getBlocks(filename):
        y, sr = librosa.load(filename)
        dur = y.size / FramesSec
        blocks_amount = (int)(math.ceil(dur / SecSlice))
        print(dur)
        print(blocks_amount)
        onset_frames = librosa.onset.onset_detect(y=y, sr=sr, units='samples')
         / FramesSec
        blocks = np.zeros(blocks_amount)
        for onset in onset_frames:
                indexofonset = (int)(onset / SecSlice)
                blocks[indexofonset] = blocks[indexofonset] + 1
        return blocks

b1 = getBlocks(fileName1)
b2 = getBlocks(fileName2)

#iguala tamanios
difsize = b1.size - b2.size
if difsize < 0:
        b1 = np.append(b1, np.zeros(abs(difsize)))
if difsize > 0:
        b2 = np.append(b2, np.zeros(difsize))

dif = np.fabs(b1 - b2) #Compute the absolute values element-wise.
dif = np.sum(dif)
print ("La diferencia entre las densidades sonoras de los archivos es de " + str(dif))
```

4.2.9. Resultados obtenidos con Python y K-Means

Para probar los códigos descritos anteriormente, se hicieron cuatro experimentos; en el primero se analizó *Sink into Return* de Clare Cooper; el segundo con un solo de guitarra eléctrica de Tetuzi Akiyama; el tercero con una improvisación de Fernando Vigueraz, *Coral Continuo*; y el cuarto un corpus de algunas improvisaciones solistas bastante disimiles entre sí, aquí se usaron las tres improvisaciones anteriores más otras de autores como Derek Bailey, Remi Alvarez, Juan Pablo Villa, Okkyung Lee, Wilfrido Terrazas, Eli Kesler, Nicolas Collins, Yan Leguay, Mike Majkowski y Maja Ratkje.

Antes de seguir adelante, hay que mencionar que la idea de gesto propuesta en secciones anteriores de este mismo capítulo, es un poco trivial para el reconocimiento y clasificación de la máquina. Similar a lo que sucede en el análisis de imágenes donde se requiere analizar pixel por pixel para generar una conceptualización que sea interpretada por la máquina, en el caso del audio sucede algo similar. Entre mayor sea la resolución de los cortes, el sistema podría identificar y clasificar de forma más certera los diferentes componentes sonoros de una improvisación. De manera que si un ataque esta seguido inmediatamente de un silencio, el programa de segmentación sin duda debería disociar entre ambos casos. A este respecto, se optó por segmentar la siguiente improvisación en la mayor cantidad de onsets detectados.

Con el fin de comprobar qué tan exacto podría ser el sistema en su totalidad se realizó un experimento que consistió en la identificación entre dos elementos muy sencillos de *Sink into Return*: sonidos fuertes y silencios o sonidos casi silentes. En el caso de la improvisación de

Clare Cooper se obtuvieron un total de 1044 archivos en segmentos de audio que van de 0.090 hasta 0.210 segundos.[23] A continuación se muestra una representación gráfica de como fue segmentado el archivo, donde las lineas rojas de la figura 4.12 representan todos los onsets detectados por el algoritmo (aunque más bien se aprecian como franjas en esta visualización, debido a la enorme cantidad de lineas y la resolución pequeña de la imagen) . En la figura 4.13 se muestra el mismo criterio de visualización pero ahora amplificado, esto permite observar cómo el algoritmo de segmentación fue cortando los distintos fragmentos de audio, asimismo se muestra un espectrograma para su comparación en ambos casos.

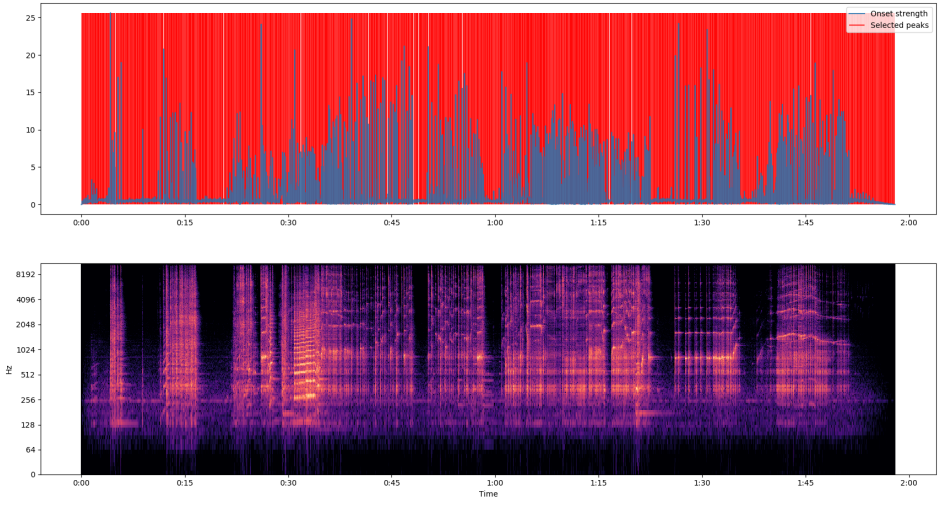

Figura 4.12: Onsets encontrados por el segmentador en Sink into Return

[23]La configuración modificada en el segmentador para alcanzar este número de fragmentos fue peaks = librosa.util.peak_pick(onset_env, 1, 1, 1, 1, 0.0945, 1)

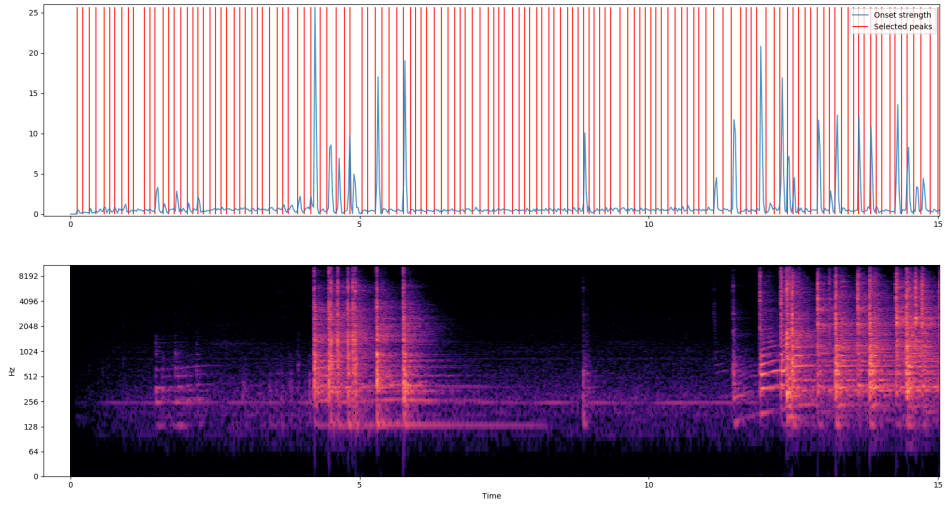

Figura 4.13: Visualización amplificada de los primeros 15 segundos

Posteriormente se decidió probar con 2, 13 y 60 vectores MFCCs para analizar cada uno de los fragmentos, debido a que en las evaluaciones auditivas si se empleaban el contraste espectral y el centroide espectral, el sistema de clasificación comenzaba a confundir y mezclar información sonora de forma errónea de acuerdo con mi escucha. Después se entrenó el sistema de clasificación con los archivos generados por el extractor de características declarando dos clases en el código de clasificación y 200 iteraciones, de manera que el sistema colocara dentro de una clase las muestras silentes y las densamente sonoras en otra. Los resultados sonoros fueron sumamente similares entre cada una de las configuraciones con 2, 13 y 60 vectores MFCCs, teniendo variaciones mínimas por lo que se optó seguir trabajando solo con 2 vectores MFCCs debido a que se simplifica y se hacen más rápidos los cálculos de los procesos involucrados. Estos resultados pueden ser

escuchados en el siguiente enlace a partir de la improvisación de Clare Cooper.[24] A continuación se muestran las formas de onda generadas por los descriptores en donde puede apreciarse las decisiones que tomó el algoritmo de clasificación de acuerdo con los criterios señalados.

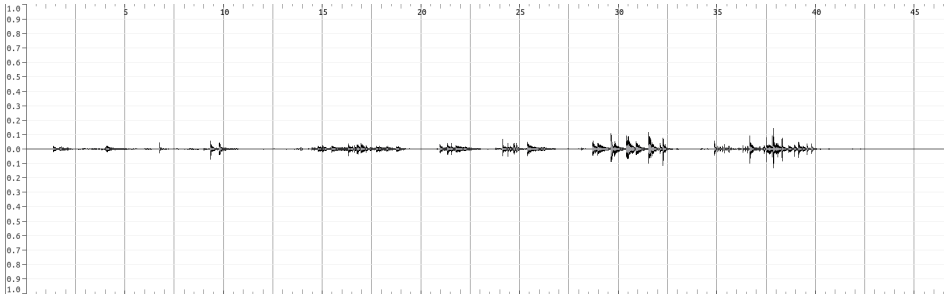

Figura 4.14: Clase 0 compuesta por sonidos en *pppp* y silencios

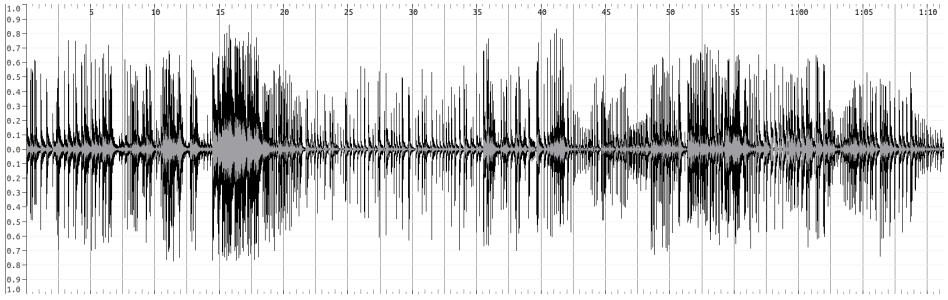

Figura 4.15: Clase 1 compuesta por sonidos superiores a *pppp*

Es importante entender porqué el algoritmo de clasificación tomó la decisión de agrupar los sonidos casi silentes y los silencios juntos en lugar de clasificar todos los silencios en un solo archivo y clasificar

[24] https://archive.org/details/pruebas_con_distintos_descriptores

los sonidos casi silentes y los más densos en otro. La respuesta a esto es que los valores encontrados en los archivos casi silentes y los que realmente contienen silencio, después de haber sacado la media de cada uno de los MFCCs encontrados, resultan más similares entre si que los densamente sonoros y los casi silentes. Es por esto que el algoritmo decidió clasificarlos de esta manera. Adicionalmente se muestra la estructura temporal por clases dibujada por el programa y una gráfica de pastel que da cuenta de los porcentajes en los que las diferentes clases fueron empleadas a lo largo de la improvisación, donde la clase 0 son los sonidos casi silentes y silencios y la clase 1 son los sonidos más densos. Adicionalmente se comparten los audios generados por el sistema en la siguiente liga.[25]

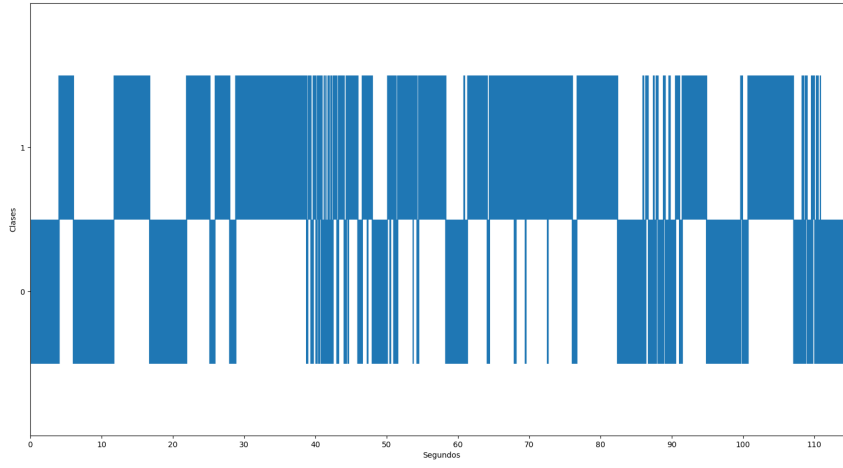

Figura 4.16: Estructura temporal de *Sink into Return* dividida en 2 clases

[25] https://archive.org/details/sink_into_return-Clare_Cooper_CLASE_1_2MFCC/sink_into_return-Clare_Cooper_CLASE_1_2MFCC.wav

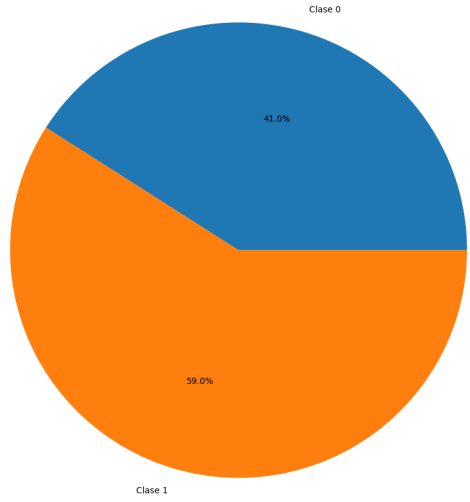

Figura 4.17: Pastel de 2 clases

Posteriormente se realizó la misma prueba pero ahora con 3 clases, aquí el resultado fue distinto ya que los sonidos que tienden a ser silenciosos están en la clase 0, los que tienden a ir de *mf-ff* están la clase 1 y los sonidos que tienden a una dinámica entre *pp-mf* están en la clase 2. Es interesante notar que dentro de la clase 2 también están presentes las resonancias que quedan después de un ataque, de manera que solo queda el remanente de estos en su desenvolvimiento acústico por el espacio. Estos casos pueden escucharse en el siguiente enlace.[26] Asimismo se muestra la estructura temporal dividida en tres clases y el gráfico de pastel correspondiente.

[26] https://archive.org/details/sink_into_return-Clare_Cooper_CLASE_1

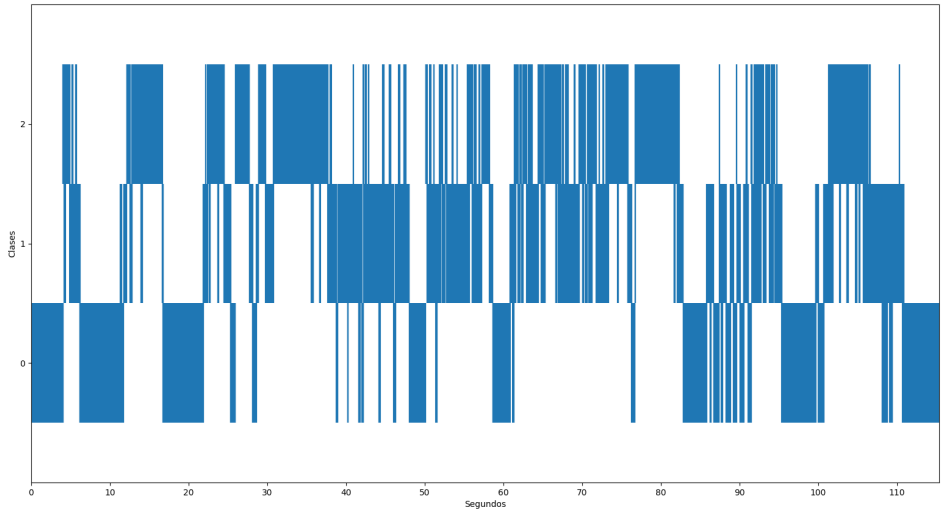

Figura 4.18: Estructura temporal en 3 clases

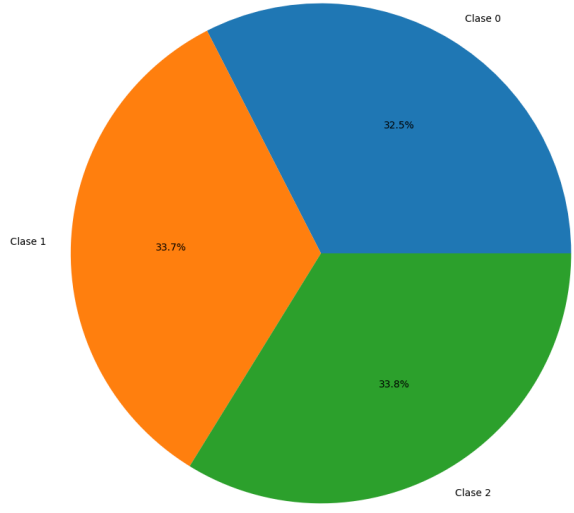

Figura 4.19: Pastel de 3 clases

Después se realizó una clasificación en 4 clases de acuerdo con lo que percibí anteriormente al analizar esta improvisación: 1) el uso deliberado del silencio, 2) un sonido que me recuerda a una tarjeta de plástico que golpea las cuerdas al imprimir presión sobre ellas, 3) sonidos aislados percutidos con una amplitud baja, 4) un juego entre tensión y distensión producida tal vez al estirar y aflojar la afinación de las cuerdas generando glisandos irregulares.

Podría decirse que el resultado en la clasificación coincide en los dos primeros aspectos detectados que corresponden a la clase 0 y 1. La clase 2 tiene sonidos percutidos que tienden a una dinámica entre pp y mp, asimismo se colaron ciertas resonancias de residuos de los ataques en secciones de poco volumen. La clase 3 consta de pequeños ataques en el registro agudo, también están presentes los glisandos irregulares, el juego entre tensión y distensión de las cuerdas, asimismo cuenta con algunas resonancias similares a las de la clase 2 con la diferencia de que tienen una mayor amplitud y corresponden al registro grave. Los resultados sonoros pueden escucharse en el siguiente enlace.[27] Siguiendo estas conclusiones se optó por generar más clases de acuerdo a las mezclas detectadas en cada una de ellas, por ejemplo: la clase 2 que tiene los residuos resonantes de los ataques y los sonidos percutidos con amplitud baja podrían dividirse en dos clases y siguiendo con ese razonamiento la clase 3 podría dividirse en 4 clases, lo que generó una nueva clasificación basada en 8 clases distintas.

[27] https://archive.org/details/sink_into_return-Clare_Cooper_4CLASES

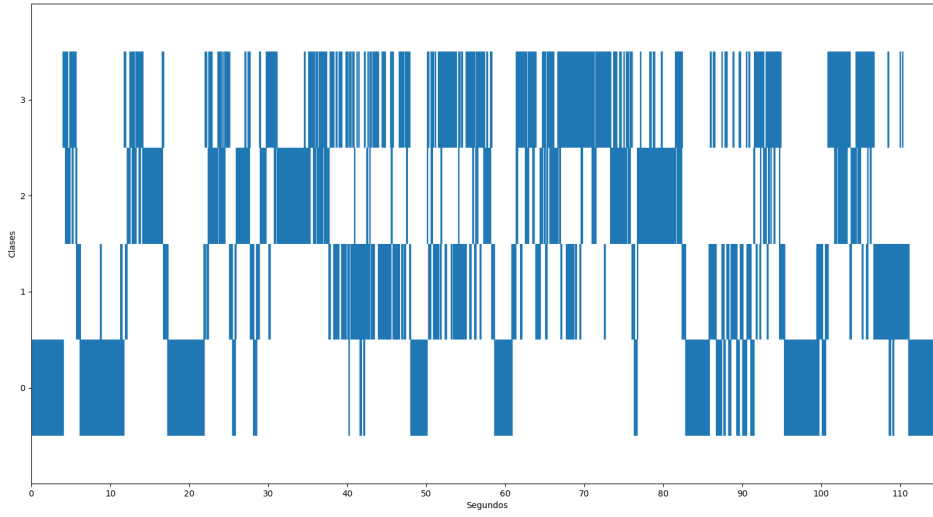

Figura 4.20: Estructura temporal en 4 clases

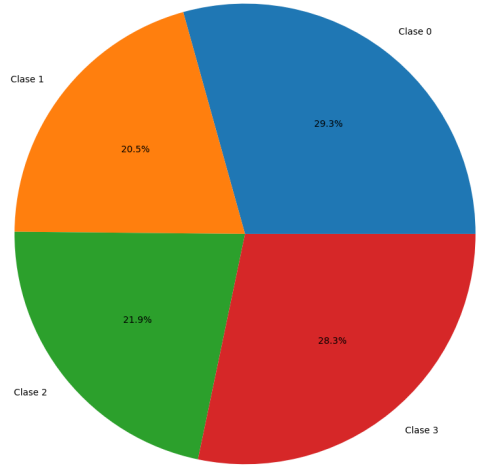

Figura 4.21: Pastel de 4 clases

A continuación se muestra nuevamente la forma de onda de *Sink into Return* basada en 8 clases que generaron 8 segmentos de segmentos creados por el sistema:

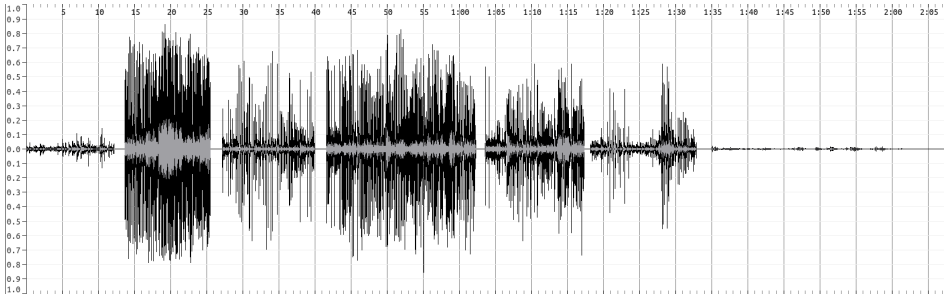

Figura 4.22: Forma de onda de las 8 clases detectadas

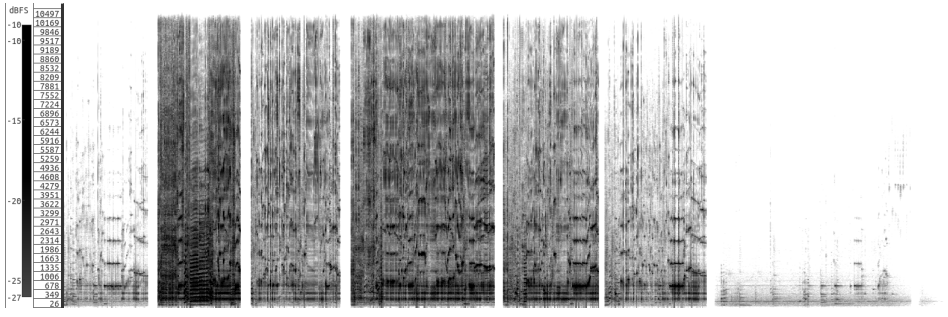

Figura 4.23: Espectrograma dividido en 8 clases

Como puede apreciarse en el espectrograma y en la forma de onda anteriores, las decisiones tomadas por el clasificador parecieran adecuadas debido a las claras diferencias visuales entre cada uno, sin embargo para mi escucha no es así en todos los casos. En esta clasificación se encontró que la clase 0 contiene las resonancias de los ataques,

en la clase 1 quedó el sonido que recuerda a una tarjeta de plástico que golpea las cuerdas, en la clase 2 están los sonidos percutidos en registro agudo, en la clase 3 el juego entre tensión y distensión en las cuerdas del registro agudo y algunas resonancias de la clase 1, en la clase 4, algunas resonancias y sonidos percutidos continuamente en el registro agudo, en la clase 5 quedaron algunas resonancias y sonidos percutidos continuamente en el registro agudo, en la clase 6, silencios y sonidos con sonoridades mínimas y en la clase 7 silencios.

Deribado de esto, considero que no fue muy clara la clasificación en esta prueba, ya que se han encontrado varios sonidos revueltos en distintas clases provocando poca coherencia en los resultados obtenidos. Una hipótesis de porqué ocurrieron estas discrepancias fue que se segmentó demasiado el archivo original de audio provocando una dislocación entre los sonidos percutidos y su resonancia en el espacio, generando una clase que ni siquiera había sido prevista en un principio. Además de que estos residuos de los ataques tendieron a generar toda una gama de clases, guardan cierto parecido con los ataques iniciales que están en una dinámica más alta, es decir la resonancia de un ataque *fortissimo* fue clasificada con un ataque *mezzo-forte*.

Estos resultados fueron muy útiles para entender un poco más las lógicas seguidas por el programa, además considero importante que dicha aproximación basada en la segmentación con alta resolución, produjo en un principio resultados insospechados que solamente pudieron ser evidentes tras realizar las pruebas. A este respecto surge la necesidad de integrar esas resonancias a donde pertenecen, los ataques. Podría ser una discusión interesante decidir que tipo de metodología sería la adecuada para integrarlas, una hipótesis sería agregar un sistema de visualización espectrográfica que pueda incluso segmentar

partiendo del contorno de la resonancia de un ataque, aunque esto implica un trabajo que no será desarrollado en la tesis por cuestiones de tiempo y complejidad técnica. Lo que se propone, contradiciendo el planteamiento inicial de segmentar en la mayor cantidad de fragmentos posibles cada una de las improvisaciones, es buscar un mayor equilibrio entre los cortes, de manera que pudieran respetar los inicios y finales de cada sonido.

Siguiendo con lo anterior se buscó a través de modificaciones en la configuración del segmentador conservar los ataques seguidos por su resonancia y colocar los silencios en otros segmentos, esto con el fin de generar una clasificación más acotada y especifica sobre cada clase. Ciertamente hay un cambio sustancial en esta clasificación debido a las modificaciones y ajustes que se hicieron a los parametros del objeto de *Librosa peak pick*, ya que este determina varias condiciones que tienen que ser cumplidas para encontrar los onsets. En las condiciones a cumplir se debe definir el:

número de muestras antes de n (siendo n el *onset*) sobre el cual se calcula el máximo, número de muestras después de n sobre el cual se calcula el máximo, número de muestras antes de n con respecto a qué media se calcula, número de muestras después de n sobre qué media se calcula, desplazamiento del umbral para la media y número de muestras para esperar después de recoger un pico.

Esta fue la configuración modificada: peaks = librosa.util.peak_pick (onset_env, 3, 10, 7, 7, 0.20, 0.01).

Con esta nueva segmentación se detectaron y crearon 321 muestras, las cuales fueron agrupadas en 4 y 8 clases, los resultados son mucho más coherentes dado que los momentos que son timbricamente similares quedaron en un solo archivo de audio y además fueron

clasificados por su densidad sonora de manera que la clase 0 es la menos densa, la clase 1 es poco densa, la clase 2 es densa y la clase 3 es mucho más densa sonoramente hablando. Sin embargo, en la clasificación donde fueron depositados los sonidos casi silentes y los silencios aún hay algunas resonancias y artefactos sonoros que de momento no fue posible aislar plenamente debido a que no hay ningún sonido que sea de utilidad para segmentar el último momento de un ataque y su continuidad con el silencio. Asimismo, es posible apreciar una mayor claridad y ordenamiento en la estructura temporal por clases de la figura 4.24 que en la versión anterior de la figura 4.20, ya que en esta última incluso pueden ser visualizados los artefactos y sonidos de resonancias mostrados como pequeñas rayas delgadas que acompañan de forma más compleja dicha visualización. En el siguiente enlace pueden ser escuchados los resultados sonoros de la clasificación.[28]

Para continuar, el siguiente experimento se realizó con un solo de guitarra eléctrica realizado por Tetuzi Akiyama, esta es una improvisación libre caracterizada por tener momentos sonoros muy sutiles y momentos de mucha agresividad sonora contrastados con silencios prolongados dispersos a lo largo de toda la interpretación. Debido a que esta improvisación es mucho más compleja que la anterior por a su carácter etéreo y abstracto, desde un principio se esperaba que no fuera muy clara en su segmentación, aún así resulta de sumo interés analizarla para identificar los fuertes y las debilidades de esta aproximación para detectar momentos en las improvisaciones libres.

[28] https://archive.org/details/sink_into_return-Clare_Cooper_321_cortes_8_clases https://archive.org/details/sink_into_return-Clare_Cooper_4_clases_321_segmentos

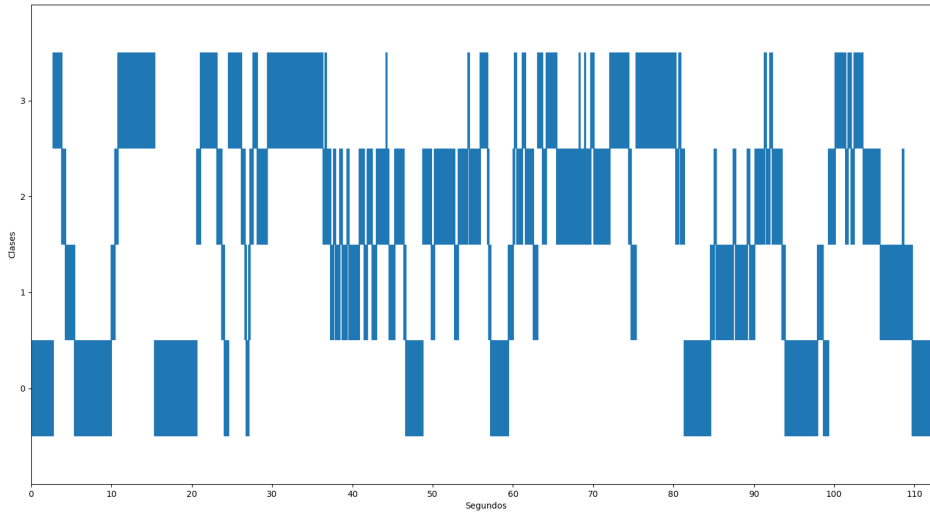

Figura 4.24: Estructura temporal dividida en 4 clases con nueva segmentación

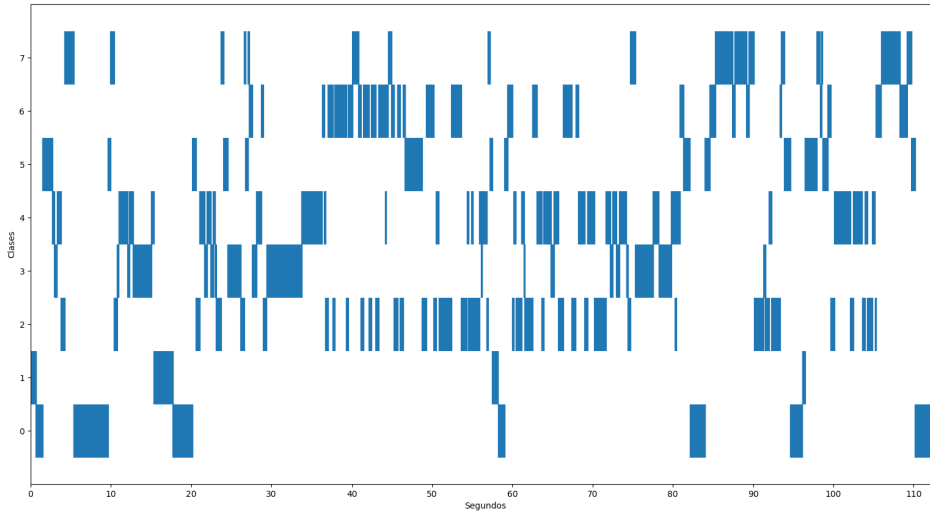

Figura 4.25: Estructura temporal dividida en 8 clases con nueva segmentación

Al realizar varias pruebas con diferentes combinaciones de descriptores, al igual que en las pruebas anteriores, se encontró que trabajar con MFCCs produce los mejores resultados al momento de clasificar, además después de haber probado con 3 hasta 9 clases diferentes, encontré que fue más coherente designar 3 clases ya que así fue posible tener una mejor clasificación entre los diferentes materiales encontrados en los 335 segmentos generados por el programa. La clase cero contiene la mayor parte de los momentos de inacción caracterizada por tender hacia la producción de sonidos mínimos y ruido de fondo (hum) de guitarra eléctrica. En la clase uno es posible encontrar los momentos más densamente sonoros: feedbacks, ataques secos y resonantes *fortissimos* tanto en registro agudo como en registro grave y ataques generados con las pastillas. En la clase dos en general hay acciones espaciadas continuadas por sus resonancias que tienden a un volumen más bajo que los sonidos de la clase uno. En las figuras 4.24 y 4.25 puede observar la forma en la que Akiyama se aproxima en esta improvisación usando la mayor parte del tiempo el elemento de la inactividad contrastándolo con momentos más o menos densos, abruptos y agresivos que rompen con el estatismo acompañado del ruido de fondo de la guitarra. En el siguiente enlace pueden ser escuchadas las clases generadas.

El siguiente experimento fue realizado con una pieza/improvisación (no queda muy claro) de Fernando Vigueras *Coral Continuo*. La improvisación esta generada por un continuo general que va transformándose paulatinamente en otros continuos resaltando distintas cualidades texturales, tímbricas y armónicas. Resulta importante analizar esta improvisación ya que es muy distinta a las 2 anteriores debido a que no cuenta con una construcción ensamblada por la interacción con la

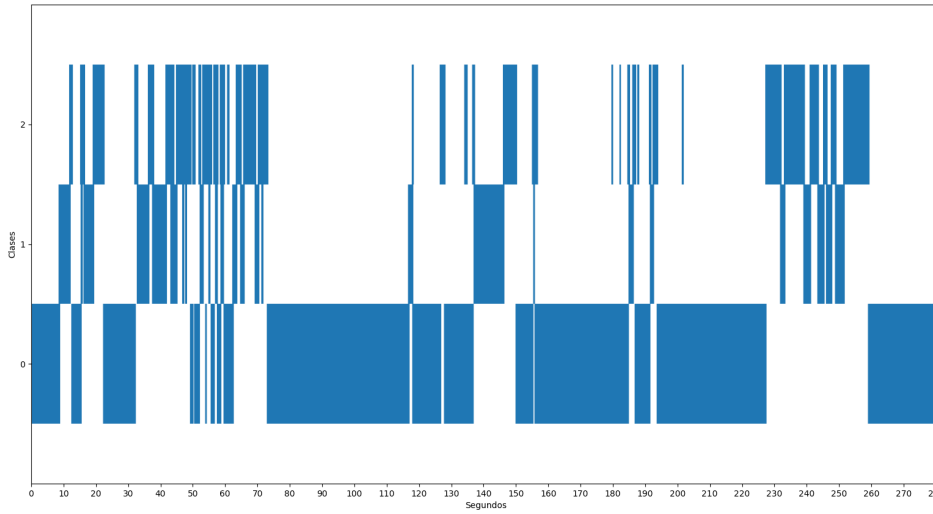

Figura 4.26: Estructura temporal del solo de Akiyama dividida en 3 clases

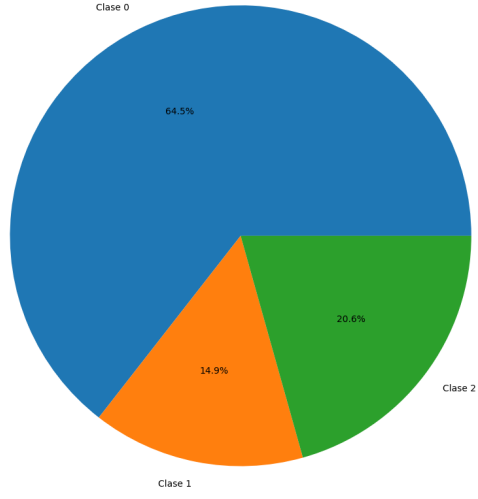

Figura 4.27: Pastel de 3 clases del solo de Akiyama

inacción, el silencio y las resonancias del espacio, sino más bien es una masa sonora ininterrumpida de principio a fin.

Al clasificar en 3 tipos de materiales los diferentes archivos generados por el segmentador estos fueron los resultados obtenidos: en la clase uno están los momentos de la improvisación de menor densidad sonora, tremolandos en una dinámica muy baja. En la clase dos el material principal es el mismo con el que empieza la clase uno además se superpone una construcción rítmica de graves al fondo, sumado a esto progresivamente se van agregando nuevas capas armónicas y al final se suma un tremolando más limpio que recuerda al sonido de una guitarra acústica sin efectos siendo frotada por un arco. En la tercer clase hay una fuerte concentración energética que tiende a la retroalimentación constante de todos los materiales anteriores, además se agregan a todo esto algunos sonidos percutidos en el registro agudo que del mismo modo se repiten con intervalos periódicos de tiempo, finalmente se encuentra la melodía principal en su momento más álgido y saturado debido a las superposiciones de las repeticiones que, por momentos, tienden a subir sutilmente de altura. En la siguiente representación puede observarse más de la mitad del tiempo la improvisación tiende a concentrarse en la clase dos, osea el momento de mayor saturación sonora, despues la clase uno y finalmente la cero con que corresponden a las regiones medianamente densas y las secciones de menor densidad sonora respectivamente.

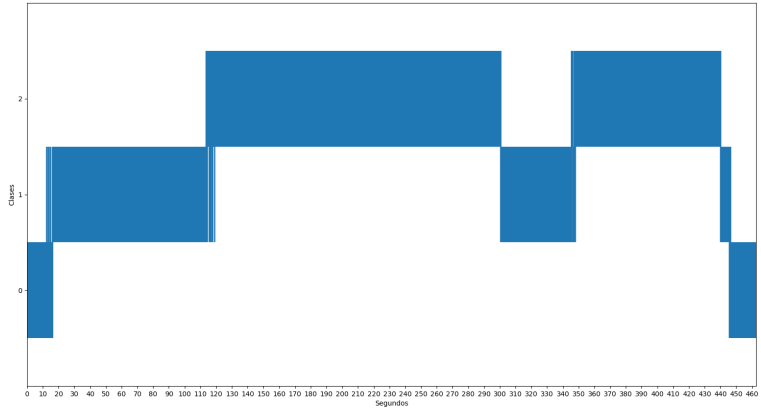

Figura 4.28: Estructura temporal en 3 clases de Coral Continuo de Fernando Vigueras

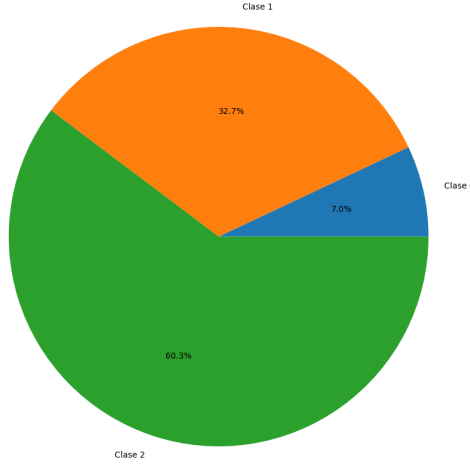

Figura 4.29: Pastel de Coral Continuo de Fernando Vigueras

En la estructura temporal de esta improvisación es posible ver un modelo de construcción arquetípico en la improvisación libre, es-

te modelo podría ser pensado como una forma AbBbA (aunque de acuerdo con la estructura temporal esta improvisación tiene una forma AbBbBbA), pensando en que comienza a construir el discurso sonoro desde el silencio y después con poca actividad, paulatinamente evoluciona en densidad hasta llegar al punto más álgido que en este caso es el de mayor duración como puede ser observado en el gráfico de pastel (acción que también es bastante recurrente en la improvisación). Después de un tiempo considerable sigue esta misma estrategia pero ahora invertida, desencadenando nuevamente en actividad mínima hasta terminar con el silencio.

Un comportamiento interesante y que aparentemente es un resultado emergente de las variables de entrada y las propiedades de los descriptores y las configuraciones de K-Means elegidas, es que pareciera que esta considerando las diferencias entre el promedio dinámico, densidad y timbre de cada una de las secciones de esta improvisación. Además un comportamiento extraño fue que los cambios en las dimensiones de los vectores, por lo menos en el análisis con K-Means, no resultaron en cambios muy sustanciales o predecibles. De hecho, los resultados con 12 vectores MFCCs o 60 o más, eran exactamente los mismos, habría que probar con otros clasificadores para corroborar si este comportamiento permanece constante.

Una vez realizados los experimentos anteriores con un solo archivo de audio a la vez, se propuso hacer una prueba final con un corpus de improvisaciones de 1 hora y 25 minutos de duración. Entre los autores de este corpus se encuentran improvisaciones de Derek Bailey, Remi Alvarez, Juan Pablo Villa, Okkyung Lee, Wilfrido Terrazas, Eli Kesler, Nicolas Collins, Yan Leguay, Mike Majkowski, Maja Ratkje, Tetuzi Akiyama, Clare Cooper, Fernando Vigueras y Davey Williams.

En total se obtuvieron 15957 fragmentos cortados con el detector de onset, los cuales fueron dispuestos para ser analizados y agrupados con la misma metodología pero con algunas modificaciones.

Primero se normalizaron todos los archivos de audio de acuerdo con la máxima amplitud, para que estuvieran en igualdad de condiciones debido a que algunas de las improvisaciones no estaban normalizadas, además se exportaron en un solo archivo. Después se propuso organizar toda la información en 15 clases respondiendo a los diferentes instrumentos de las improvisaciones que son 14 y una más por los momentos de silencio presentes en las improvisaciones. A pesar de que se probó con diferentes combinaciones de descriptores y distintas configuraciones en las variables de estado del clasificador, los resultados fueron que la clasificación con K-Means tuvo un desempeño medianamente aceptable ya que se obtuvieron mezclas tímbricas y de densidad muy disimiles en algunos casos, aunque otros de los archivos generados sí comparten ciertas características. Si a través de mi escucha pudiera indicar un índice de certeza en la clasificación del corpus sacando un promedio de los aciertos y fallos de la metodología, sería muy similar al obtenido en el análisis con Weka, entre 40 y 50 %, lo que demuestra que el análisis "realizado por la máquina" a pesar de no haber sido escuchado se acerca a lo que yo percibo como similar en cuanto a densidad sonora y similaridades tímbricas.

Además, se detectaron micro-fragmentos sonoros al inicio y al final de algunos archivos de audio generados que no tienen ninguna relación con el audio en general, como una constante emergente que se produjo de la interrelación de todos los elementos de esta metodología. Por otro lado se encuentra cierta consistencia en algunos fragmentos de los audios generados pero que del mismo modo no con-

servan ninguna relación con la clase que se supone más general. Por ejemplo, en la clase 3 es posible apreciar una coherencia en cuanto a densidad de materiales y actividad sonora se refiere, desde el principio maneja una altísima densidad sonora que alcanza Okkyung Lee con su cello, después es complementada con una densidad muy similar interpretada por Remy Alvarez con el Sax, pero, de pronto aparecen materiales como el Feedback de Akiyama que después es continuado con el de Derek Bailey, en seguida otro momento de mucha acción sonora con la percusión frenética de Eli Kesler, después viene un grito de Juan Pablo Villa que contrasta con la actividad anterior, aunque inmediatamente después la sonoridad vuelve a ser muy similar con los momentos altamente densos de esa misma improvisación, después el algoritmo agrega la voz de Maja Ratkje la cual contiene una densidad sonora congruente con los materiales anteriores, y finalmente un fragmento de la improvisación de Yann Leguay con la tornamesa intervenida que comparte las características generales de esta clase; alta actividad sonora y mucha agresividad. En el siguiente enlace pueden ser consultados los resultados de las improvisaciones. [29]

Otro ejemplo interesante es la clase 4 basado en una idea general de pulsaciones y ataques fluidos, rápidos sobre una construcción tímbrica similar, e incluso una altura constante con un efecto oscilante, hay mucha relación en varios de los gestos de esta clase. Además hacia la mitad de este archivo de audio, genera un bloque de sonidos que son semejantes entre sí pero que no corresponden al gesto común que es la pulsación constante.

Es interesante preguntarse ¿porqué tomó estas decisiones el algoritmo de clasificación? ¿qué encuentra similar en estos elementos?

[29] https://archive.org/details/corpusimprovisacioneslibres.

¿qué o quién le sugirió dicho ordenamiento? Es importante dejar en claro que este sistema sigue un proceso determinista, ya que no incluye procesos aleatorios, sino más bien opera desde variables de entrada, configuraciones y operaciones estadísticas bien definidas, además de que los resultados pueden ser exactamente replicables por cualquiera si las condiciones iniciales y los parámetros en los descriptores y algoritmos de clasificación son los mismos. En esta aproximación incluso los resultados audibles están definidos por dichas propiedades de entrada. La cuestión es entender que la relación entre la información de entrada y los resultados audibles no son caóticos, no lineales o impredecibles, sino más bien, difíciles de comprender, relacionar o anticipar, debido a la enorme base de datos que requiere analizarse y por la complicada forma en que se interrelacionan sus elementos. En este sentido el ordenamiento propuesto por el algoritmo de clasificación está determinado por todas las decisiones tomadas para construir el algoritmo del sistema de la máquina que escucha, las variables de entrada, sus propiedades intrínsecas, las configuraciones empleadas y el filtro de mi escucha, el cual calibró al sistema para producir dichos resultados. El hecho de que los resultados no sean del todo claros o evidentes en algunos momentos, responde a la gran cantidad de información que es procesada por todas las etapas encadenadas que involucran: la segmentación de los audios, su descripción numérica, la clasificación mediante algoritmos de aprendizaje de máquinas, la generación de los modelos visuales y la representación sonora de las clases identificadas por el algoritmo clasificador.

4.3. Conclusiones sobre el sistema en su estado actual

En este capítulo se pusieron a discusión algunas posturas de diferentes autores respecto a la libre improvisación para comprender a qué hace referencia la palabra libre y de este modo intentar saber qué características deberían ser consideradas al momento de generar una máquina que escucha improvisaciones libres. Se llegó a comprender que esa libertad está mediada por procesos de organización autogestiva y descentralizada, es decir, sin un centro que dicte lo que se va a hacer ni un rol muy específico de interacción que determine las funciones de cada integrante dentro de un ensamble como generalmente sucede en otras músicas. Ya en la práctica fueron detectadas algunas tendencias estéticas y estilísticas que la caracterizan y que sirvieron de guía para conceptualizar varias formas en que esta música puede analizarse partiendo de su estructura, densidades sonoras, tímbrica y formas de interacción. Posteriormente se generaron dos metodologías para analizar varias improvisaciones libres solistas, la primera partiendo de la aproximación del aprendizaje supervisado, del cual es posible concluir que es mucho más acertada que el aprendizaje sin supervisión basado en K-Means, el cual abarcó la segunda metodología. Además de que se hicieron evidentes las características emergentes de sistemas basados en aprendizaje automático a partir de las modificaciones en las variables de entrada y las propiedades tanto de los descriptores de audio como de los algoritmos de clasificación, ejemplo de ello puede ser escuchado en en la segunda metodología. Además, se encontró que realizar el análisis de clasificación con el perceptor multicapa y una configuración muy específica en los descriptores de audio resulta

mucho más eficiente, obteniendo un porcentaje de certeza de 91.75 %, respecto a otros descriptores y otros algoritmos de clasificación como K-Means y DL4MlpClassifier (aprendizaje profundo). A partir de esta aproximación se generó un modelo que puede ser utilizado para futuros desarrollos; asimismo, con el fin de corroborar la información expuesta por Weka se programó un segundo sistema (usando Python y las librerías Tensorflow y Librosa) de comprobación de resultados sonoros basado en el clasificador K-Means que permitió la escucha de las clasificaciones generadas por el sistema.

Los resultados obtenidos en pequeña escala, con un solo archivo, fueron bastante aceptables de acuerdo con mi escucha ya que se logra una clasificación basada en timbres amplitudes y densidad sonora generada de forma aparentemente emergente, aunque a gran escala (con el gran corpus de improvisaciones) los resultados fueron muy similares a los generados por Weka al usar el clasificador K-Means, obteniendo un índice de certeza en la clasificación de 59.53 %.[30] Una posible respuesta de por qué funciona analizar audios individuales pero no analizar el corpus más grande con K-Means es que una cantidad muy diversa de información es capaz de sobrepasar las capacidades de segmentación y clasificación que el algoritmo puede generar, considero que en este caso y como ha sido posible constatar en la clasificación con Weka y el algoritmo del perceptor multicapa, entre más complejo sea el problema a resolver, más compleja tendría que ser la herramienta a utilizar.

[30]Esto es más o menos una suposición debido a que por los tiempos de la maestría no fue posible sistematizar con un grupo de control estos resultados, de manera que varias personas escucharan y evaluaran con un porcentaje la capacidad del sistema para clasificar.

Asimismo es posible entablar en este momento una analogía entre los elementos que compartirían la clasificación con K-Means o el "aprendizaje no supervisado" y la libre improvisación, ya que ambos pueden proveer estrategias que denotan características emergentes, no predecibles por ninguno de los integrantes y que también surgen de manera espontánea por la interacción de todos sus componentes. Por ejemplo, al analizar el gran corpus de improvisaciones fue más evidente este resultado debido a que podría decirse que el sistema por sí mismo generó una suerte de composiciones que llegan a tener cierto sentido de orden estético.

Finalmente, me gustaría establecer algunos puntos comparativos entre la libre improvisación y el aprendizaje automático. A ambos los definen ciertos elementos constitutivos (que han sido descritos a lo largo del trabajo); sin embargo muchos otros quedan abiertos: como la forma de la improvisación, las interacciones que ocurrirán entre los músicos y el momento en que concluirá la improvisación, por mencionar algunos. En el caso del aprendizaje sin supervisión queda abierto el cómo se toman las decisiones de segmentación de clases y qué sonidos serán destinados a cada clase. En el caso de la libre improvisación, es en estos procesos de indefinición donde se encuentra el potencial para que surjan resultados insospechados que no pueden ser explicados simplemente como procesos causales (es decir que no necesariamente a cierta acción o inacción sonora le corresponde una determinada acción o consecuencia por parte de otro improvisador, esto debido a la naturaleza propia de la improvisación). En contraste, pese a que la máquina llega a generar resultados que parecieran insospechados, espontáneos o difíciles de predecir para la percepción humana, están fríamente calculados: no hay lugar para la incertidumbre y la impre-

decibilidad. Lo que parece suceder accidentalmente, está al margen de un fin perseguido por la máquina, con lo cual la máquina no tiene ninguna deriva contemplada. Lo que se puede afirmar es que la máquina está tomando "ciertas decisiones" al empalmar los archivos de audio proveídos, como agrupar aquellos que tienen una densidad sonora muy alta en una sola clase y los que tienen densidad sonora baja en otra; en este sentido la máquina logra su propia coherencia, que en algunos casos resultó bastante interesante escuchar. Queda abierta la discusión de si la máquina puede ser capaz de "percibir" combinaciones de sonidos y silencios que no son apreciables por la escucha humana, y que debido a esto haya llegado a resultados como los mostrados en la segunda metodología. Si la máquina puede escuchar más que nosotros, entonces tal vez hubo un sesgo en la forma en que me aproximé a la clasificación de la libre improvisación; por consiguiente pareciera que la máquina, debido a sus cualidades constitutivas, está escuchando diferente a nosotros. Lo que quedaría entonces por investigar es acercar la escucha de la máquina a una escucha más humana a través de procesos relacionales mucho más complejos que den cuenta de una escucha subjetiva.

Como se ha mencionado, este sistema se caracterizó por analizar una base de datos para generar modelos de escucha que describen las formas usuales o arquetípicas de aproximarse a la libre improvisación. Una utilidad que encontré al generar estos modelos basados en la aproximación de algunos improvisadores a la densidad sonora, la amplitud y el timbre, fue entender cómo estos manejan sus materiales de improvisación generando una estructura temporal que se hace evidente con el sistema de clasificación. Esta estructura queda accesible a través de las herramientas producidas en esta investiga-

ción para su uso y estudio en cualquier otro momento; estructura que puede ser modificada, transformada o visualizada como una partitura para ser incluso interpretada por otro músico atendiendo a los mismos parámetros o asignando otros materiales.

4.3.1. Propuestas para continuar trabajando en el sistema

Otros componentes importantes a analizar que quedaron pendientes en esta investigación son los roles de interacción en la improvisación con más de un solo músico (ya sea escucha, imitación, proposición (de una nueva idea musical), acompañamiento, ruptura y solo) aunque estos solo podrían ser analizados partiendo de grabaciones multicanal en donde cada improvisador genere su propio archivo para ser analizado posteriormente en su conjunto. Asimismo, sigo explorando de qué forma realizar una implementación de redes neuronales de aprendizaje profundo al análisis de improvisaciones libres para obtener un resultado sonoro audible que sea más útil que solamente la obtención de un índice de certeza como fue el caso del análisis con Weka.

Considero sumamente necesario realizar dos mesas de discusión con distintos improvisadores que analicen las controversias, criterios e implicaciones que cada uno tiene respecto a la forma de aproximarse a una unidad gestual en la improvisación libre y una segunda mesa que que analice la propuesta de clasificación generada por el sistema automático de escucha de improvisaciones libres, estas tendrían el objetivo de retomar e integrar las ideas generadas en estas discusiones para seguir desarrollando el sistema de clasificación basado no solamente en mi percepción sino también en la de otros.

Finalmente vendría la programación reactiva del sistema, aquí las formas de reaccionar de este estarían necesariamente mediadas por el análisis de lo que escucha, esto implica escuchar en tiempo real a los músicos y además las acciones sonoras que el sistema mismo produce; en este sentido si la máquina detecta la suma de varios de los componentes sonoros de una improvisación, respondería a través de diferentes técnicas de síntesis de audio acoplando sus acciones sonoras en tiempo real respecto a lo que escucha. Esta parte resulta una tarea bastante ardua, ya que implica la composición de estos sonidos y su respectiva programación o generación de reglas para ser activados, de manera que la máquina pueda fijar un rol de interacción dependiendo del contexto que se genere con los otros improvisadores.

Para cerrar, el sistema sonoro interactivo propuesto contendría tres subsistemas principales:

1) Un subsistema de escucha automática enfocado en el reconocimiento del timbre, la densidad y la amplitud del sonido. Se agregará al subsistema la posibilidad de grabar en tiempo-real lo que hacen las improvisadores que interactúen con él, además de analizar e integrar a su base de datos la nueva información almacenada para seguir ampliando su corpus sonoro. De esta forma, el sistema estaría en constante actualización y su adaptación a diferentes contextos sonoros-interactivos sería cada vez mayor.

2) Un subsistema de procesamiento de información que contendría los modelos generados por el sistema de clasificación basados en momentos sonoros de la libre improvisación. Para la generación de modelos interconectados más complejos se propone la utilización de redes de aprendizaje profundo. Los materiales sonoros a analizar serían: la base de datos de modelos de improvisación generada en esta investi-

gación, además de la ampliación de esa misma base de datos (tomada de grabaciones propias y de otros improvisadores libres). También se propone integrar grabaciones de paisajes sonoros propios o de otros autores.

3) Un subsistema de reacción, encargado de interpretar los modelos descritos por el subsistema anterior así como de activar las respuestas que el sistema entablará en diferentes contextos humanos. Este último subsistema contiene a su vez seis sub-subsistemas que son cada uno de los estados posibles de interacción propuestos: escucha, imitación, proposición (de una nueva idea musical), acompañamiento, ruptura y solo. A través de estos estados el sistema, idealmente, sería capaz de retar/cuestionar/acompañar/proponer/imponer(se) sonoramente frente a un humano, y, generar una dinámica entre los diferentes modos de interacción posibles.

Una propuesta metodológica de construcción de los sub-subsistemas es la siguiente: para el estado de imitación se propone que el sistema tenga la capacidad de analizar en tiempo real el timbre, la densidad y la amplitud sonora de lo que un humano está tocando. Después se haría un análisis comparativo mediante pesos (weights) que responden a los índices de similaridad tímbrica, de densidad y amplitud entre lo detectado y sus modelos en la base de datos. Si el sistema detecta el sonido con cierto índice de precisión mediante una escala asignada, se activará alguno de los modos de interacción propuestos. Por ejemplo, si el sistema obtuvo un índice de detección del noventa por ciento podría activar el estado de imitación, alterando sonoramente materiales que fueron grabados momentos antes de su acción. Otro ejemplo: si el sistema tuvo un índice de detección de cincuenta por ciento, podría activar el modo de proposición, seleccionando un tim-

bre, amplitud y densidad sonora similar a lo que escuchó, tomando como referencia los distintos archivos de audio de su base de datos. Último ejemplo: el modo de interacción de ruptura o solista, podría ser activado si el índice de detección llega apenas a 30 por ciento de certeza en la identificación de lo escuchado. En esos dos modos se propone que el sistema mediante el uso invertido de GANs (redes adversariales generativas) pueda responder a las nuevas acciones. Es decir, en vez de intentar construir un modelo sonoro descriptivo, intentaría a través de un modelo específico reconstruir otro(s) modelos sonoros. Ejemplos de ello se han realizado ampliamente en los últimos años en las artes visuales pero aún no con sonido. Una amplia lista de artículos e implementaciones técnicas se pueden encontrar en el siguiente repositorio de github.[31]

[31] https://github.com/nightrome/really-awesome-gan

Conclusiones

La investigación consistió en programar una serie de herramientas capaces de generar modelos descriptivos que dieron cuenta de las formas en que algunos improvisadores libres nos acercamos a dicha práctica, esto desde una aproximación basada en la escucha y el aprendizaje automático, que devino en resultados de análisis dirigidos hacia la densidad, la tímbrica y la amplitud sonora. Cabe señalar, que la metodología planteada en la subsección 4.2.9 presentó ciertas limitantes aunque también ciertas virtudes, difiriendo de otras metodologías al incluir un fuerte carácter empírico, basado principalmente en comprobar la clasificación de materiales por parte del sistema a partir de mi escucha atenta. Además, se describió de manera más técnica cómo es que funcionan algunas de las aproximaciones que hicieron posible la generación de estos modelos descriptivos de improvisaciones libres. Paralelamente se generó un recuento de proyectos similares destacando las aproximaciones metodológicas que fueron útiles para conocer el estado del arte en términos de la experimentación creativa dentro de sistemas interactivos aplicados al modelado de formas para improvisar. Asimismo se propuso pensar estas máquinas como instrumentos improductivos e ineficientes, capaces de producir en los ámbitos estéticos desviaciones que evoquen experiencias diferentes en contra

de la clásica idea prometéica del progreso que acompaña la eficiencia tecno-científica. Esta aproximación para generar modelos descriptivos de improvisaciones libres, además de las discusiones y menciones sobre proyectos similares, dio cuenta de cómo las herramientas del aprendizaje y la escucha de máquinas, más allá de sus fines y aplicaciones originarias, pueden ser encaminadas hacia la generación de proyectos artísticos y estéticos capaces de fomentar el surgimiento de otros imaginarios posibles. La tecnología no necesariamente debe tener un funcionamiento óptimo para tales fines, sino que es posible emplear estas precariedades, errores u omisiones para transformar la misma práctica artística así como las utilidades finales de estas herramientas, y, de este modo, ampliar el panorama de acción entre el cruce del arte, la tecnología y la ciencia.

Por otra parte, desde una mirada más crítica sobre el contexto de dichas herramientas, se discutieron las implicaciones que el aprendizaje y la escucha automática tienen en algunos campos como la recuperación de información musical, el análisis y la recopilación de datos destinados a la vigilancia, el control y la guerra. Se analizaron algunos escenarios actuales y se plantearon otros más catastróficos, llegando a la conclusión de que una alternativa es apropiarse y conocer a fondo la tecnología que nos domina. Adicionalmente se analizaron los imaginarios implicados detrás de todo ello, partiendo de la necesidad por la simulación del mundo a través de entrañables historias con autómatas que siguen resonando hasta nuestros días en los mecanismos capaces de simular funciones biológicas y humanas con el objetivo de tener el control sobre cada uno de los aspectos del mundo a través de los medios de reproducción sonora, visual, de nosotros mismos, nuestra subjetividad y la propia realidad.

Glosario de términos

4.4. TensorFLow

TensorFlow es una librería de programación de código abierto comúnmente usada en la investigación y la producción de Google para la creación de aplicaciones con aprendizaje de máquinas y redes neuronales de aprendizaje profundo. Fue desarrollada por el equipo de investigadores de Google Brain en la investigación de aplicaciones de inteligencia artificial. "Los nodos en el gráfico representan operaciones matemáticas, mientras que los bordes del gráfico representan los conjuntos de datos multidimensionales (tensores) comunicados entre ellos. La arquitectura flexible le permite implementar cálculos en una o más CPU o GPU (tarjeta de video) en una computadora de escritorio, servidor o dispositivo móvil con una sola API."[32]

4.5. Weka

Weka (Waikato Environment for Knowledge Analysis) es un un software libre escrito en java que incluye varias herramientas enfoca-

[32] https://www.tensorflow.org/

das en el aprendizaje de máquinas, fue desarrollado en la Universidad de Weikato en Nueva Zelanda desde 1993. A diferencia de otros proyectos de aprendizaje automático, el énfasis en WEKA está en proporcionar un entorno de trabajo para los especialistas en un tema especifico en lugar del experto en aprendizaje automático. WEKA incluye una gran cantidad de herramientas interactivas para la manipulación de datos, visualización de resultados, vinculación de bases de datos, validación cruzada y comparación de conjuntos de reglas, para complementar las herramientas básicas de aprendizaje automático.[33]

4.6. Wekinator

Wekinator es un software con licencia u código abierto desarrollado en 2009 por Rebecca Fiebrink, su enfoque principal es acercar a los artistas de formas prácticas e intuitivas a el aprendizaje de máquinas para la construcción de nuevos instrumentos, controladores e interfaces, sus aplicaciones pueden ser muy amplias. Wekinator a sido ampliamente utilizado por muchos artistas para crear obras basadas en visión computacional y sistemas de escucha automática.[34]

[33] https://www.cs.waikato.ac.nz/ml/weka/
[34] http://www.wekinator.org/

Anexos

4.7. Resultados del modelo generado por WEKA

```
=== Predictions on test set ===

    instance  actual  predicted error  prediction
        1    1:Cello 4:Feedback   +   1
        2    1:Cello 4:Feedback   +   0.751
        3    1:Cello 5:Guitarra_Electrica   +   0.998
        4    1:Cello 3:Tam+Objetos   +   0.567
        5    1:Cello    1:Cello       0.987
        6    1:Cello 3:Tam+Objetos   +   0.825
        7    1:Cello    1:Cello       0.999
        8    1:Cello    1:Cello       0.616
        9    1:Cello    1:Cello       0.997
       10    1:Cello     2:Arpa   +   0.687
       11    1:Cello     2:Arpa   +   0.999
       12    1:Cello 4:Feedback   +   0.974
       13     2:Arpa 6:Silencios   +   0.977
       14     2:Arpa 5:Guitarra_Electrica   +   0.559
```

```
15     2:Arpa 3:Tam+Objetos   +   0.975
16     2:Arpa 3:Tam+Objetos   +   0.524
17     2:Arpa 5:Guitarra_Electrica   +   0.664
18     2:Arpa 5:Guitarra_Electrica   +   0.843
19     2:Arpa 5:Guitarra_Electrica   +   0.588
20     2:Arpa 3:Tam+Objetos   +   0.91
21     2:Arpa 5:Guitarra_Electrica   +   0.902
22     2:Arpa 5:Guitarra_Electrica   +   0.955
23     2:Arpa 5:Guitarra_Electrica   +   0.924
24     2:Arpa 3:Tam+Objetos   +   0.526
25     2:Arpa    1:Cello   +   0.647
26     2:Arpa 5:Guitarra_Electrica   +   0.65
27     2:Arpa 5:Guitarra_Electrica   +   0.999
28     2:Arpa 5:Guitarra_Electrica   +   0.85
29     2:Arpa 5:Guitarra_Electrica   +   0.454
30     2:Arpa 5:Guitarra_Electrica   +   0.618
31     2:Arpa     2:Arpa       0.988
32     2:Arpa 5:Guitarra_Electrica   +   0.556
33     2:Arpa 5:Guitarra_Electrica   +   0.992
34 3:Tam+Objetos 6:Silencios   +   0.758
35 3:Tam+Objetos 4:Feedback   +   0.764
36 3:Tam+Objetos 6:Silencios   +   0.952
37 3:Tam+Objetos 5:Guitarra_Electrica   +   0.508
38 3:Tam+Objetos 5:Guitarra_Electrica   +   0.998
39 3:Tam+Objetos 4:Feedback   +   0.511
40 3:Tam+Objetos 3:Tam+Objetos       1
41 3:Tam+Objetos 5:Guitarra_Electrica   +   0.701
42 3:Tam+Objetos 6:Silencios   +   1
43 3:Tam+Objetos 6:Silencios   +   0.847
44 3:Tam+Objetos 3:Tam+Objetos       0.851
45 3:Tam+Objetos 6:Silencios   +   1
46 3:Tam+Objetos 3:Tam+Objetos       1
```

```
47 3:Tam+Objetos 6:Silencios   +   0.504
48 3:Tam+Objetos 5:Guitarra_Electrica   +   0.851
49 3:Tam+Objetos 5:Guitarra_Electrica   +   0.999
50 3:Tam+Objetos 5:Guitarra_Electrica   +   0.974
51 3:Tam+Objetos 4:Feedback   +   0.898
52 3:Tam+Objetos     2:Arpa   +   0.993
53 3:Tam+Objetos     2:Arpa   +   0.65
54 3:Tam+Objetos 4:Feedback   +   1
55 3:Tam+Objetos 4:Feedback   +   0.951
56 3:Tam+Objetos 3:Tam+Objetos       0.993
57 3:Tam+Objetos     2:Arpa   +   0.77
58 3:Tam+Objetos 4:Feedback   +   0.513
59 3:Tam+Objetos     2:Arpa   +   0.827
60 3:Tam+Objetos     2:Arpa   +   0.594
61 3:Tam+Objetos     2:Arpa   +   0.613
62 3:Tam+Objetos     2:Arpa   +   0.971
63 3:Tam+Objetos    1:Cello   +   0.502
64 3:Tam+Objetos     2:Arpa   +   0.979
65 3:Tam+Objetos 5:Guitarra_Electrica   +   0.99
66 4:Feedback 4:Feedback       1
67 4:Feedback 4:Feedback       1
68 4:Feedback 4:Feedback       1
69 4:Feedback 4:Feedback       1
70 4:Feedback 4:Feedback       0.819
71 4:Feedback 4:Feedback       0.999
72 4:Feedback 4:Feedback       0.981
73 4:Feedback 4:Feedback       0.99
74 4:Feedback 4:Feedback       1
75 4:Feedback 4:Feedback       1
76 5:Guitarra_Electrica 4:Feedback   +   0.989
77 5:Guitarra_Electrica 5:Guitarra_Electrica       0.51
78 5:Guitarra_Electrica     2:Arpa   +   0.611
```

```
 79 5:Guitarra_Electrica 5:Guitarra_Electrica       0.813
 80 5:Guitarra_Electrica     2:Arpa   +   0.976
 81 5:Guitarra_Electrica 4:Feedback   +   0.996
 82 5:Guitarra_Electrica 5:Guitarra_Electrica       0.989
 83 5:Guitarra_Electrica 5:Guitarra_Electrica       0.999
 84 5:Guitarra_Electrica 5:Guitarra_Electrica       0.992
 85 5:Guitarra_Electrica 5:Guitarra_Electrica       0.998
 86 5:Guitarra_Electrica 4:Feedback   +   0.973
 87 5:Guitarra_Electrica 5:Guitarra_Electrica       0.481
 88 5:Guitarra_Electrica     2:Arpa   +   0.804
 89 5:Guitarra_Electrica 4:Feedback   +   0.997
 90 5:Guitarra_Electrica 5:Guitarra_Electrica       0.986
 91 5:Guitarra_Electrica 4:Feedback   +   0.987
 92 5:Guitarra_Electrica 5:Guitarra_Electrica       0.773
 93 5:Guitarra_Electrica 4:Feedback   +   0.998
 94 5:Guitarra_Electrica 4:Feedback   +   0.829
 95 5:Guitarra_Electrica 4:Feedback   +   0.995
 96 5:Guitarra_Electrica 5:Guitarra_Electrica       0.996
 97 5:Guitarra_Electrica 4:Feedback   +   0.657
 98 5:Guitarra_Electrica 4:Feedback   +   0.842
 99 5:Guitarra_Electrica    1:Cello   +   1
100 5:Guitarra_Electrica 4:Feedback   +   0.999
101 5:Guitarra_Electrica 5:Guitarra_Electrica       0.866
102 5:Guitarra_Electrica 5:Guitarra_Electrica       0.953
103 5:Guitarra_Electrica 4:Feedback   +   0.983
104 5:Guitarra_Electrica     2:Arpa   +   0.889
105 5:Guitarra_Electrica 5:Guitarra_Electrica       0.986
106 5:Guitarra_Electrica    1:Cello   +   0.989
107 5:Guitarra_Electrica 5:Guitarra_Electrica       1
108 5:Guitarra_Electrica 3:Tam+Objetos   +   0.946
109 5:Guitarra_Electrica 5:Guitarra_Electrica       0.745
110 5:Guitarra_Electrica 4:Feedback   +   1
```

```
111 5:Guitarra_Electrica 5:Guitarra_Electrica       0.996
112 5:Guitarra_Electrica 5:Guitarra_Electrica       0.98
113 6:Silencios 6:Silencios       0.948
114 6:Silencios 6:Silencios       0.998
115 6:Silencios 6:Silencios       0.985
116 6:Silencios 6:Silencios       0.995
117 6:Silencios 6:Silencios       0.879
118 6:Silencios 6:Silencios       0.862
119 6:Silencios 6:Silencios       0.984
120 6:Silencios 6:Silencios       1
121 6:Silencios 6:Silencios       0.994
122 6:Silencios 6:Silencios       0.998
123 6:Silencios 6:Silencios       0.995
124 6:Silencios 6:Silencios       1
125 6:Silencios 6:Silencios       0.921
126 6:Silencios 6:Silencios       1
127 6:Silencios 6:Silencios       0.971
128 6:Silencios 6:Silencios       0.997
129 6:Silencios 6:Silencios       0.999
130 6:Silencios 4:Feedback   +   0.506
131 6:Silencios 6:Silencios       0.941
132 6:Silencios 6:Silencios       0.997
133 6:Silencios 6:Silencios       0.957
134 6:Silencios 6:Silencios       1
135 6:Silencios 6:Silencios       0.987
136 6:Silencios 6:Silencios       1
137 6:Silencios 6:Silencios       1
138 6:Silencios 6:Silencios       1
139 6:Silencios 6:Silencios       1
140 6:Silencios 6:Silencios       0.997
141 6:Silencios 6:Silencios       1
142 6:Silencios 6:Silencios       1
```

```
      143 6:Silencios 6:Silencios       1
      144 6:Silencios 6:Silencios       1
      145 6:Silencios 6:Silencios       1
      146 6:Silencios 6:Silencios       1
      147 6:Silencios 6:Silencios       0.998
      148 6:Silencios 6:Silencios       1
      149 6:Silencios 6:Silencios       0.965
      150 6:Silencios 6:Silencios       0.98

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances          73               48.6667 %
Incorrectly Classified Instances        77               51.3333 %
Kappa statistic                          0.3686
Mean absolute error                      0.1722
Root mean squared error                  0.3835
Relative absolute error                 64.14   %
Root relative squared error            104.607  %
Total Number of Instances              150
```

4.8. Reconocimiento al momento con Wekinator

Para tener una idea más cercana de como sería el reconocimiento de elementos sonoros de la libre improvisación en tiempo-real, decidí

efectuar aparte de los experimentos realizados en esta tesis, el mismo experimento de clasificación con Wekinator, programa de aprendizaje supervisado basado en WEKA desarrollado por Rebecca Fiebrink. Su característica principal es que está diseñado para usarse de forma muy intuitiva sin la necesidad de conocer a profundidad los temas del aprendizaje automático, además de que está optimizado para reconocer en tiempo real caudales de información tan diversa como sensores haya; Wekinator puede aprender de video, audio, sensores lumínicos, acelerómetros , etc. Otra de sus características es que funciona de forma muy artesanal, ya que requiere que las nuevas muestras sean proporcionadas y clasificadas una a una de forma meramente manual. Las clasificaciones son guardadas en archivos que le permiten a Wekinator identificar ciertos estados que posteriormente pueden ser reconocidos por el sistema. Para poder echar a andar los experimentos con Wekinator se tuvieron que realizar varias tareas manualmente en distintos programas o aplicaciones; generar un sistema de reconocimiento de audio (con los descriptores de audio MFCC y centroide espectral) en Supercollider. Reproducir un archivo de audio o registrarlo mediante un micrófono, mandar toda la información reconocida por Supercollider a Wekinator via OSC. Indicar a Wekinator manualmente a cuál de las posibles clases corresponde el momento sonoro mostrado. Entrenar a la máquina, lo cual consiste en aplicar un algoritmo de segmentación para diferenciar entre las diferentes clases disponibles,[35] Finalmente, activar el modelo y probarlo con nuevas

[35]El sistema cuenta con varios algoritmos como k-means, k-nearest neighbors, redes neuronales, máquinas de soporte vectorial (Suport Vector Machines), entre otros, que posibilitan hacer una segmentación en la información de manera que al entrar nuevos ejemplos sean de audio o cualquier otro tipo de dato, el sistema se encargue de clasificarlo y determinar a qué clase pertenece.

muestras sonoras. "[...] hay que enseñarle a la computadora cómo el sonido cambia a lo largo del tiempo a través de cuadros o ventanas de análisis que pueden ir desde muy pocas hasta miles de muestras por segundo, después se tienen que calcular las medidas o cambios significativos en las muestras de la ventana y repetir el proceso cuantas veces sea necesario alrededor del archivo de audio."[36] Hecho lo anterior, se programó un sistema de reacción en Supercollider de acuerdo con las clases identificadas por Wekinator. Por ejemplo, si lo que detectó fue un arpegio realizado con cuerdas de nylon, el sistema inmediatamente manda una señal OSC a Supercollider quien se encarga de recibir la información correspondiente para activar sintetizadores y efectos en el audio de entrada. Cada vez que aparezca un nuevo sonido el sistema intercambiará entre estos dos modos de reacción.

Esta exploración con Wekinator fue bastante útil para reconocer las limitantes que esta metodología tiene, ya que al igual que WEKA requieren de procesos que toman mucho tiempo debido a su enfoque meramente manual y a que resulta algo ineficaz por la cantidad de pasos a realizar y el uso de varias aplicaciones. Asimismo, los resultados obtenidos mediante esta aproximación fueron algo limitados ya que el sistema de reconocimiento resultó poco robusto para identificar timbres complejos generados en tiempo-real, y además, la función de mandar esa información, desde y hacia Supercollider para generar reacciones sonoras que se adecuaran a los sonidos y las interacciones producidas por el intérprete, no fue del todo convincente. Si bien, en algunos momentos se obtuvieron resultados interesantes a nivel de in-

[36] Rebecca Fiebrink, Machine Learning for Musicians and Artists, https://www.kadenze.com/courses/machine-learning-for-musicians-and-artists/info

teracción o transformación tímbrica —muchos de ellos fueron producto del azar, la contingencia o la confusión para clasificar e identificar audios por parte del sistema—, no fueron del todo satisfactorios a un nivel más amplio en un contexto de improvisación libre ya que las posibilidades de interacción no se desarrollaron mucho más en esta etapa debido a la poca claridad percibida desde el inicio en la identificación de los momentos sonoros.

A este respecto surge la siguiente pregunta: ¿El sistema tiene que estar completamente optimizado o terminado para poder ser útil en un contexto artístico? La respuesta a esta pregunta sería doble, ya que depende lo que se esté buscando. Sí, porque precisamente se está buscando hacer un sistema inteligente que pueda conceptualizar y modelar la forma en la que improvisamos los humanos, y no necesariamente si lo que se está buscando es crear un sistema caótico de interacciones. Pude haber seguido por este camino y llevar a sus últimas consecuencias un sistema de reacción a momentos sonoros sin que necesariamente los reconociera de forma certera, pero se estaría jugando con el terreno de la ilusión y lo ficticio a través de un sistema que parcialmente responde de manera adecuada en algunos momentos y en otros no, a causa de la coincidencia o el azar. Además, un sistema de estas características fue desarrollado antes en el proyecto de licenciatura (del cual hago mención en la introducción). Más bien, lo que se está buscando aquí es generar un sistema capaz de reconocer de forma precisa momentos de la improvisación y tener una idea más o menos clara de los arquetipos improvisatorios comunes en la práctica.

Bibliografía

Vincent Akkermans, Joan Serrà, and Perfecto Herrera. Shape-based spectral contrast descriptor. In *Sound and Music Computing Conference*, pages 143–148, Porto, Portugal., 25/07/2009 2009.

Chefa Alonso. *Improvisación libre la composición en movimiento.* 2007.

Ethem Alpaydin. *Machine Learning: The New AI.* The MIT Press Essential Knowledge series. MIT Press, 2016.

Francis Bacon. *Bacon's Advancement of Learning and the New Atlantis.* Lulu Enterprises Incorporated, 2010.

Derek Bailey. *Improvisation. Its Nature and Practice in Music.* Moorland, 1980.

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, Sept 2005.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

Mason Bretan and Gil Weinberg. A survey of robotic musicianship. *Commun. ACM*, 59(5):100–109, April 2016.

Fritjof Capra. *La trama de la vida: Una nueva perspectiva de los sistemas vivos*. Colección compactos. Editorial Anagrama S.A., 2009.

Nick Collins. SCMIR: A SuperCollider music information retrieval library. In *Proceedings of the International Computer Music Conference 2011*, pages 499–502, 2011.

Nick Collins. Towards machine musicians who have listened to more music than us : audio database-led algorithmic criticism for automatic composition and live concert systems. *Computers in entertainment.*, 14(3):2, December 2016.

Kate Crawford. *Following You: Disciplines of Listening in Social Media*. The Sound Studies Reader. Taylor & Francis, 2012.

Derek J. de Solla Price. Automata and the origins of mechanism and mechanistic philosophy. *Technology and Culture*, 5(1):9–23, 1964.

Paul Doornbusch. Computer sound synthesis in 1951: The music of csirac. *Computer Music Journal*, 28:10–25.

E. F. Evans. Auditory processing of complex sounds: An overview. *Philosophical Transactions: Biological Sciences*, 336(1278):295–306, 1992.

Ira J. Hirsh. Auditory perception of temporal order. *The Journal of the Acoustical Society of America*, 31(6):759–767, 1959.

Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature.

In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, 2002.

Marc Leman, Federico Avanzini, Alain de Cheveigné, and Emmanuel Bigand. The societal contexts for sound and music computing: Research, education, industry, and socio-culture. *Journal of New Music Research*, 36(3):149–167, 2007.

Alexander Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Wiley, 2012.

B. Levy. *OMax The Software Improviser*, 2004-2012.

Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

Wade Matthews. ¡escucha! claves para entender la libre improvisación. 2001. URL www.wadematthews.info/.

Wade Matthews. Y la libre improvisación, qué tiene de improvisada. 2002. URL www.wadematthews.info/.

Meinard Müller. *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 2007.

Doug Van Nort, Pauline Oliveros, and Jonas Braasch. Electro/acoustic improvisation and deeply listening machines. *Journal of New Music Research*, 42(4):303–324, 2013.

Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 01 2004.

Juan Martín Prada. *Prácticas artísticas e Internet en la época de la redes sociales.* AKAL, 2015.

Bruno H. Repp. Patterns of note onset asynchronies in expressive piano performance. *The Journal of the Acoustical Society of America*, 100(6):3917–3932, 1996.

Luigi Russolo. *The art of noise.* A Great Bear Pamphlet, 1913.

Pierre Schaeffer. *Tratado de los objetos musicales.* Alianza música, 1996.

Isaac Schankler, Jordan B. L. Smith, Alexandre R. J. François, and Elaine Chew. *Emergent Formal Structures of Factor Oracle-Driven Musical Improvisations*, pages 241–254. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-21590-2.

Eric D. Scheirer. *Music-Listening Systems.* PhD thesis, 2000.

Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard. Machine listening techniques as a complement to video image analysis in forensics. *IEEE International Conference on Image Processing*, 2016.

Nishant Shukla. *Machine Learning with TensorFlow.* Manning Publications, 2017.

Jonathan Sterne. *The Audible Past: Cultural Origins of Sound Reproduction.* Duke University Press, 2003.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Dan Stowell and Mark Plumbley. Adaptive whitening for improved real-time audio onset detection. *International Computer Music Conference (ICMC)*, 2007.

Emily Thompson. *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933*. ACLS Humanities E-Book. MIT Press, 2004.

David Toop. *Into the Maelstrom: Music, Improvisation and the Dream of Freedom Before 1970*. 2016.

Eguzki Urteaga. La teoría de sistemas de niklas luhmann. *Univerisdad del País Vasco*, 2009.

Jun Yang, Fa-Long Luo, and Arye Nehorai. Spectral contrast enhancement: Algorithms and comparisons. *Speech Communication*, 39 (1):33–46, January 2003.