ISSN 1688-2806

Universidad de la República
Facultad de Ingeniería

Alineación entre audio y partitura para obras del repertorio de la flauta traversa

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Juan P. Braga Brum

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Eléctrica.

Director de Tesis
DSc. Luiz W. P. Biscainho ..... Univ. Federal de Rio de Janeiro

Tribunal
Dr. Ing. Pablo Cancela ..... Universidad de la República
PhD. Ignacio Ramírez ..... Universidad de la República
DSc. Leonardo Nunez (Revisor Externo) ... Microsoft Lab Research
DMus. Osvaldo Budón ..... Universidad de la República

Director Académico
Dr. Ing. Federico Lecumberry ..... Universidad de la República

Montevideo
lunes 26 noviembre, 2018

Alineación entre audio y partitura para obras del repertorio de la flauta traversa,
Juan P. Braga Brum.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 117 páginas.
Compilada el lunes 26 noviembre, 2018.
http://iie.fing.edu.uy/

Agradecimientos

El agradecimiento va para todas aquellas personas que me acompañaron en esta etapa. En especial a Sandra, Hugo y Nicolás mi familia más cercana que se encuentran a mi lado desde siempre, y para siempre. A Macarena, mi pareja y compañera de vida que brindó su apoyo día tras día y se hizo importante en los momentos más difíciles. A mis amigos. A todos ellos, sin su cariño y paciencia nada de esto podría haber sido posible.

Al Grupo de Procesamiento de Audio del IIE por su apoyo, y por compartir el gusto de la unión entre tecnología y música. Una mención especial a Martín Rocamora por impulsarme en la definición del tema de tesis e introducirme a Luiz, mi director de tesis.

A mis colegas, compañeros de ruta y amigos de la vida Martín Etchart, Pablo Flores y Juan Pablo Garella que con su amistad y experiencia propia me mostraron el camino. También mencionar a Guillermo Carbajal que siempre estuvo con la opinión justa.

A mi director de tesis, Luiz Wagner Pereira Biscainho, que con su paciencia, aplomo y experiencia me acompañó en todo el proceso. A la orientación Musicológica de Osvaldo Budón y su constante apoyo. También a mi director académico Federico Lecumberry que siempre estuvo cuando se lo necesitaba. Sin ellos no podría haber sido posible.

Por último, agradecer a la Agencia Nacional de Investigación e Innovación (ANII) y la Comisión Académica de Posgrados (CAP) por confiar en mi.

Esta página ha sido intencionalmente dejada en blanco.

Al misterio de la música.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

La presente tesis aborda el problema de alineación entre audio y partitura. Éste se define como la sincronización entre una computadora y la interpretación de una pieza musical de partitura conocida. En otras palabras, es la asociación entre dos tipos de datos: muestras de audio digital y notación simbólica de música. Es un tema de investigación que ha captado la atención durante más de 30 años de la comunidad científica en áreas como el Procesamiento de Audio, Machine Learning y Computer Music.

Por otro lado, la flauta traversa es elegida por muchos compositores para la creación de música para medios mixtos[1] (i.e. Música Electroacústica). Teniendo en cuenta que sistemas basados en algoritmos de alineación audio partitura tienen directa implicación en esta corriente musical, se define trabajar con señales de flauta traversa. Dando lugar a uno de los aportes de ésta tesis, la creación de una base de datos de flauta traversa para evaluación de algoritmos de alineación audio partitura. Compuesta de 30 fragmentos de grabaciones reales, con anotaciones manuales y sus respectivos archivos de notación simbólica, se hace pública para su uso con fines académicos.

La última arista del trabajo aparece frente al repertorio contemporáneo de la flauta, donde las técnicas extendidas llevan el material sonoro del instrumento más allá de lo representable con alturas y duraciones (i.e. notas musicales). Con el objetivo de la alineación entre audio y partitura en el repertorio contemporáneo, se hace necesaria la exploración de representaciones matemáticas que sean capaces de extraer las características propias de la nueva sonoridad. Como aproximación al problema, se presenta un caso de estudio sobre Aliento/Arrugas, obra para flauta contemporánea del compositor argentino Marcelo Toledo.

[1] Música donde se combina el material sonoro generado por una computadora con la ejecución de instrumentos musicales.

Esta página ha sido intencionalmente dejada en blanco.

Tabla de contenidos

Capítulo 1

Introducción

1.1. Definición del problema

El seguimiento de partitura es la sincronización entre una computadora y la interpretación de una pieza musical de partitura conocida. Es la asociación entre dos tipos de datos: muestras de audio digital y notación simbólica de música. Es un tema de investigación que ha captado la atención durante más de 30 años, de la comunidad científica en áreas como el Procesamiento de Audio, Machine Learning y Computer Music [OLS03]. El problema se puede dividir en dos grandes enfoques de resolución en *online*[1] y *offline*, cada uno con aplicaciones diferentes y características propias de la estrategia utilizada.

El enfoque offline cuenta con toda la interpretación de la obra mediante un archivo de audio al momento de procesamiento, siendo posible analizar de forma no causal y lograr mayor precisión en la alineación partitura y audio. Se puede ver como el indexado de las muestras de audio según la información de la partitura. En otras palabras, asociar los eventos simbolizados en la partitura con las correspondientes muestras de audio de una grabación, como se esquematiza en la Figura 1.1. La resolución del problema offline tiene diversas aplicaciones de interés como los editores de audio inteligentes que acceden al audio a través de compases y notas de la partitura, búsquedas asistidas en grandes bases de datos a partir de fragmentos de notación musical, herramientas para el análisis automático de parámetros expresivos como son las dinámicas, variaciones de tempo, articulaciones, entre otros [DR06].

Por otro lado la resolución del problema en tiempo real tiene como principal motivación transformar la interacción entre computadora-humano en una experiencia bidireccional, simulando el comportamiento de una interpretación de un músico con otro. En la literatura es usualmente denominado también como acompañamiento automático o músico sintético [Ver84,VP85]. Tiene directa implicación en la música electroacústica de medios mixtos, donde se combina el material sonoro generado por una computadora con la ejecución de instrumentos musicales.

[1] Refiere a análisis en tiempo real

1.2. Alcance y desarrollo de la tesis

Con el objetivo de acotar el alcance de la tesis, se trabaja con el enfoque offline de resolución del problema. Por otro lado, teniendo en cuenta que la flauta traversa es un instrumento de uso extendido en obras compuestas para medios mixtos[2] se define abarcar la problemática desde las señales de flauta. Para esto se construye en el marco de la tesis una base de datos de señales de flauta a partir de grabaciones reales de obras de referencia del repertorio. Siendo éste, un aporte original de la presente tesis que se hace disponible con fines académicos como recurso web[3].

En adición a lo anterior la flauta traversa cuenta hoy en día con un diccionario amplio de técnicas modernas denominadas extendidas. El empleo de estas técnicas define al repertorio contemporáneo con un material sonoro innovador, que escapa de lo representable con las clásicas notas musicales. La alineación entre audio y partitura para este tipo de obras depende entonces, de la exploración otras representaciones matemáticas que no estén basadas en las alturas musicales. Objetivo que será abordado en un caso de estudio sobre la obra contemporánea Aliento/Arrugas para flauta con técnicas extendidas compuestas por el argentino Marcelo Toledo[4].

La tesis se desarrolla de la siguiente manera: en lo que queda de este capítulo 1 se desarrolla sobre el estado del arte y las aplicaciones más extendidas. A continuación, en los capítulos 2 y 3 se presentan las principales características de la flauta traversa con el objetivo de ahondar sobre su naturaleza sonora y el proceso de compilación de la base de datos original de la presente tesis. El capítulo 4 está dedicado a describir la solución propuesta así como el principal algoritmo *Dynamic Time Warping* para alineación de series temporales. En los capítulos 5 y 6 se detalla sobre la conversión del audio y la partitura (respectivamente) a una representación matemática donde pueden ser comparadas entre si por el algoritmo de alineación. Luego, en el capítulo 7 se presentan los resultados de los experimentos con la base de datos de flauta traversa de varias estrategias de alineación, así como el desempeño de la implementación de un tercero con fines comparativos. El caso de estudio de representaciones matemáticas para el material sonoro de la flauta contemporánea se detalla en el capítulo 8. El documento finaliza con conclusiones y trabajo a futuro en el capítulo 9. Además, en los anexos se presentan las partituras de las obras utilizadas en el marco de la tesis.

[2]Numerosos ejemplos de obras para flauta y medios mixtos existen, algunos de renombre: (1) Davidovsky, Mario. 1963. Synchronisms 1. Flauta y sonidos electroacústicos sobre cinta analógica; (2) Lanza, Alcides. 1977. Acufenos III. Flauta, piano y sonidos electroacústicos sobre cinta analógica; (3) Truax, Barry. 1981. East Wind. Flauta y sonidos electroacústicos sobre cinta analógica

[3]Link: https://www.kaggle.com/jbraga/traditional-flute-dataset

[4]http://www.marcelotoledomusic.com/

Figura 1.1: Esquema conceptual de la alineación entre audio y notación simbólica.

1.3. Estado del arte

La resolución del problema de alineación entre audio y partitura es generalmente dividida en dos etapas. En primer lugar, ambas representaciones de la misma pieza musical (i.e. grabación y notación simbólica) deben ser llevadas a un espacio de características donde puedan ser comparables, usualmente llamado como representación intermedia. Ésta transformación genera a la salida dos series temporales de vectores con la misma dimensión. Para posteriormente definir la correspondencia punto a punto entre la representación del audio y la notación simbólica mediante algún algoritmo de alineación.

En 1984 en la conferencia ICMC (*International Computer Music Conference*) aparecen las primeras publicaciones que dieron comienzo a esta linea de investigación. Ambas se basaban en una estrategia de *string matching* para generar la alineación en tiempo real. Dannenberg con la publicación [Dan84] describía un sistema basado en programación dinámica y una representación simbólica de la música de alto nivel. Para esto la partitura y eventos midi generados por el instrumentista en tiempo real eran comparados y alineados. Por otro lado, Vercoe publicó en [Ver84] su intérprete sintético (*Synthetic Performer* por su denominación en inglés) para acompañamiento de flauta traversa. La estrategia se basaba en representar audio y partitura como una lista de alturas (*pitches* por su denominación en inglés). Como en ese entonces los detectores de pitch no eran ni lo suficientemente rápidos ni robustos para trabajar en tiempo real, la estimación de pitch desde el instrumento se hacía con la colocación de llaves que enviaban señales midi al sistema con la posición de los dedos. Posteriormente, en 1990 se introdujo *EXPLODE* por parte de Puckette [Puc90]. Varias piezas musicales para medios mixtos[5] fueron escritas para interpretación basada en este sistema. Si bien la experiencia fue exitosa, los compositores debían sacrificar aspectos musicales para asegurar el correcto funcionamiento del sistema a la hora de dar un concierto.

[5]En Echo de Philippe Manoury para soprano y computadora, tal vez es de las más conocidas.

Teniendo en cuenta que las observaciones (i.e. medidas de la señal de audio) nunca son exactas sea por errores en las técnicas computacionales o errores en la interpretación por parte del músico, aparecieron los enfoques estadísticos para la resolución del problema. La estrategia más extendida en este caso es basada en HMM (de su denominación en inglés *Hidden Markov Models*). No es objeto de esta tesis entrar en detalle de la técnica de HMM, para eso se recomienda el tutorial de Rabiner en la publicación [Rab89]. Por otro lado algunas referencias de su aplicación en alineación entre audio y partitura se encuentran en [MO09, Rap99, OD01].

El enfoque de resolución que se implementa en esta tesis es offline y esta basado en *Dynamic Time Warping (DTW)* debido a que los mayores desempeños reportados en los últimos 10 años lo utilizan (ver tabla 1.1 resumen de la competencia MIREX en alineación audio partitura en tiempo real [CSSR07]). Ésta técnica además, se ha aplicado con éxito en la resolución de problemas de *Speech Recognition* [RJ93].

Existen diversas implementaciones de sistemas de alineación audio partitura con DTW, en la publicación [OS01] una estructura espectral de picos es generada a partir de la partitura y es utilizada para el cálculo de distancia con las ventanas de audio analizadas. Aseguran que esta metodología es aplicable a señales polifónicas logrando mejores resultados y mayor robustez que las técnicas basada en extracción de pitch. Por otro lado en [DR06] se propone la utilización de DTW con extracción de características basadas en la representación tiempo frecuencia denominada como Chromagrama, este mismo sistema es presentado para la resolución del problema de *Music Retrieval* en grandes bases de datos. Dixon en la publicación [Dix05] es el primero en proponer una variante de DTW para la resolución del problema en tiempo real con la información disponible a cada instante, sacrificando desempeño del algoritmo. Además, en [GLB07] el camino óptimo de alineación es calculado a partir de información de alto nivel como es chroma y una estimación de la duración y ritmo local a partir de la señal de análisis.

Un salto cualitativo en los resultados del MIREX (observar tabla 1.1) fue logrado por el algoritmo implementado por J. Carbias y detallado en la publicación [CORSVC$^+$15]. El sistema está separado en dos etapas: una etapa de procesamiento y a continuación la de alineación. En la primera etapa se hace la síntesis de la notación simbólica y mediante el análisis se obtienen patrones espectrales asociados a cada unidad de la partitura. Estos son aprendidos desde el audio generado por la síntesis, mediante la factorización espectral basada en NMF (*Non-Negative Matrix Factorization* por su denominación en inglés). En la segunda etapa la descomposición espectral de la magnitud del espectrograma es realizada con los patrones aprendidos previamente resultando en una matriz de distorsión, que es utilizada como matriz de costo para el cómputo de DTW de forma online. La alternativa presentada por FJ Rodriguez Serrano [RSCOVCMM17], que actualmente tiene el mejor resultado en la competencia, define el estado del arte. El algoritmo esta basado en el de J. Carabias donde el cómputo de la alineación se hace

con DTW incorporando información del tempo de la interpretación, mejorando notoriamente los resultados.

1.4. Aplicaciones

Existen diversas aplicaciones prácticas en procesos musicales que involucran tecnología que pueden ser implementadas mediante la alineación audio partitura.

1.4.1. Enfoque offline

El enfoque offline cuenta con toda la interpretación de la obra mediante un archivo de audio al momento de procesamiento, siendo posible analizar de forma no causal y lograr mayor precisión en la alineación partitura y audio. A continuación se presentan algunas de sus aplicaciones más importantes.

Síntesis expresiva

La síntesis expresiva por ejemplo, tiene como insumo principal anotaciones realizadas grabaciones reales. Ésta tarea involucra mucho tiempo y tedio si es realizada por un ser humano. La alineación entre audio y partitura sirve como método de automatización de éste trabajo [RE16], problema denominado comúnmente por la comunidad científica como etiquetado automático de audio.

Herramientas musicológicas para análisis de la expresividad

La Musicología por otro lado, se vale de herramientas tecnológicas que asisten en los procesos de análisis e investigación. En particular el estudio de la expresividad en la interpretación musical, es abordado mediante la cuantificación de parámetros tales como el tempo, las dinámicas, los recursos idiomáticos de los instrumentos[6] (ej. vibrato, color del tono, etc.). En esta línea, existen herramientas para la visualización [DGW02] de la expresividad y la extracción [Dix01] de estos parámetros.

Editores de audio inteligentes

La edición de audio digital permite la manipulación del material sonoro por ejemplo en la corrección de desafinaciones o el ajuste temporal del inicio de las notas. La realización de estos procedimientos requieren tiempo y experiencia. Mediante el uso de los algoritmos que aquí se abordan la corrección grabaciones puede ser automatizada, lo que se denomina editores de audio inteligente [Dan07].

[6]Recursos musicales propios de la construcción del instrumento y las técnicas particulares de ejecución.

MIREX Real-Time Score Following Task (mejores resultados)

Año	Autor	Representación intermedia	Alineación	Resultado
2017	Francisco J. Bris Peñalver	Alternating Non-Negative Least Squares	Online DTW	94.16 %
2016	Francisco José Rodriguez Serrano	Spectral Factorization	Online DTW	97.43 %
2015	Francisco José Rodriguez Serrano	Spectral Factorization	Online DTW	95.70 %
2014	Chunta Chen*	CQT y Onset Strength Function	Offline DTW	91.46 %
2013	Julio-José Carabias Orti	Non-Negative matrix factorization	Online DTW	86.70 %
2012	Julio-José Carabias Orti	Non-Negative Matrix Factorization	Online DTW	83.01 %
2011	Kosuke Suzuki	Chromagrama y diferencia de primer orden	Online DTW	67.11 %

*El algoritmo presentado por Chunta Chen utiliza la versión offline de DTW a pesar de ser una competencia de algoritmos online.

Tabla 1.1: Resultados de la competencia anual en Real-Time Score Following organizada por Mirex, desde el 2011 al 2017

Separación de fuentes

Por último, la separación de fuentes tiene el objetivo de segregar a los instrumentos musicales a partir del audio con la mezcla. En la publicación [RSMCV+15] se propone un algoritmo que sus autores definen como el estado del arte en el área de lo que se denomina como separación de fuentes asistida.

1.4.2. Enfoque online

Por el contrario el enfoque online por su propia naturaleza, sacrifica el resultado de la alineación para obtener una estimación en tiempo real con la señal de audio disponible. Existen varias aplicaciones interesantes de estos algoritmos, a continuación se detallan las más relevantes.

Acompañamiento automático

Quizás la más importante y ambiciosa de las aplicaciones es la de sistemas capaces de generar un acompañamiento automático. Éste es el objetivo por el cual comienza la línea de investigación hace más de 30 años con [Dan84, Ver84]. Actualmente el software de usuario final más reconocido en el área de Computer Music es el Antescofo [Con08].

Pasador de páginas automático

En la publicación [Arz08] se presenta una aplicación tan interesante como útil, que resuelve el problema a los pianistas de pasar las páginas de las partituras en los momentos justos.

Acompañamiento de Orquesta para despliegue de información

En la publicación [PGHK13] se presenta un sistema que permite a los escuchas de un concierto de orquesta, tener información contextual en tiempo real aumentando la experiencia del concierto. Para esto se utiliza una grabación previa de la obra musical etiquetada y características espectrales para la comparación con la interpretación en vivo.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 2

Flauta traversa

En la presente tesis se trabaja con señales de música generadas a partir de la ejecución de la flauta traversa. En este capítulo se abordará el estudio de la naturaleza de estas señales para fundamentar las decisiones tomadas en el desarrollo de los algoritmos propuestos. Es razonable entonces, entrar en aspectos constructivos del instrumento, así como los principios acústicos de la generación del sonido. Asimismo, se considerarán también aspectos musicales, analizando la interacción entre el compositor y el instrumento como medio expresivo. Es así, que en el presente capítulo además de presentar a la flauta traversa con sus características constructivas, se definen los conceptos de técnicas tradicionales y técnicas extendidas en la ejecución del instrumento.

2.1. Introducción

La flauta traversa pertenece a la clase de instrumentos musicales denominada de *Viento* [Pis55][1]. De forma más específica, integra un subconjunto de esa clase, que se denomina los *Vientos-Madera*[2]. Asimismo, no es la única en su especie ya que con los mismos principios acústicos y diferente tamaño la acompañan el piccolo, la flauta alta y la flauta baja. En lo que sigue, se realiza una reseña de la evolución histórica del instrumento, desde la flauta antigua hasta tal como la conocemos hoy, la flauta moderna.

Es el instrumento más antiguo del que se tiene registro, ha sido construida al menos desde el Paleolítico a esta parte. Las primeras fueron hechas de hueso (ver Figura 2.1), con taladro aproximadamente cilíndrico y con 6 agujeros como máximo [D$^+$07]. La flauta moderna tal como la conocemos hoy, apareció a mediados

[1]Los instrumentos de viento o aerófonos son una familia de instrumentos musicales que producen el sonido por la vibración de la masa de aire en su interior.

[2]Familia compuesta por instrumentos de viento con características heterogéneas. Incluso su nombre *Vientos-Madera* no describe con exactitud a los integrantes. Si bien en algún momento de la historia, la mayoría (menos los Saxos) fueron construidos de madera, al día de hoy se suelen utilizar otro tipo de materiales.

del siglo *XIX* con los principios introducidos por Theobald Boehm[3] en el año 1847. El sistema tenía tres pilares básicos. El primero, promovía la forma cilíndrica en el taladro[4] con excepción de la cabeza del instrumento. El segundo, enunciaba que debía existir un agujero para todas las notas cromáticas en sus posiciones acústicamente correctas, siendo lo más grande posible para mejorar la entonación y el sonido. Por último, el mecanismo debía disponerse de manera que todos los agujeros fueran controlados mediante llaves [Dic75]. Estos principios con mas de 150 años, siguen vigentes y son los que rigen la construcción de la flauta traversa hasta el día de hoy.

Figura 2.1: Evolución de la flauta a lo largo de la historia. (a) flauta antigua de hueso de buitre construida hace más de 35000 años (extraído de página web CBS News). (b) flauta barroca de madera. (c) flauta clásica de madera. (d) flauta moderna de metal, basada en el sistema de Bohem. Imágenes (b), (c) y (d) tomadas de la disertación [D$^+$07]

2.1.1. Partes

La flauta se compone de tres piezas: la cabeza, el cuerpo del instrumento y el pie, como se detalla en la Figura 2.2. El cuerpo es la pieza de mayor extensión y donde se encuentran la mayoría de las llaves. Por otro lado, el pie representa una extensión añadida al cuerpo y aporta los sonidos graves adicionales $C\#4$[5] y $C4$ en el caso de *pie en C*, y $C\#4$, $C4$ y $B3$ para el caso de *pie en B*. Por último la cabeza de la flauta es la que lleva la embocadura, responsable de convertir el flujo de aire del intérprete la excitación periódica que pone a resonar la columna de aire interior al instrumento.

[3]Theobald Boehm (Alemania, 1794-1881): Fue un músico, flautista, compositor, inventor, fabricante de instrumentos y especialista en acústica.

[4]El taladro de un instrumento de viento refiere a la cámara interior, la cual define el camino que describe el aire.

[5]Se utiliza el sistema anglosajón para la nomenclatura, además el índice acústico utilizado cumple que $A4 = 400Hz$.

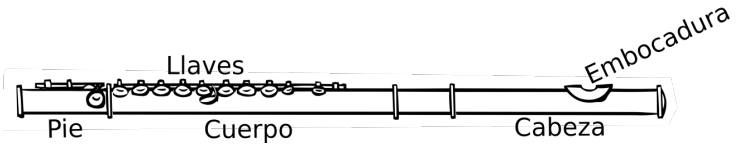


Figura 2.2: Partes de la flauta moderna.

2.2. Material sonoro

Las técnicas para la generación de sonido con el instrumento se clasifican en dos grupos. Por un lado existen las técnicas tradicionales, que son las de uso común del instrumento, asociadas a la generación de música principalmente basada en alturas y duraciones. En complemento a lo anterior, los compositores y flautistas contemporáneos han definido una nueva clase de técnicas llamadas extendidas[6]. Éstas, como su nombre lo indica, extienden las capacidades sónicas del instrumento generando material sonoro que va más allá del definido con alturas y duraciones. No es objetivo de esta sección hacer una descripción exhaustiva del material sonoro generado con las distintas técnicas, para eso existe extensa bibliografía [Pis55, Sam02, Dic75]. Por el contrario, presentar los aspectos relevantes para la comprensión de la naturaleza de las señales de flauta, y los desafíos que presentan las técnicas extendidas en áreas científicas de investigación con el *Music Information Retrieval* (por su denominación en inglés).

2.2.1. Técnicas tradicionales

Las técnicas tradicionales de la flauta son aquellas con las cuales el material sonoro ejecutado es definible mediante los parámetros de altura y duración. Como su nombre lo indica, las mismas refieren al uso tradicional de la flauta traversa y sus mecanismos de producción de sonido. Para comprender la naturaleza de las señales generadas con técnicas tradicionales se comienza identificando dos elementos esenciales: por un lado la producción de la excitación periódica, y en complemento a lo anterior el largo de la columna de aire. Además, se hace un esbozo de las características tímbricas detallando los modos de vibración del instrumento, y se finaliza dejando en claro las limitaciones en registro de la flauta traversa.

Producción del sonido

Existen dos procesos independientes que son los encargados de definir la nota que emite la flauta. Por un lado, la generación de la excitación periódica que pone a resonar la columna de aire y en complemento, el largo de la misma determinada por la configuración de las llaves presionadas por el instrumentista.

[6]El desarrollo de técnicas extendidas tiene estrecha relación con el desarrollo de la música electroacústica desde mediados del siglo XX. La síntesis electrónica y la manipulación electroacústica del sonido introdujeron en la creación musical nuevas sonoridades, que suministraron modelos para la experimentación sonora en la música instrumental

Para poner en oscilación al instrumento, el flautista debe soplar superando en el interior de su boca la presión atmosférica. El trabajo[7] necesario para subir la presión interna y acelerar el aire es la fuente de energía de entrada al instrumento, por lo que al intérprete se lo puede modelar como una fuente continua de energía. Sin embargo, las notas musicales se generan a partir de un movimiento oscilatorio. Estas fluctuaciones periódicas de energía son generadas a partir de la colisión del flujo de aire con el filo del agujero de la embocadura. En otras palabras, la turbulencia provocada por la colisión genera una onda viajera que se traslada a través del flujo de aire. Ésta es la que pone a resonar la columna de aire interior al tubo de la flauta. De esta forma, se genera entonces un sonido de naturaleza periódica denominado como nota musical.

Puesto a resonar el tubo, la frecuencia fundamental de la nota emitida depende estrictamente del largo de la columna de aire oscilatoria. Para el control de este parámetro, existen las llaves del instrumento (en la flauta moderna) que tapan o liberan los agujeros del tubo. Un agujero libre significa la imposición de presión atmosférica en ese punto de la columna de aire, definiendo de esta forma el largo de la misma.

Modos de Vibración

Además de la mecánica de producción de sonido, es de relevancia mencionar aspectos tímbricos del sonido de la flauta. En la práctica, una configuración de llaves en el instrumento permite más de un modo de vibración[8], generando otras alturas musicales que se suman a la de frecuencia fundamental. Se tiene que para una columna de aire con largo determinado (i.e. posición de llaves determinada por el instrumentista) resonando en el interior del tubo, existe emisión simultánea de otras alturas musicales por encima de la de frecuencia fundamental. Éstas dan un sonido característico y se las denomina armónicos. En la Figura 2.3 se observa el espectrograma de una señal de flauta donde se identifican visualmente la frecuencia fundamental con sus armónicos.

De forma teórica se pueden deducir los armónicos permitidos por la construcción del instrumento, modelándolo como un tubo cilíndrico con sus dos extremos abiertos. De esta forma, el modelo impone que la presión en los extremos sea la atmosférica, definiendo dos nodos[9] de presión. Por el contrario, en el interior del tubo la presión no está impuesta, siendo posible las variaciones de energía. De lo anterior, se deduce que la onda de mayor longitud que soporta las condiciones de borde[10] tiene una longitud de dos veces la distancia entre los nodos de presión

[7] En su acepción como concepto de la Física.

[8] Refiere a las ondas estacionarias que un medio de propagación y sus características permiten.

[9] El nodo de una onda refiere, a un punto donde la variación de energía es nula a lo largo del período.

[10] Las condiciones de borde refiere a la imposición de nodos de presión en los extremos del tubo cilíndrico.

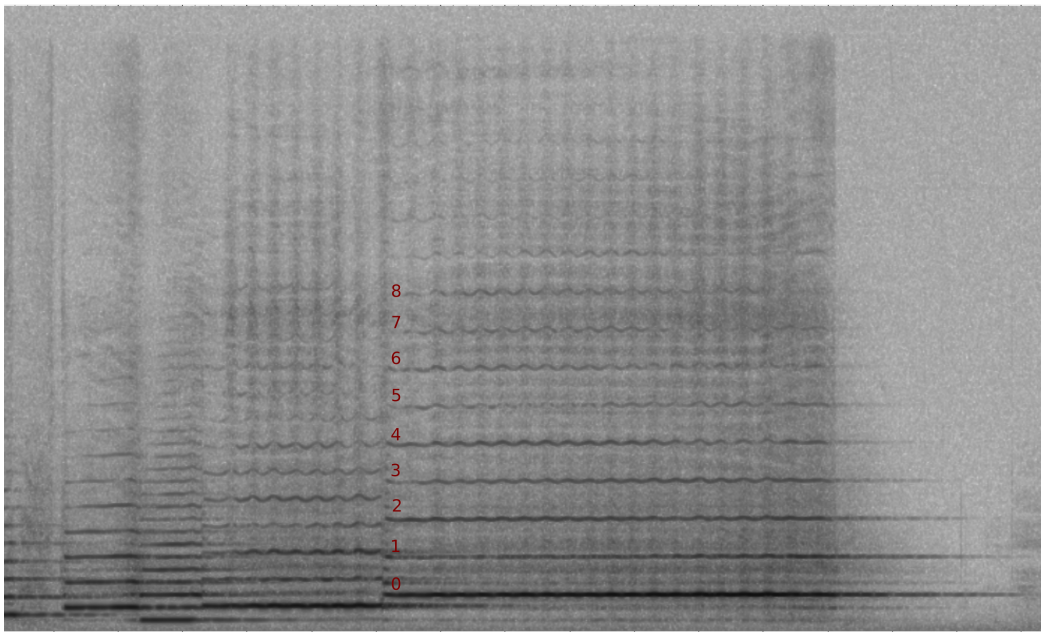


Figura 2.3: Espectrograma de una señal de flauta, se observa la serie armónica numerada sobre el espectro. Audio extraído de la grabación de Density 21.5 por parte del intérprete Jacques Zoon. Imagen generada con Sonic Visualizer.

(en su forma matemática se escribe como $\lambda = 2L$). De la misma forma, se deduce que existen otras longitudes de onda permitidas en este modelo, y se demuestra matemáticamente que cumplen $\lambda = 2L/k$, con $k \in N$.

Por otro lado, la frecuencia del modo de vibración se calcula como la velocidad de propagación de la onda sobre la longitud de la misma, matemáticamente se expresa como $f = v/\lambda$. De la relación anterior se deduce en primer lugar, que la mayor longitud de onda provoca la altura más baja, que en el caso particular de la flauta es la frecuencia fundamental. En segundo lugar, que la estructura armónica de la flauta se puede expresar como $f_i = (i+1)f_0$, donde $i \in N$ y f_0 es la frecuencia fundamental en Hertz.

Registro

Por último, se especifican las notas musicales que son emitibles por la flauta traversa. Esta característica asociada al instrumento se denomina registro, y determina el rango de frecuencias posibles en el instrumento. La cota inferior del registro queda determinada por el pie elegido, para el caso de *pie en C* el limite es el $C4$, por el contrario para *pie en B* es el $B3$. Del otro lado, en la parte alta del registro, la flauta moderna alcanza notas superiores a $C7$, en particular $C\#7$ y $D7$ (observar Figura 2.4). La producción del sonido a partir de $A6$ se vuelve dificultosa, siendo posible para flautistas expertos [Sam02].

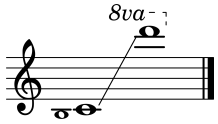


Figura 2.4: Registro de la flauta traversa. Se observa la cota inferior del registro para flauta con *pie en* B y *pie en* C.

2.2.2. Técnicas extendidas

Con el afán de extender el lenguaje musical, los compositores contemporáneos[11] se han dedicado a explorar las capacidades sónicas de los instrumentos musicales. Para esto, se siguen procedimientos como la intervención[12] mecánica de instrumentos o la definición métodos no ortodoxos de ejecución. Es así, que en el caso particular de la flauta se ha derribado el mito de que la sonoridad del instrumento es limitada, y hoy en día existe un diccionario bien definido de técnicas reproducibles, denominadas extendidas [Dic75].

No es objeto de esta sección el detalle exhaustivo de las técnicas extendidas para flauta traversa, mencionar los procedimientos de ejecución más extendidos en la flauta contemporánea. Algunas de las técnicas más conocidas se en listan a continuación (se utilizan las denominaciones en inglés):

- **Flutter Tonguing:** Refiere a la generación del soplido con aleteo de la lengua.
- **Tongue Noises:** Ruidos con la lengua dentro de la embocadura.
- **Percussive Sounds:** Presión de las llaves de forma percusiva.
- **Microtonal Inflections:** Inflecciones microtonales.
- **Multiphonics:** Sonidos multifónicos (más de una nota a la vez con el instrumento).
- **Cantar y tocar a la vez:** Como su nombre lo hace explicito la ejecución de dos alturas a la vez mediante el canto y el instrumento.

La exploración en la música contemporánea abarca también el control tímbrico y la calidad del sonido como un parámetro notado por el compositor. En esta línea, la ejecución de sonidos de banda angosta y banda ancha son muchas veces

[11] E.g. George Crumb (Estados Unidos, 1929), Helmut Lachenmann (Alemania 1935), Salvatore Sciarrino (Italia, 1947).

[12] Denominado también como la preparación de instrumentos. Por ejemplo, el piano preparado de John Cage (Estados Unidos, 1912-1992) y la cabeza móvil en la flauta traversa (*Glissando Headjoint* por su denominación original) de Robert Dick (Estados Unidos, 1950).

elementos buscados desde la composición. En el caso particular de la flauta, este nuevo material sonoro se ejecuta mediante el control de la embocadura[13] así como la presión de aire. En el Capítulo 8 se explora la capacidad de algunas técnicas computacionales en la extracción automática de la embocadura (i.e. detectar automáticamente el tipo de embocadura a partir de el análisis de las muestras de audio) con el objetivo de evaluar la capacidad de representación de algunas características clásicas en el material sonoro generado con técnicas extendidas. Como caso de estudio se utilizan grabaciones de la obra *Aliento/Arrugas* (1998) del compositor contemporáneo Marcelo Toledo (Argentina, 1964).

[13]El término embocadura refiere al aparato de producción de la excitación de la columna de aire, en conjunto con la técnica de soplido.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 3

Base de datos

El objetivo del capítulo es presentar uno de los aportes de esta tesis: una base de datos con señales etiquetadas de flauta traversa ejecutada con técnicas tradicionales. Los datos fueron generados a partir de grabaciones reales de obras referentes del repertorio musical. El capítulo comienza con la descripción de las obras y el porqué de su elección. Se describe posteriormente, el proceso de anotación de las grabaciones, así como la traducción de la partitura a una notación simbólica utilizable por los algoritmos de alineación. Se finaliza presentando algunas características cuantitativas de la base generada y el acceso disponible como recurso web para uso con fines académicos.

3.1. Obras

La flauta traversa cuenta con un repertorio vasto de obras musicales asociado a su larga historia. Diversos compositores en todas las épocas han compuesto para este instrumento musical. En esta sección, se detalla las obras seleccionadas para la evaluación de los algoritmos desarrollados en la presente tesis. Con ese objetivo, se hizo una elección de cuatro obras ejecutadas con técnicas tradicionales, de compositores relevantes de la historia de la música.

Como queda claro en las siguientes secciones, las obras seleccionadas pertenecen a compositores y períodos musicales distintos. De esta forma, el estilo musical resulta sustancialmente diferente para cada pieza. Más aún, si se ordenan de forma cronológica, es notoria la complejización de los recursos musicales utilizados en el procesos compositivos. Este último aspecto plantea un desafío creciente en el propósito de alineación de notación simbólica con audio. En lo que sigue se presentan las obras seleccionadas.

Allemande de BWV 1013

La más antigua de las obras seleccionadas es *La Partita en La menor BWV 1013* esta pieza fue compuesta por Johann Sebastian Bach, alrededor del año 1725. La pieza cuenta con 4 movimientos, de los cuales *Allemande* es el primero y

el único considerado para la base de datos. Como característica principal presenta una estructura rítmica en la que la gran mayoría de las notas tienen el mismo valor de duración (la semicorchea). Además no presenta indicaciones de dinámica ni ornamentaciones notadas por el compositor. Estas características la hacen la más simple de las obras de análisis.

Syrinx

Le sigue *Syrinx* del francés Claude Debussy, obra compuesta para flauta moderna en el año 1913. La interpretación requiere la agilidad brindada por el sistema Bohem, para ejecución de figuras como la fusa y apoyaturas. Además, el compositor agrega indicaciones de dinámica que generan grandes cambios en la energía de la señal, agregando una complejidad extra. También aparecen subdivisiones irregulares como el tresillo.

Density 21.5

Otro francés/norteamericano compone la lista, Edgard Varèse con su composición *Density 21.5* del año 1936. Esta obra también requiere las capacidades de la flauta moderna. Al igual que *Syrinx*, presenta apoyaturas y amplio rango dinámico por indicaciones de intensidad dadas por el compositor. Como característica distintiva, esta obra tiene la aparición del golpe de llave como efecto percusivo y es la que presenta la nota de altura mayor (un D7).

Sequenza I

Por último, la composición de Luciano Berio realizada 1958. De las obras con técnicas tradicionales, es la más compleja en su estructura rítmica. Su versión original presenta notación simbólica no convencional, donde las duraciones no son notadas mediante el uso de las figuras convencionales. Por el contrario, éstas quedan determinadas por proximidad entre las apariciones de las notas. En la Figura 3.1 se observa el comienzo de la partitura. El resultado sonoro es de una alta complejidad rítmica.

3.2. Base de Datos

En esta sección se especifica sobre la base de datos de señales de flauta traversa, generada a partir de grabaciones reales de obras referentes del repertorio musical. Cada audio de la base, viene asociado con un archivo de notación simbólica y otro de anotaciones. En lo que sigue, se explica el papel de cada uno de estos archivos, así como se expresan algunas de las principales características del conjunto de datos.

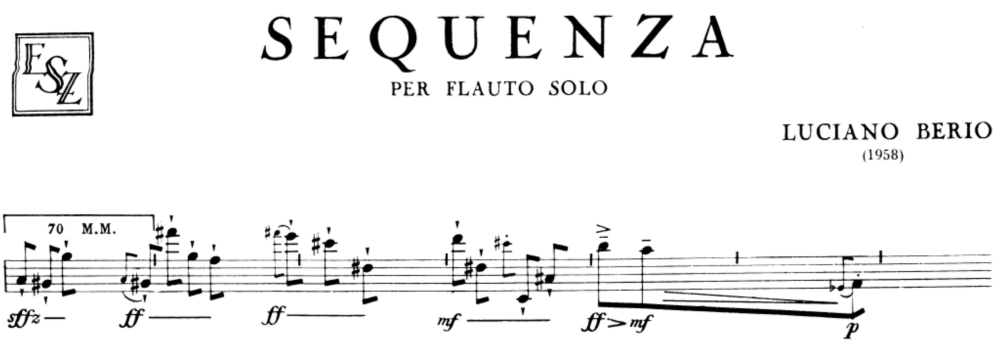


Figura 3.1: Imagen extraída de la versión original de la partitura de Sequenza I. Se observa la notación no convencional para la determinación de las duraciones.

3.2.1. Audio

Los audios de la base de datos están formados por grabaciones de diferentes intérpretes. De forma de agregar variabilidad de interpretación, se tienen distintas grabaciones de flautistas de una misma composición. En concreto, la base de datos se encuentra compuesta por fragmentos de estas grabaciones. Para esto, la elección de los fragmentos se hizo de forma que un fragmento como mínimo engloba una frase musical[1]. Sin intenciones de entrar en una discusión musical, se define que un fragmento es válido en el marco de la tesis, si al escucharla como elemento aislado (sin el contexto de la totalidad de la pieza) da sensación de completitud en el sentido de frase musical. En otras palabras, una idea que comienza, se desarrolla y termina con el transcurrir del fragmento de audio.

A continuación en las Tablas 3.1, 3.2, 3.3 y 3.4 se especifica la cantidad de fragmentos por obra y además los intérpretes de la grabación correspondiente.

Primer	Segundo	Tercer	Cuarto	Quinto
Maxence Larrieu & Aurèle Nicolet	Jean Claude Gerard & Stephen Preston	Aurèle Nicolet & Jean Pierre Rampal	Jean Claude Gerard & Maxence Larrieu	Jean Pierre Rampal & Stephen Preston

Tabla 3.1: En la tabla se detalla para cada fragmento del movimiento Allemande de BWV 1013, los intérpretes responsables de la grabación.

Para Syrinx y Density 21.5, la suma de los fragmentos da como resultado el total de la obra, mientras que para Allemande y Sequenza I no se cumple la misma

[1]La frase musical es una de las unidades más pequeñas en una composición musical. Esta asociada a la sensación de completitud (inicio, desarrollo y fin) de una idea musical, similar a la idea de frase en la composición literaria. En el caso particular de la flauta traversa, esta generalmente asociada a la sección de música entre respiración y respiración [Gro04].

Primer	Segundo	Tercer	Cuarto	Quinto
Bridget Douglas & Philippe Bernold	Doriot Anthony Dwyer & Roger Bourdin	Paul Rhodes & Bridget Douglas	Doriot Anthony Dwyer & Philippe Bernold	Paul Rhodes & Roger Bourdin

Tabla 3.2: En la tabla se detalla para cada fragmento de Syrinx los intérpretes responsables de la grabación.

Primer & Segundo & Tercer	Cuarto & Quinto & Sexto
Jacques Zoon	Lawrence Beauregard

Tabla 3.3: En la tabla se detalla para cada fragmento de Density 21.5 los intérpretes responsables de la grabación.

Primer & Segundo & Tercer & Cuarto
Paula Robison

Tabla 3.4: En la tabla se detalla para cada fragmento de Sequenza I la intérprete responsable de la grabación.

condición. Para más detalles, ir al apéndice A donde se especifica en notación musical cada uno de los fragmentos que conforman la base de datos.

3.2.2. Notación Simbólica

Cada una de las obras tiene asociada una partitura, que en notación musical expresa alturas, duraciones, silencios, y otros parámetros expresivos como son las dinámicas, variaciones de tempo y articulaciones. El objetivo de la partitura es el de notar los aspectos musicales necesarios para la ejecución de la pieza. En la Figura 3.2 se observa un fragmento de la partitura de Partita en La menor, de J. S. Bach.

Con el objetivo de generar archivos de notación musical[2] que cumplan con el doble rol, de por un lado seguir las convenciones visuales de las partituras y por otro pueda ser exportado para utilizar como insumo en los algoritmos de esta tesis, se hizo la transcripción de los fragmentos utilizando la herramienta Lilypond [NN03]. En la Figura A.1 se observa los gráficos generados con la herramienta, mientras que como insumo para los algoritmos se exporta un archivo de texto con la información de altura y duración, en formato *csv* (*comma-separated values* por su denominación en ingles).

[2]La notación musical refiere en la más general de sus acepciones al arte de expresar ideas musicales por medio de la escritura. Aparece como forma de grabación y transmisión de ideas musicales, sustituyendo a la transmisión oral usada hasta ese entonces [Gro04].

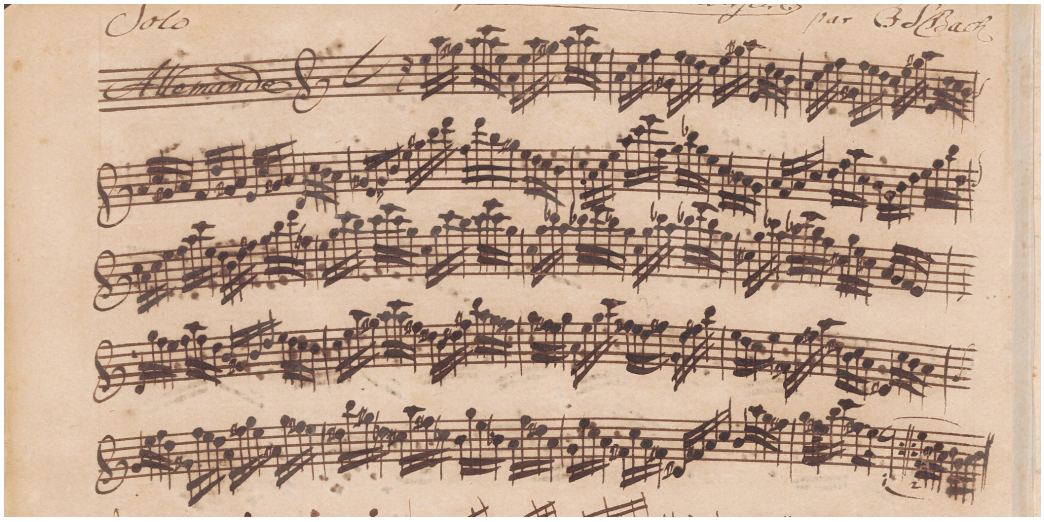


Figura 3.2: Primeros compases del movimiento Allemande de la Partita en La menor BWV 1013. Foto de una partitura escrita a mano.

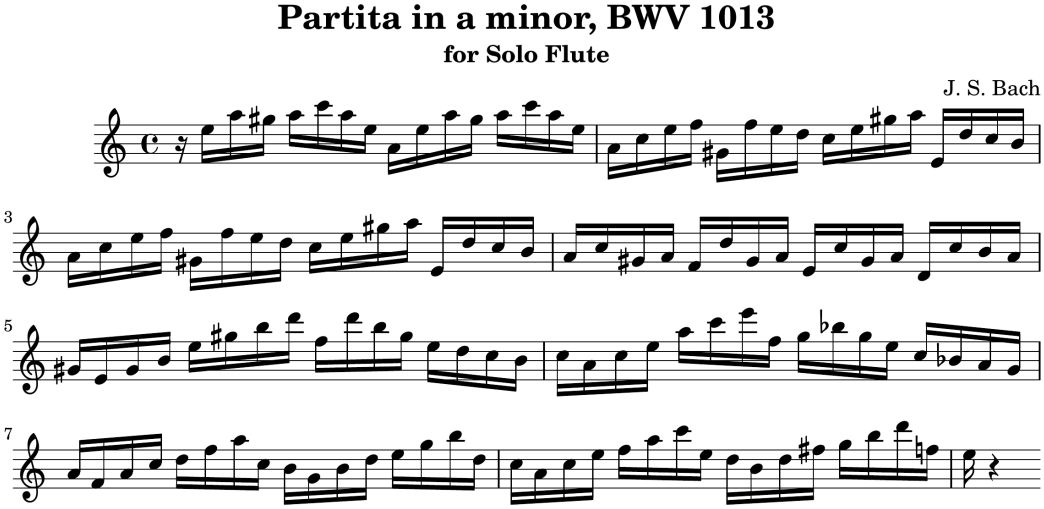


Figura 3.3: Primeros compases del movimiento Allemande de la Partita en La menor BWV 1013. La partitura de la imagen fue generada utilizando la herramienta Lilypond.

3.2.3. Anotaciones

Las anotaciones especifican para cada audio, el comienzo y final de los eventos musicales de interés en forma de etiqueta. De esa forma, sirven como *ground truth*[3] para la evaluación de desempeño de algoritmos. Teniendo en cuenta solamente las alturas y silencios musicales, los archivos de anotaciones especifican el momento

[3]Denominación común (proveniente del inglés) para los archivos de referencia en evaluación de algoritmos.

de comienzo, el tipo de evento[4] y la duración.

Si bien las anotaciones manuales sobre audio significan un proceso largo y tedioso (sobre todo para piezas complejas como es el caso de Sequenza I de L. Berio), la necesidad surge por la falta de disponibilidad de audio etiquetado en grabaciones de las obras seleccionadas. El método de marcado utilizado fue mediante la escucha y la visualización de unas representación tiempo-frecuencia, con el apoyo de la partitura. Este proceso se llevó adelante utilizando la herramienta *Sonic Visualizer* [CLS10]. En la Figura **??** se observa el espectrograma de un fragmento musical con la anotación de notas musicales superpuesta.

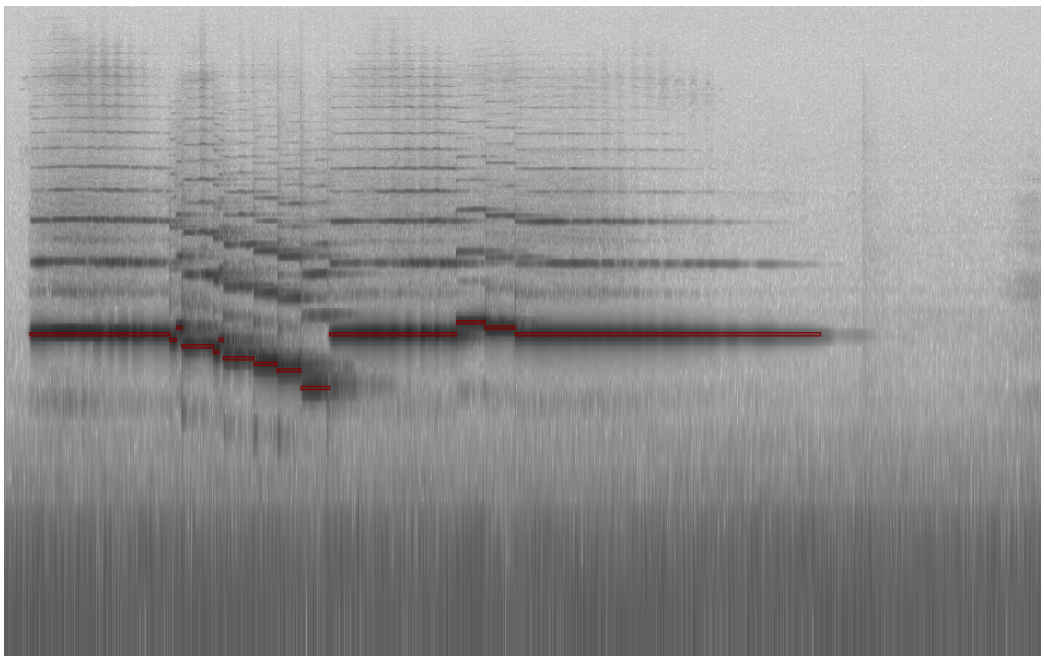


Figura 3.4: Espectrograma de audio con las anotaciones asociadas, se observan notas y silencios musicales. Fragmento musical extraído de la interpretación Philippe Bernold de Syrinx, Debussy. Imagen generada con Sonic Visualizer.

3.2.4. Características de la base de datos

La base de datos cuenta en total con 30 tríos de archivos. Es así que para cada fragmento de audio, le corresponde un archivo de texto con notación simbólica y otro archivo de texto con anotaciones manuales. Además las notas que aparecen en la base van desde $C4$ a $D7$. En total existen 2245 eventos, entre notas y silencios. En la Figura 3.5 se observa el histograma general de la base, donde el eje horizontal especifica el evento musical (notas o silencios) y el vertical

[4]Nota musical o silencio musical. En caso de ser una nota se aclara además la misma.

la cantidad. La base se encuentra accesible para su uso con fines académicos en: https://www.kaggle.com/jbraga/traditional-flute-dataset.

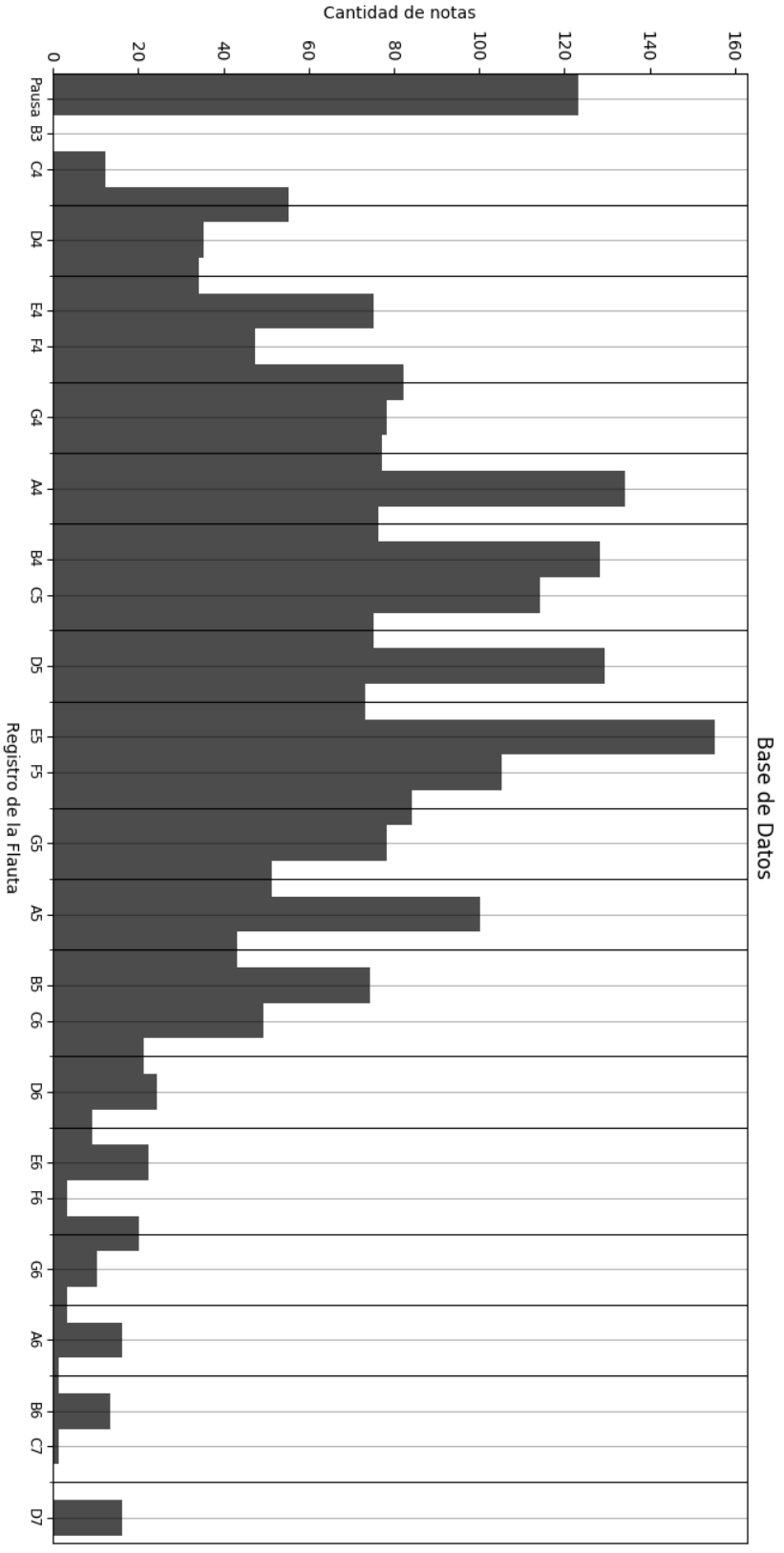


Figura 3.5: Histograma de notas presente la base de datos. Se cuentan solamente las etiquetas, obviando la relación temporal relativa. En el eje horizontal se observa el registro de la flauta, mientras que el vertical la cantidad de etiquetas con el mismo nombre.

Capítulo 4

Solución propuesta

La solución que aquí se propone se puede dividir conceptualmente en dos partes. Por un lado, notación simbólica y audio deben ser transformados de forma que puedan ser comparables. Este es el objetivo de lo que se denomina como representación intermedia (RI). A continuación un algoritmo de alineación de series temporales, como se observa en la Figura 4.1, tiene como entrada las representaciones intermedias de audio y partitura y a la salida una correspondencia temporal punto a punto entre ambas. Para ésta toma de decisión se utiliza el algoritmo clásico denominado como DTW (*Dynamic Time Warping* por su denominación en inglés) [SC78, RRL78, SC07]. En la Figura 4.1 se observa un diagrama general con los bloques recién mencionados.

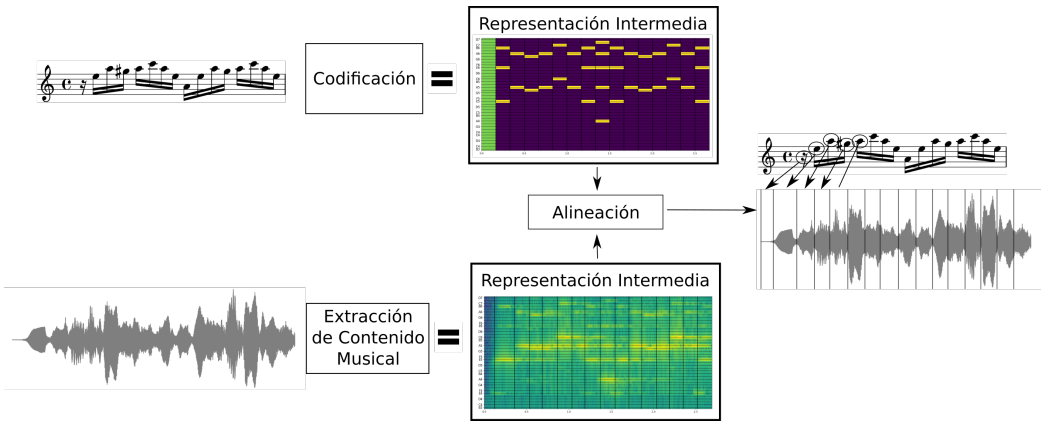


Figura 4.1: Esquema general de la solución del problema de alineación entre audio y partitura utilizada en el presente trabajo.

4.1. Representación intermedia

Una grabación y su notación simbólica son tipos de datos con estrecha relación, pero no comparables de forma directa. Por un lado, la notación simbólica es una

abstracción del material sonoro a ejecutar por el intérprete, dado como una lista de símbolos musicales. Por otro lado, la grabación se la puede pensar como una foto de una ejecución particular de la notación simbólica, inmortalizada en forma de señal digital de audio. La representación intermedia permite computar una distancia en un espacio matemático donde ambas son comparables.

Las señales de flauta con las que se trabaja en la presente tesis son ejecutadas con técnicas tradicionales, como se detalla en en el capítulo 3. Esto implica necesariamente que el material sonoro esté organizado en alturas de la escala cromática de 12 tonos (también llamada como la escala de música occidental). Estos sonidos, por su naturaleza se organizan y simbolizan de dos formas: por alturas absolutas o por clases de altura. Es en estas dos organizaciones de los sonidos musicales es que se basa la representación intermedia utilizada para la resolución del problema de alineación entre audio y partitura.

Se implementaron dos bloques independientes, como se observa en la Figura 4.1 del diagrama general. El primero, denominado bloque de extracción de contenido musical, tiene el objetivo de transformar muestras de audio en representación intermedia y se detalla en el capítulo 5. El otro, denominado bloque de codificación de notación simbólica, así como su nombre lo especifica lleva la partitura a la misma representación y es abordado en el capítulo 6. A continuación se presentan los parámetros más relevantes en el cómputo de la representación intermedia de audio y partitura mediante la solución propuesta en el marco de la tesis:

- **Organización de las alturas musicales:** La representación intermedia puede identificar las alturas de forma absoluta o en clases de altura. Esta elección tiene implicación en ambos bloques de transformación a RI. En primer lugar, para la extracción de contenido musical el algoritmo utilizado en el caso de alturas absolutas es la transformada CQT (sigla de su denominación del inglés *Constant-Q transform*). Mientras que, para clases de alturas mediante una operación conveniente se colapsa la CQT a una octava, obteniendo la representación tiempo frecuencia denominada Chromagrama. En segundo lugar, el bloque de codificación de notación simbólica debe mapear las notas de la partitura en la grilla correspondiente. En la Figura 4.2 se muestra un ejemplo para cada organización de alturas, se utilizó el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.
- **Resolución temporal:** Determina el salto temporal entre vectores de representación intermedia. En otras palabras, la distancia en segundos entre dos vectores consecutivos. Al igual que el anterior, éste afecta ambos bloques de transformación. Por el lado de la codificación, mayor resolución temporal resulta en un aumento proporcional de la cantidad de muestras en todos los símbolos musicales. Del otro lado, un aumento significa un salto menor entre ventanas de análisis generando mayor cantidad de vectores en la RI.
- **Bins en una octava:** Determina la resolución en frecuencia de la representación intermedia. Es así que cuando este valor es igual a 12, cada bin está

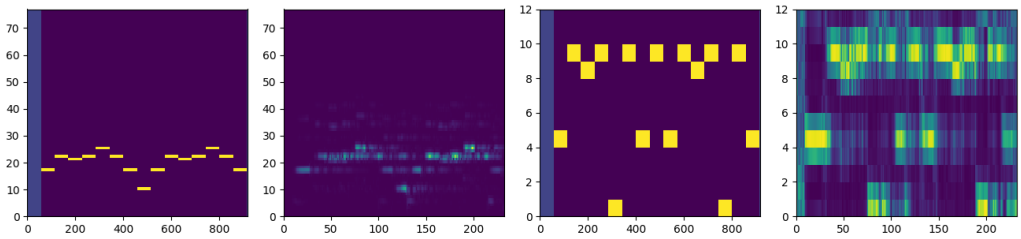


Figura 4.2: Parámetro de organización de las alturas. A la izquierda organización en alturas absolutas para codificación y extracción respectivamente. A la derecha organización en clases de altura para codificación y extracción respectivamente. Para la generación de las imágenes se utilizó el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

directamente asociado a una nota de la escala cromática. Este parámetro afecta a ambos bloques de transformación. En la Figura 4.3 se ejemplifica la acción de este parámetro sobre la representación intermedia. Es importante resaltar, como se vio en el capítulo 5, del lado de la extracción de contenido musical la elección determina el compromiso tiempo-frecuencia. Es decir, con mayor cantidad de bins se mejora la resolución en frecuencia a costo de analizar ventanas temporales que se alejan de la hipótesis de estacionareidad.

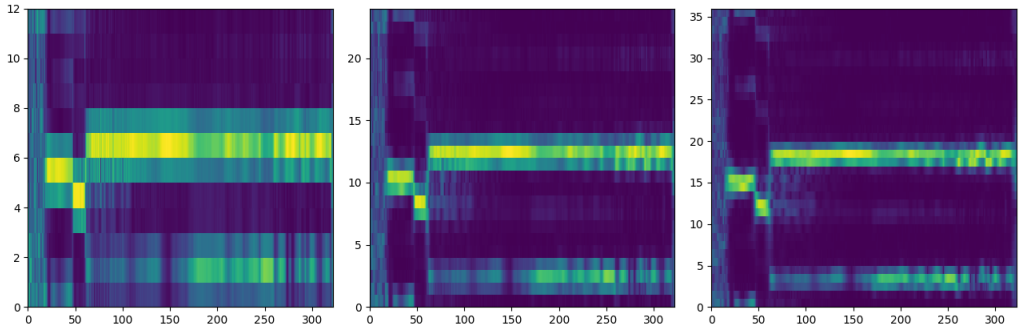


Figura 4.3: Parámetro de resolución en frecuencia. Se utiliza para ejemplificar la organización en clases de altura, es similar para alturas absolutas. De izquierda a derecha, 12, 24 y 36 bins por octava. Para la generación de las imágenes se utilizó el primer compás de Density 21.5 de E. Varese.

- **Cantidad de armónicos:** Como se vio en el Capítulo 2 según el modelo de cilindro con ambos extremos abiertos, la flauta traversa presenta todos los armónicos. De forma que al codificar las notas de la partitura se puede definir cuántos tener en cuenta, en la Figura 4.4 se ejemplifica lo dicho.
- β **(beta):** Al silencio musical se lo codifica como de amplitud constante a lo largo del eje de las frecuencias. Se define entonces el parámetro $\beta \in (0, 1]$ que determina ésta amplitud. En la Figura 4.5 se muestra la influencia de este parámetro sobre la representación intermedia.

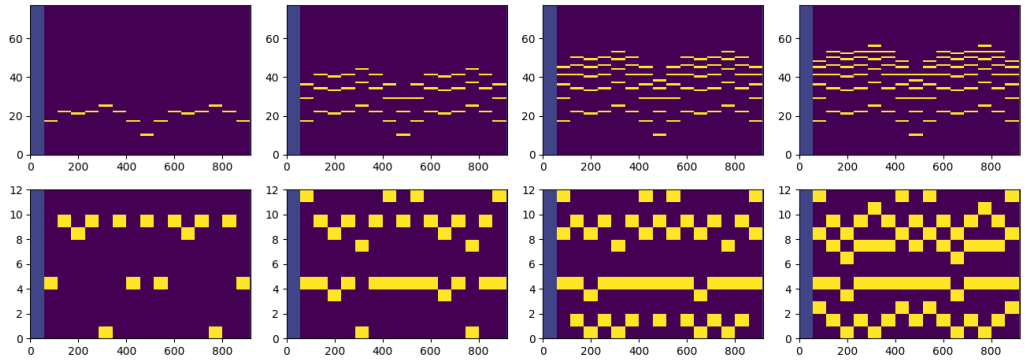


Figura 4.4: Parámetro de cantidad de armónicos en la representación intermedia. Arriba se observa con representación en alturas absolutas, abajo en clases de altura. De izquierda a derecha los armónicos elegidos son 0, 2, 4 y 6. Para la generación de las imágenes se utilizó el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

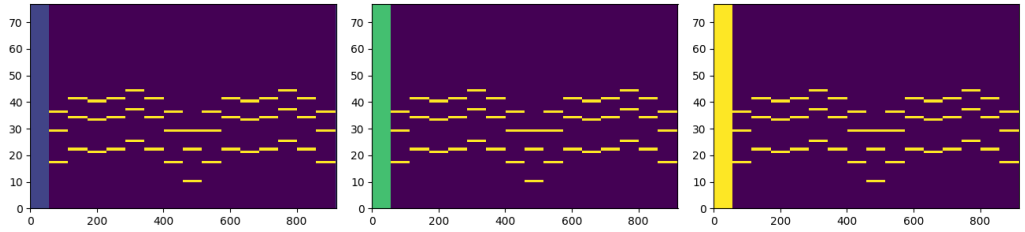


Figura 4.5: Parámetro β que define la amplitud de la representación del silencio musical. De izquierda a derecha respectivamente, se varía el parámetro en $0{,}2$, $0{,}7$ y 1. Para la generación de las imágenes se utilizó el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

- **Sparsity:** Define la proporción de energía del núcleo espectral (núcleo espectral fue definido en el capítulo 5) que es despreciada. Al núcleo espectral se lo puede pensar como un banco de filtros selectivos en frecuencia que ponderan la energía en cada banda de la CQT. Con el aumento de este parámetro, los filtros se vuelven más selectivos enfatizando la energía central del bin. En contraposición, existe posibilidad de no ponderar energía de fuentes existentes que sen encuentren en los limites del bin. En la Figura **??** se observa el efecto de este parámetro en la extracción de contenido musical.

A la salida de los bloques de extracción de contenido musical y codificación de la partitura se obtienen dos series temporales, digamos $\vec{X}$ e $\vec{Y}$ con vectores temporales asociados t_X y t_Y respectivamente. Siendo estos insumo para el algoritmo de alineación que se presenta en la sección contigua.

4.2. Alineación

Para la definición del problema de alineación de forma matemática, supóngase que se tienen dos series temporales $\vec{X} \in \mathbb{R}^{M \times D}$ y $\vec{Y} \in \mathbb{R}^{N \times D}$, donde D es la dimensión del vector de características, y M y N el largo de las mismas respectivamente. La alineación está dada por dos secuencias, dígase $p, q \in \mathbb{N}^L$, que definen la correspondencia punto a punto entre $\vec{X}$ e $\vec{Y}$. Por lo que, de forma matemática se dice que $\vec{X}[p[i]]$ y $\vec{Y}[q[i]]$ están alineados. Para encontrar la correspondencia entre series se debe resolver el siguiente problema de minimización:

$$p, q = \underset{p,q}{\operatorname{argmin}} \sum_{i=1}^{L} d(\vec{X}[p[i]], \vec{Y}[q[i]]) \tag{4.1}$$

Este problema de minimización, con algunas restricciones sobre las secuencias p y q, es resoluble con el algoritmo denominado como DTW (*Dynamic Time Warping* por su denominación en inglés). Esta es una técnica consolidada para la alineación de series numéricas con fuerte correspondencia temporal, como es el caso de señales de voz hablada [RRL78, SC78]. También es el caso, en el problema de alineación entre audio y partitura como se vio en la sección 1.3 dedicada al estado del arte. El tipo de restricciones definen variantes de DTW que son detalladas más adelante en la presente sección.

Por otro lado, DTW es una técnica de programación dinámica por lo que se divide el problema en muchos subproblemas cada uno de los cuales contribuye al cálculo de la distancia total de forma acumulativa. El primer paso es el cómputo de D la matriz de similaridad, que depende estrictamente de la distancia utilizada, el cálculo se define matemáticamente como

$$D[i, j] = d(\vec{X}[i], \vec{Y}[j]) \tag{4.2}$$

donde $D[i, j]$ tiene $M \times N$ elementos donde representan la distancia entre todos los pares de puntos de las series temporales $\vec{X}$ e $\vec{Y}$.

El segundo paso corresponde al computo de C la matriz de costo acumulada. El cálculo se hace de forma recursiva como muestra la siguiente ecuación (esta no es la única forma de calcular la matriz de costo como se verá más adelante),

$$C[i, j] = min \begin{cases} C[i, j-1] + w_h \cdot D[i, j] \\ C[i-1, j] + w_v \cdot D[i, j] \\ C[i-1, j-1] + w_d \cdot D[i, j] \end{cases} \tag{4.3}$$

donde $C[i, j]$ es el costo del camino menos costoso, desde el punto $(1, 1)$ hasta el (i, j). Además $C[1, 1] = d(\vec{X}[1], \vec{Y}[1])$. Los valores $\vec{w} = (w_h, w_v, w_d)$[1] son factores de penalización, donde valores mayores que 1 desalientan movimientos en la dirección correspondiente. A efectos de los cálculos en la presente tesis, siguiendo las recomendaciones de [SC78], se utiliza $\vec{w} = (1, 1, 2)$ sin penalizar ninguna dirección.

[1]Notar que los subíndices refieren respectivamente a dirección horizontal, vertical y diagonal

Luego que se completa el cómputo de la matriz C, se busca el camino de menor costo obteniendo la alineación entre series dada por p y q. Éste se encuentra haciendo recursión hacia atrás desde $C[M, N]$ hasta $C[1, 1]$. El algoritmo se compone de decisiones locales óptimas bajo el supuesto de que el resultado será un mínimo global. En concreto, se comienza desde $C[M, N]$ evaluando todas las celdas vecinas buscando el mínimo, éste se agrega al comienzo del camino y de forma sucesiva el procedimiento finaliza al llegar a $C[1, 1]$. En la Figura 4.7 se observa a modo de ejemplo una matrices C y D.

Por otro lado en las ecuaciones 4.4 y 4.5 se definen de forma matemática dos distancias de uso común en la literatura para la resolución del problema y las usadas en la presente tesis.

$$d_{coseno}(\vec{X}, \vec{Y}) = 1 - \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\|_2 \|\vec{Y}\|_2} \tag{4.4}$$

$$d_{euclideana}(\vec{X}, \vec{Y}) = \|\vec{X} - \vec{Y}\|_2 \tag{4.5}$$

4.2.1. Restricciones sobre el camino de mínimo costo

Las restricciones sobre el camino acotan el universo de posibilidades en la búsqueda del mínimo, disminuyendo el costo computacional en el cálculo de la alineación. La elección correcta de estas restricciones está asociada al conocimiento a priori del problema que se quiera resolver, es así que se pueden aplicar sin atentar contra el resultado final. En lo que sigue se hará mención solamente de las restricciones que fueron aplicadas para los experimentos de la tesis, para más detalle se recomienda el libro clásico de L. Rabiner et al [RJ93] o el más reciente de M. Muller [Mül07].

En el caso particular de alineación entre audio y partitura existe una correspondencia directa entre notación simbólica y las grabaciones de audio. Es claro si se tiene en cuenta que el músico ejecuta la pieza mediante la lectura de la partitura. Ésta característica de las series temporales en el problema planteado, permiten aplicar las restricciones que se especifican a continuación:

- **Limites**: Los limites de la alineación deben cumplir la siguiente condición: $p[1] = q[1] = 0$ y $p[L] = M, q[L] = N$. Es razonable suponer que la grabación empieza y termina con la ejecución del comienzo y el final de la partitura.
- **Monotonicidad**: Las secuencias deben cumplir: $p[i + 1] \geq p[i]$ y $q[i + 1] \geq q[i]$. Teniendo en cuenta que la ejecución de la partitura se hace en una lectura direccionada (i.e. de izquierda a derecha) sin cambios a la dirección contraria parece una restricción acorde.

- **Continuidad**: Por último se impone: $p[i+1] \leq p[i]+1$ y $q[i+1] \leq q[i]+1$. Suponiendo que el intérprete no realiza ningún salto en la lectura de la partitura durante la ejecución no debería resultar en el descarte de una solución válida.

- **Ventana de ajuste**: Teniendo en cuenta que las fluctuaciones temporales entre audio y partitura nunca serán excesivas se puede limitar el cómputo de la matriz de costo a una ventaja de ajuste. Existen varias formulaciones en la literatura, en la presente tesis se trabaja con la denominada de Paliwal en honor a su autor [PAS82]. Matemáticamente se escribe como: $|p[i]\frac{N}{M} - q[i]| < r$, donde el término r es usualmente denominado como radio y define el tamaño de la ventana de ajuste. Notar que la ecuación presenta el factor de escala $\frac{N}{M}$ que permite penalizar de la misma forma ambas dimensiones.

- **Pendiente**: Si la pendiente no se limita el mejor camino puede contener fragmentos puramente horizontales o verticales, permitiendo saltear fragmentos enteros de la partitura o grabación. Esto puede ser favorable en casos donde el intérprete se saltea una parte de la partitura, o ejecuta un material sonoro que no se encuentra escrito. Por el contrario, puede ser desfavorable en casos donde haya correspondencia directa entre partitura y ejecución. Existen varias estrategias para limitar la pendiente del mejor camino, los experimentos realizados en la presente tesis utilizan las propuestas en [SC78]. A continuación se las define matemáticamente:

  - $P = 0$:
  $$C[i,j] = min \begin{cases} C[i,j-1] + D[i,j] \\ C[i-1,j] + D[i,j] \\ C[i-1,j-1] + 2 \cdot D[i,j] \end{cases}$$

  - $P = 0{,}5$:
  $$C[i,j] = min \begin{cases} C[i-1,j-3] + 2 \cdot D[i,j-2] + D[i,j-1] + D[i,j] \\ C[i-1,j-2] + 2 \cdot D[i,j-1] + D[i,j] \\ C[i-1.j-1] + 2 \cdot D[i,j] \\ C[i-2,j-1] + 2 \cdot D[i-1,j] + D[i,j] \\ C[i-3,j-1] + 2 \cdot D[i-2,j] + D[i-1,j] + D[i,j] \end{cases}$$

  - $P = 1$:
  $$C[i,j] = min \begin{cases} C[i-1,j-3] + 2 \cdot D[i,j-2] + D[i,j-1] + D[i,j] \\ C[i-1,j-1] + 2 \cdot D[i,j] \\ C[i-3,j-1] + 2 \cdot D[i-2,j] + D[i-1,j] + D[i,j] \end{cases}$$

  - $P = 2$:
  $$C[i,j] = min \begin{cases} C[i-2,j-3] + 2 \cdot D[i-1,j-2] + 2 \cdot D[i,j-1] + D[i,j] \\ C[i-1,j-1] + 2 \cdot D[i,j] \\ C[i-3,j-2] + 2 \cdot D[i-2,j-1] + 2 \cdot D[i-1,j] + D[i,j] \end{cases}$$

Por último vale resaltar el procedimiento con el cual se llega al resultado final de alineación. Definiendo t_X y t_Y como los vectores de tiempo asociados a $\vec{X}$ e $\vec{Y}$ respectivamente, y además suponiendo que la serie $\vec{X}$ es el resultado de la etapa de extracción contenido musical y el vector $\vec{Y}$ de la codificación de la partitura, el resultado final esta dado por la serie temporal $Y[\vec{q}]$ asociada al vector de tiempo dado por $t_x[p]$. Por último el resultado es transformado de la forma de serie temporal a una lista donde se detalla el comienzo, altura y duración de las notas (ambas son representaciones equivalentes) como se ve en la Figura 4.6.

tiempo(s)	frecuencia (midi)	duración (s)
0.000	0.000	0.150
0.150	76.000	0.242
0.391	81.000	0.150
0.541	80.000	0.178
0.719	81.000	0.150
0.869	84.000	0.207
1.076	81.000	0.150
1.226	76.000	0.150
1.375	69.000	0.207

Figura 4.6: Ejemplo del resultado de etapa de alineación. Se observa la serie temporal representada como comienzo, altura y duración de las notas musicales.

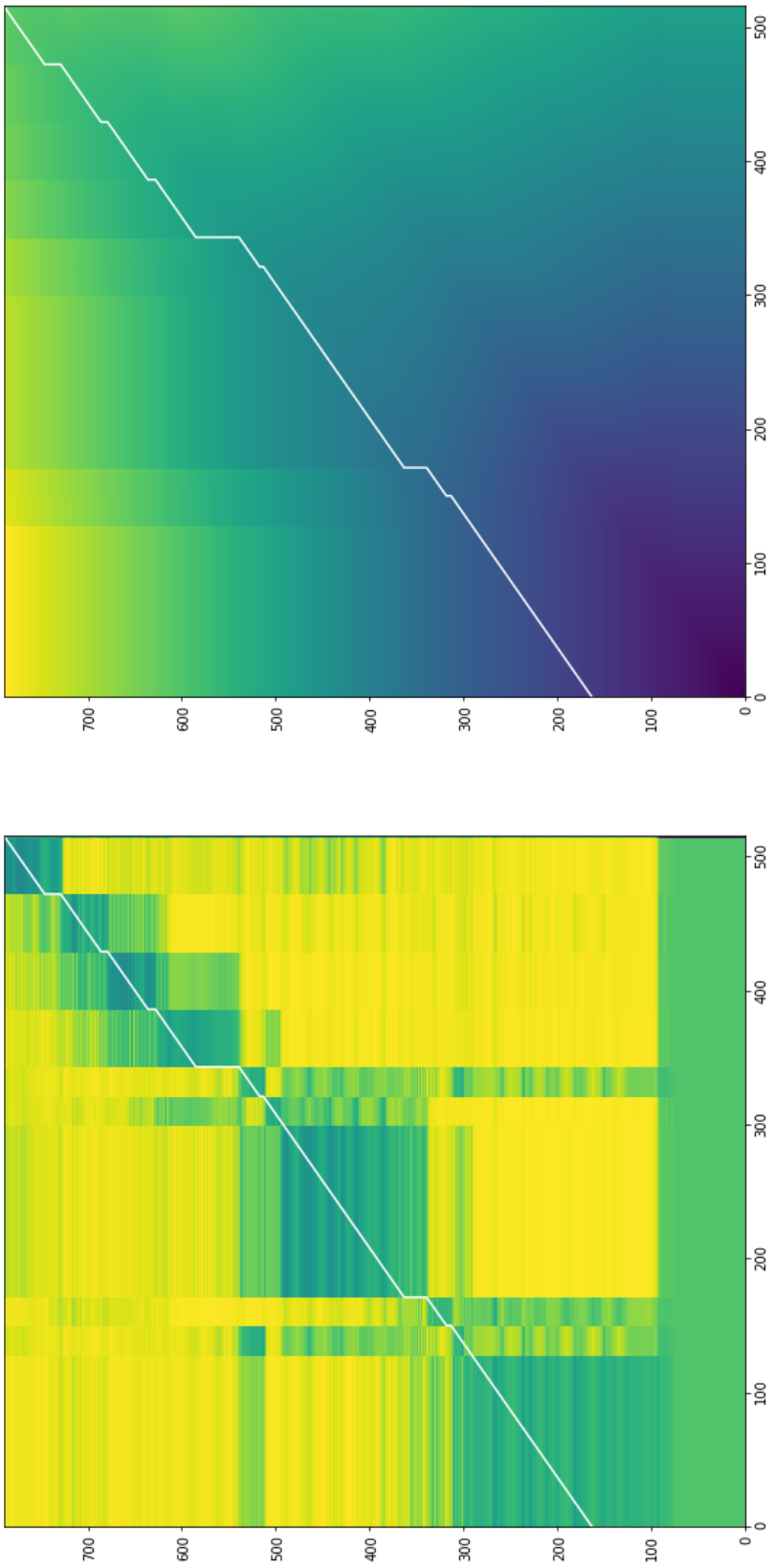


Figura 4.7: Se observa el cómputo de la matriz de similaridad a la izquierda y la matriz C a la derecha.

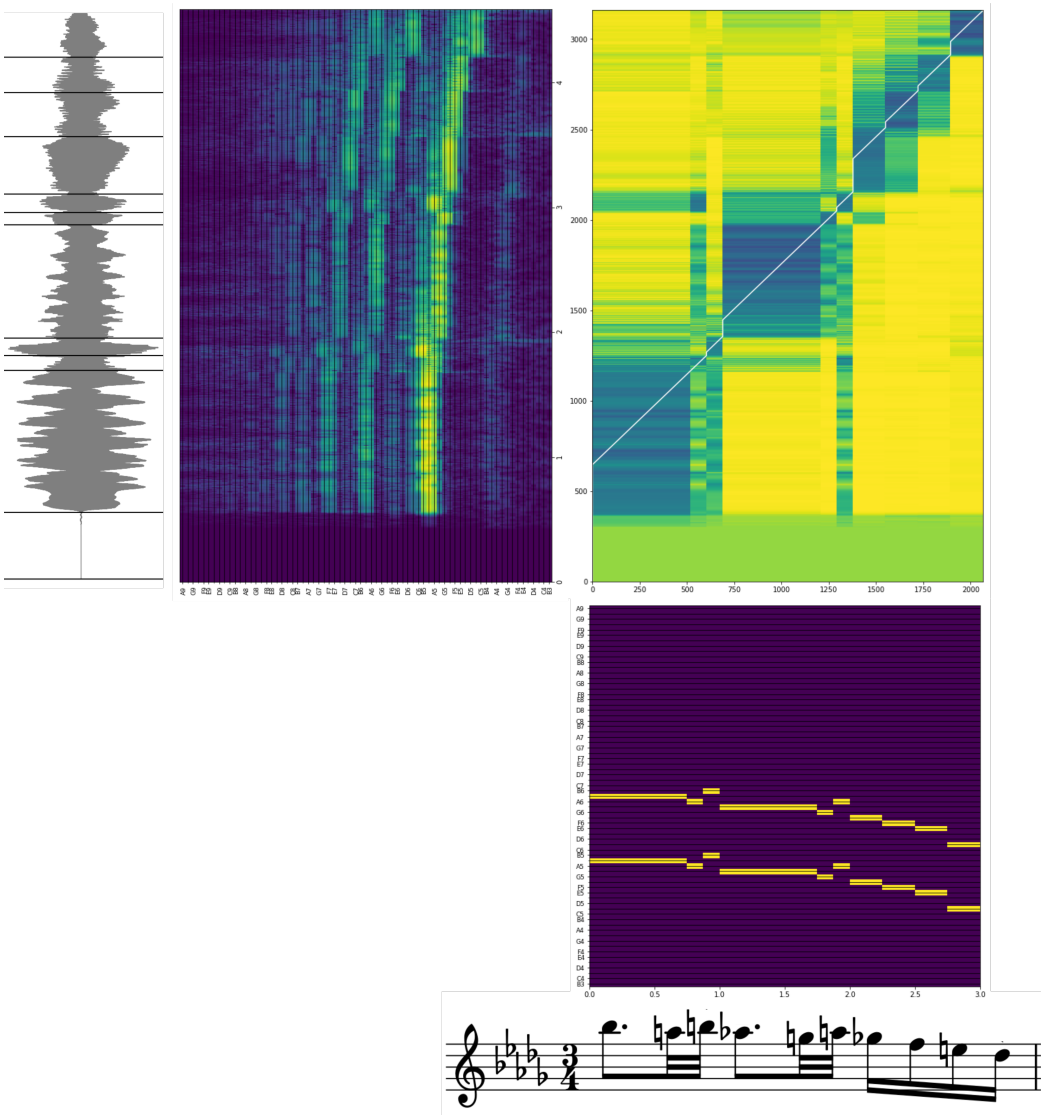


Figura 4.8: Se observa como ejemplo la alineación entre audio y partitura para el primer compás de la obra Syrinx de C. Debussy. Se detalla la partitura y audio del fragmento, además la representación intermedia de cada una de las partes y la matriz de similaridad con el camino óptimo de alineación.

Capítulo 5

Extracción de contenido musical

En esta sección se detalla los algoritmos utilizados en la transformación del audio a representación intermedia. La elección esta basada en la organización básica de los sonidos musicales generados con técnicas tradicionales en la flauta (i.e. la altura). En base a esta relación, los sonidos son organizados de dos formas: de forma absoluta o por clases de altura. Es así que se eligieron dos algoritmos que transforman las muestras de audio en una representación tiempo-frecuencia que cuantifica la energía de la señal en bandas asociadas a la escala musical. En primer lugar se detalla el algoritmo denominado como CQT (sigla de su denominación en inglés *Constant-Q transform*), donde el resultado final es una grilla espectral asociada a alturas absolutas. Para luego seguir con la representación denominada como Chromagrama que determina la energía espectral en bandas directamente relacionadas con las 12 notas de la escala cromática.

5.1. Trasformada espectral Q-constante (CQT)

La escala con *Temperamento Igual* es la organización básica de los sonidos en la Música Occidental. La naturaleza de esta organización esta directamente vinculada a la relación de octava entre sonidos armónicos. En una relación de octava una frecuencia duplica a la otra, es así por ejemplo que las notas $C3$ y $C4$ tienen frecuencias 130,8Hz y 261,6Hz respectivamente. La escala con *Temperamento Igual* tiene 12 sonidos por octava, donde las frecuencias fundamentales se distribuyen de forma geométricamente espaciada. Suponiendo afinación estándar de 440Hz la escala se puede representar matemáticamente como,

$$F_k = 440Hz \times 2^{\,k/12} \text{ con } k \in [-50, 40] \tag{5.1}$$

Por otro lado, la clásica Transformada Discreta de Fourier (DFT por su sigla en inglés *Discrete Fourier Transform*) es de las técnicas más extendidas para el análisis tiempo-frecuencia de señales. Aunque el algoritmo FFT la hace eficiente desde el punto de vista computacional, la resolución constante (en el dominio de la frecuencia) no se adapta de buena forma a la organización de los sonidos en la

	Q Constante	DFT
Frecuencia: f_k	$f_{min}2^{\frac{k-1}{B}}$	$k\Delta_f$
Tamaño de Ventana: $N[k]$	f_sQ/f_k (variable)	N (constante)
Resolución: Δ_{f_k}	f_k/Q (variable)	f_s/N (constante)
Factor de calidad: $\frac{f_k}{\Delta_{f_k}}$	Q (constante)	k (variable)

Tabla 5.1: Comparación de variables relevantes en calculo de CQT y DFT. Extraído de [Bro91].

Música Occidental. A modo de ejemplo, se puede suponer que se tiene una señal de flauta traversa con frecuencia de muestreo $f_s = 44{,}1KHz$ y se realiza la DFT con tamaño de ventana $N = 1024$, para estos valores la resolución de la representación en frecuencia es aproximadamente de $\Delta_f = 43Hz$. La parte baja del registro de la flauta traversa alcanza notas como C_4 y $C\#_4$ de frecuencias $261{,}7Hz$ y $277{,}2Hz$ respectivamente. Es claro que la resolución frecuencial no permite distinguir entre el semitono en la parte baja del registro de la flauta. Por otro lado, en la parte alta del registro se alcanzan notas como $B6$ y $C7$ de frecuencias $1975{,}5Hz$ y $2093{,}0Hz$ respectivamente, donde para este caso la resolución $\Delta_f = 43Hz$ permite distinguir mas a allá del semitono.

Queda claro entonces la necesidad de una representación en frecuencia multiresolución y con *bins* geométricamente espaciados, que permita distinguir las notas de la misma forma en todo el registro musical. Para esto la resolución debe estar directamente relacionada con la frecuencia de los *bins* de forma similar al de la escala con temperamento igual, donde la relación entre el tamaño del semitono y la frecuencia fundamental es constante (aproximadamente del 6 %).

Si bien existen diversas formulaciones con aplicaciones fuera de señales musicales [YB78, Har76, Hel76, OJS71, BO74] no son objeto de esta tesis. La misma se centrará en la CQT con *bins* geométricamente espaciados y resolución mínima de 12 *bins* por octava. La implementación utilizada en los experimentos del presente trabajo [MRL+15] esta basada fuertemente en la publicaciones [BP92, SK10] como se detalla a continuación.

5.1.1. Formulación matemática

Como se detalla en la publicación [Bro91] la representación espectral CQT puede ser directamente calculada mediante una evaluación conveniente de la DFT. El *k-ésimo* componente de la DFT para una señal de análisis $x[n]$ queda determinado por la expresión matemática,

$$X[k] = \sum_{n=0}^{N-1} w[n]x[n]e^{-j2\pi kn/N} \tag{5.2}$$

donde $w[n]$ refiere a la ventana de análisis y N su tamaño. Además, como se puede corroborar en la expresión anterior, la frecuencia central del *k-ésimo* componente

Nota	Midi	Hz	Muestras	Nota	Midi	Hz	Muestras	Nota	Midi	Hz	Muestras
B3	59	246.9	3003	C5	72	523.3	1417	**D#6**	87	1244.5	596
C4	60	261.6	2835	C#5	73	554.4	1338	E6	88	1318.5	562
C#4	61	277.2	2676	D5	74	587.3	1263	F6	89	1396.9	531
D4	62	293.7	2525	D#5	75	622.3	1192	F#6	90	1480.0	501
D#4	63	311.1	2384	E5	76	659.3	1125	G6	91	1568.0	473
E4	64	329.6	2250	F5	77	698.5	1062	G#6	92	1661.2	446
F4	65	349.2	2124	F#5	78	740.0	1002	A6	93	1760.0	421
F#4	66	370.0	2004	G5	79	784.0	946	A#6	94	1864.7	398
G4	67	392.0	1892	G#5	80	830.6	893	B6	95	1975.5	375
G#4	68	415.3	1786	A5	81	880.0	843	C7	96	2093.0	354
A4	69	440.0	1686	A#5	82	932.3	795	C#7	97	2217.4	334
A#4	70	466.1	1591	B5	83	987.8	751	D7	98	2349.3	315
B4	71	493.9	1502	**C6**	84	1046.5	709	D#7	99	2489.0	298

Tabla 5.2: Tabla representativa de los filtros de la CQT. Se detalla frecuencia central y tamaño de ventana para el rango de frecuencias B3-D#7 con 12 *bins* por octava. Se omiten C#6 y D6 por facilidad para dar formato a la tabla.

es $f_k = f_s k/N$ (o su equivalente en frecuencia digital $\omega_k = 2\pi k/N$) y la resolución es constante a lo largo del dominio con valor $\Delta_{f_k} = f_s/N$. Por lo que el factor de calidad para la DFT es $Q_k = f_k/\Delta_{f_k} = k$. De lo anterior, se deduce que para el cálculo de una representación tiempo-frecuencia con Q constante, el tamaño de la ventana debe variar inversamente con la frecuencia. De esta forma N y $w[n]$ se transforman en funciones de k simbolizadas por $N[k]$ y $w[n,k]$ respectivamente. Por lo que para el *k-ésimo* componente centrado en f_k se tiene que $N[k] = f_s/\Delta_{f_k} = f_s Q/f_k$. Resultando en la expresión $f_k = f_s Q/N[k]$ que determina la frecuencia del *k-ésimo* componente, siendo su equivalente en frecuencia digital $\omega_k = 2\pi Q/N[k]$. La re-formulación de estos parámetros relevantes es resumida en la Tabla 5.1. Por último, sustituyendo ω_k, $N[k]$ y $w[n,k]$ en la expresión 5.2 se deduce la formulación para CQT que se detalla a continuación,

$$X^{CQT}[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w[n,k]x[n]e^{-j2\pi Qn/N[k]} \tag{5.3}$$

donde el factor de normalización $1/N[k]$ aparece para compensar la dependencia con k del número de términos en la sumatoria. Para los cálculos en la presente tesis, la mínima resolución esta determinada por la cantidad de semitonos en la escala cromática. Es así que como mínimo se trabaja con 12 bins por octava. Por otro lado, para analizar todo el espectro es necesario que $\Delta_{f_k} = f_{k+1} - f_k = f_k(2^{\frac{1}{B}} - 1)$ por lo que con $B = 12$ se cumple que $Q = 1/(2^{\frac{1}{12}} - 1) \approx 17$. De esta forma queda determinado el tamaño de las ventanas de análisis de la CQT en función de la cantidad de Bins por octava, siendo este parámetro el que define el compromiso tiempo frecuencia de la transformada. En la tabla 5.2 se observa el tamaño de las ventanas de análisis para el caso de 12 bins por octava en todo el registro de la flauta (i.e. B3-D#7).

5.1.2. Algoritmo

Debido a que la implementación de CQT mediante la evaluación directa de la expresión 5.3 conlleva un alto costo computacional, un algoritmo eficiente fue propuesto por Brown y Puckette en [BP92] apoyándose en el uso de la Transformada Rápida de Fourier (FFT por su sigla en inglés de *Fast Fourier Transform*). Se reescribe la expresión de CQT como una multiplicación de matrices tal que $X^{CQT} = x \cdot \mathcal{K}^*$, donde x es la señal de análisis dispuesta como vector fila de largo N (con $N \geq N[k]\ \forall k$) y $\mathcal{K}^*$ es el complejo conjugado del núcleo temporal (denominado originalmente por los autores como *Temporal Kernel*) y de expresión matemática siguiente,

$$\mathcal{K}^*[n,k] = \begin{cases} \frac{1}{N[k]} w[n,k] e^{-j2\pi Qn/N[k]}\,, & \text{con } n < N[k] \\ 0\,, & \text{en otro caso} \end{cases}$$

de la relación de Parseval para la DFT [Opp75] se obtiene la expresión equivalente,

$$X^{CQT}[k] = \sum_{n=0}^{N-1} x[n]\mathcal{K}^*[n,k] = \frac{1}{N} \sum_{k'=0}^{N-1} X[k']K^*[k',k] \tag{5.4}$$

donde $X[k']$ y $K[k',k]$ denotan la Transformada de Fourier para $x[n]$ y $\mathcal{K}^*[n,k]$ respectivamente. Además, el término $K[k',k]$ es denominado núcleo espectral (traducción de la denominación original de los autores *Spectral Kernel*).

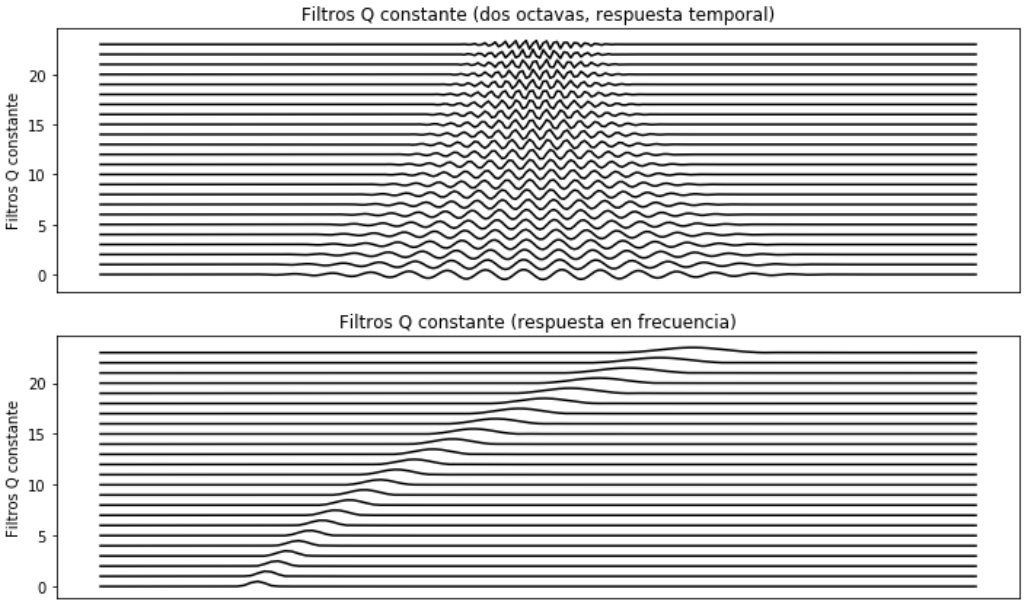


Figura 5.1: Arriba la parte real del núcleo temporal con dos octavas de 12 *bins* cada una. Abajo la representación en frecuencia o lo que es equivalente el núcleo espectral.

La eficiencia del algoritmo reside en primer lugar en que el Núcleo Espectral se calcula una sola vez, utilizando la FFT con su bajo costo computacional. En

segundo lugar teniendo en cuenta que el núcleo temporal esta conformado por sinusoides moduladas, $K^*[k', k]$ es prácticamente cero en todo el dominio espectral, por lo que mediante una umbralización conveniente resulta en pocas multiplicaciones despreciables frente a las de FFT. Además es real y simétrico con respecto al 0, dado que $\mathcal{K}^*[n, k] = \mathcal{K}[-n, k]$, por lo que el número de multiplicaciones se puede reducir a la mitad, utilizando sólo las frecuencias positivas del dominio y duplicando el resultado.

La implementación de Brown y Puckette presenta dos problemas que afectan su eficiencia. En primer lugar, cuando un rango amplio de frecuencias es considerado (por ejemplo 8 octavas de $60Hz$ a $16kHz$), aumenta significativamente la cantidad de bloques de transformación y además el núcleo espectral deja de ser mayoritariamente cero para altas frecuencias, aumentado considerablemente el número de multiplicaciones. Por otro lado, en altas frecuencias el tamaño de la ventana N_k se vuelve pequeño, por lo que para no dejar fragmentos de la señal sin analizar el salto entre ventanas debe ser pequeño también (por ejemplo $N_k/2$), generando un aumento en el costo computacional del algoritmo.

Teniendo en cuenta los aspectos mencionados anteriormente, en la publicación [SK10] se propone otra implementación eficiente basada en la de Brown y Puckette [BP92].

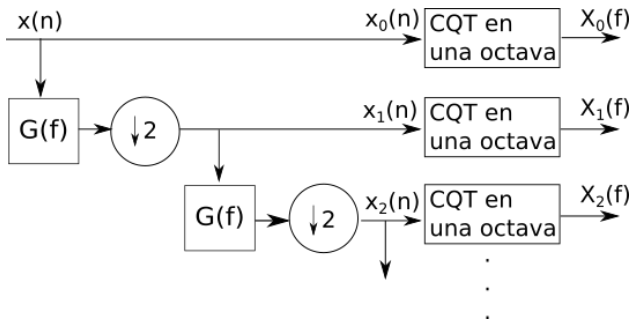


Figura 5.2: Esquema del algoritmo para cálculo de CQT. $G(f)$ es un filtro pasa-bajos y $\downarrow 2$ simboliza sub-muestreo con un factor de 2.

Los autores en [SK10] proponen una extensión al algoritmo para superar los casos donde la eficiencia se ve perjudicada. La estrategia está basada en procesar la señal por octavas. Se utiliza el núcleo espectral $\mathcal{K}$ de forma que solamente produce la CQT para la octava más alta. Luego, se aplica un filtro pasa-bajos a la señal seguido de una etapa de sub-muestreo a la mitad y se reutiliza el núcleo $\mathcal{K}$ obteniendo la CQT para la siguiente octava más alta. Este proceso es aplicado sucesivamente hasta obtener las octavas requeridas, en la Figura 5.2 se esquematiza el proceso.

Como el núcleo espectral $\mathcal{K}$ cubre solamente una octava, el bloque de transformación se puede hacer considerablemente pequeño (el número de muestras N_k

asociados el *bin* de la CQT de baja frecuencia), además la matriz $\mathcal{K}$ para los *bins* de altas frecuencias es mayoritariamente cero, resultando en una matriz rala.

5.2. Chromagrama

El chromagrama de notación matemática $s(t, c)$, es una distribución conjunta donde se representa la energía de la señal de análisis en función de las variables tiempo y el chroma. Tiene estrecha relación conceptual con las clásicas distribuciones tiempo-frecuencia (TFD por su sigla en Inglés de *Time-Frequency Distribution*) donde el eje de las frecuencias es sustituido por el de chroma.

El chromagrama está basado en la medida de chroma, que tiene estrecha relación con la música occidental. Por lo que es una herramienta matemática relevante para aplicaciones con señales de música, obteniendo un montón de ventajas frente a otras representaciones tiempo-frecuencia. Se destaca que es menos dependiente a las variaciones en el timbre y al ruido, que otras características computacionales. Además, es robusto frente a cambios en dinámica y errores de octava. Estos últimos son típicos en algoritmos de detección de *pitch*. Como contrapartida, la información absoluta de octava se pierde siendo imposible distinguir intervalos mayores o iguales al de octava. En lo que sigue se detalla el concepto de chroma, se define de chromagrama y se detalla el algoritmo utilizado para el cálculo computacional.

5.2.1. Sobre el Chroma

Shepard en la publicación [She64] reporta que la percepción humana de *pitch* (por su denominación en Inglés) necesariamente debe ser modelada con dos parámetros. En contraposición con las corrientes de la época, que modelaban con un sólo parámetro dando un carácter lineal a la percepción de *pitch*. Además, determina que la percepción del sistema auditivo es mejor representada con una curva en forma de hélice, y define dos parámetros: altura del tono y chroma, para caracterizar la dimensión vertical y angular respectivamente. La altura de tono describe el aumento en la percepción de *pitch* al aumentar la frecuencia. Por otro lado, el chroma es cíclico por naturaleza y de período la octava en frecuencia.

De acuerdo a los resultados de Shepard, la percepción de *pitch* (p) puede ser factorizada entonces en valor de chroma (c) y altura de tono (h),

$$p = 2^{c+h} \tag{5.5}$$

para que esta descomposición sea única es suficiente con que el chroma (c) sea un número real entre 0 y 1, y la altura de tono (h) un entero. Cambios lineales en c determinan cambios logarítmicos en frecuencia. Al dividir el intervalo $[0, 1)$ en 12 partes iguales, la escala cromática con temperamento igual es obtenida.

Los anteriores resultados confirman por un lado el rol central que durante siglos ha tenido el intervalo de octava en la música occidental, y por otro, la práctica común de identificar las notas por clase de altura y número de octava.

5.2.2. Algoritmo

Para su cálculo existen diversas estrategias, que se pueden dividir en transformaciones directas de la señal o transformaciones de una TFD pre-calculada [Wak99a]. Las más reportadas en la literatura son las segundas, por lo que se deja de lado en el presente documento las transformaciones directas de la señal.

Las implementaciones basadas en TFD varían de publicación en publicación pero en todos los casos se procede de la siguiente forma [Ler12]:

- se calcula una TFD a partir de la señal de análisis y se agrupa en el dominio de la frecuencia por semi-tonos
- se calcula alguna medida de saliencia para cada banda
- y por último se suman (en todas las octavas) las correspondientes a cada clase de altura

Wakefield hace un planteo matemático de lo dicho anteriormente en [Wak99b]. Para esto, define una función de agregación $G(.)$, que mapea una TFD en la entrada, en una representación chroma-frecuencia de salida,

$$s(t, c) = G(s(t, f); \forall\ f = 2^{c+h}) \tag{5.6}$$

donde $s(t, f)$ es la TFD de entrada, $c \in [0, 1)$ y $h \in \mathcal{Z}$. Quedan planteadas dos decisiones de implementación, por un lado la elección de la TFD de entrada y por otro la función de agregación $G(.)$. En el presente trabajo el cálculo de la TFD se hace con la CQT. En adición la función de agregación suma todos los bins espectrales asociados al mismo chroma, matemáticamente:

$$s(t, c) = \sum_{k} s(t, 2^{c+k}) \tag{5.7}$$

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 6

Codificación de la notación simbólica

La notación simbólica de la partitura es traducida al dominio de la representación intermedia mediante un bloque de codificación, tal como se muestra en el esquema de Figura 4.1. Para la codificación se tienen en cuenta solamente los parámetros musicales de altura y duración, dejando de lado las dinámicas, articulaciones y otros tipos de recursos musicales. En lo que sigue se detalla los elementos de notación musical que son relevantes para la conversión de la partitura. Además, se especifica el procedimiento de codificación de la misma a la representación intermedia.

6.1. Elementos de Notación Musical

Los elementos de notación musical relevantes en el marco de este experimento, son aquellos que definen el material sonoro desde los parámetros de altura y duración. Usualmente el símbolo utilizado para representar unívocamente estos aspectos del material sonoro es llamado nota[1] musical. El cometido de esta sección es presentar la nota musical, así como su relación con el pentagrama para la codificación en representación intermedia, a modo de síntesis de los elementos de notación musical relevantes para la tesis. La notación musical es un tema muy vasto, que no se pretende abarcar aquí, se recomienda el libro [RR64] para profundización en el tema.

En la Figura 6.1 se observa el elemento de notación llamado como nota musical. El mismo esta compuesto por tres partes: la cabeza, la plica y el corchete. Estos elementos determinan la duración relativa de la nota. Por otro lado la altura queda definida por la posición de la cabeza de la nota en el pentagrama.

El pentagrama consta de 5 líneas horizontales paralelas, así como los espacios determinado entre líneas. Es una organización vertical de los sonidos, donde cada una de los espacios y lineas corresponde a una altura musical, ordenadas de abajo hacia arriba en relación a la sensación perceptiva de altura. La referencia de altura

[1]Nota proveniente del latín y su significado es el de marca o signo.

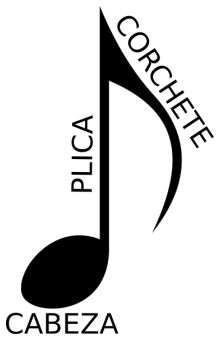


Figura 6.1: Partes de una nota musical, por claridad se elige la corchea como ejemplo ya que presenta las tres partes definidas de una nota musical.

se determina con un elemento nominado como clave, donde define la altura absoluta para una linea del pentagrama. A partir de lo anterior, de forma secuencial se deducen el resto de las alturas para cada espacio y línea del pentagrama. Para la flauta traversa se utiliza usualmente la Clave de Sol en segunda línea, que determina la altura musical G4 tal y como se observa en la Figura 6.2.

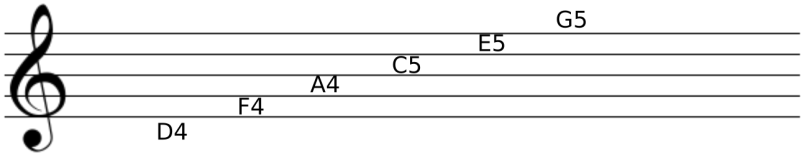


Figura 6.2: Pentagrama con Clave de Sol en segunda línea.

Por otro lado, la duración relativa de la nota queda determinada por la cabeza de la nota, la plica y el corchete. Estos elementos se van agregando de forma secuencial para la disminución de la duración a la mitad de la anterior. En la Tabla 6.1 se observa las duraciones con su símbolo correspondiente para la notación utilizada en el marco de la tesis. Se observa que están relacionadas por un factor de 2.

Además de las figuras, mencionadas anteriormente es posible operar con las duraciones definidas por la figura. Para esto existen dos elementos de notación que se presentan aquí: la ligadura y el puntillo. En primer lugar la ligadura simboliza la suma en duración entre las figuras involucradas, por otro lado el puntillo suma a la duración notada su mitad. De las definiciones anteriores es fácil deducir que la duración una negra con puntillo es igual a una negra ligada a una corchea.

El silencio musical es un elemento de notación el cual queda definido solamente por su duración. Si bien existe una convención en la posición vertical en el pentagrama para cada símbolo de pausa, ésta no es relevante en su significación.

El calderón también conocido como fermata, representa un punto de reposo musical, y se ejecuta alargando la duración de las figuras musicales a las que afecta. Su posición es inmediatamente arriba de la nota o silencio donde se aplica.

Los símbolos de duración están organizados de forma binaria y relativa, es así que una corchea es la mitad de una negra y una fusa la mitad de una semi corchea. Existen también elementos de notación que modifican la relación binaria en otros tipos de subdivisión. En función de las obras seleccionadas para la presente tesis, basta aquí con presentar el tresillo y el quintillo. El primero modifica la subdivisión binaria en ternaria, es así que un tresillo de corchea divide en tres la duración de una negra. De forma análoga, un quintillo de semi corchea subdivide a la negra en 5 partes iguales. Por claridad, la convención en notación exige que la nueva subdivisión no puede ser mayor que el doble de la figura superior, es así que el quintillo de semicorchea no divide a la corchea en 5 partes iguales, sino que como fue dicho anteriormente, divide a la negra en 5 partes iguales.

El último elemento referente a duración que se especifica en esta sección es el encargado de transformar la duración relativa en absoluta, y es denominado como tempo. El tempo se expresa generalmente en BPM (*beats per minute* por su denominación en inglés), notando cuantas negras entran en un minuto.

Las obras seleccionadas, presentan recursos musicales de los denominados ornamentos. Estos elementos de notación no son estrictamente necesarios para el devenir melódico, y son agregados por los compositores con el objetivo de adornar o embellecer. Por esta característica es que en general pueden ser obviados en la ejecución de la notación simbólica. En el marco de la tesis se trabaja con dos tipos de ornamentos: el trino y las notas de gracia. En primer lugar el trino es una alternación rápida entre dos notas adyacentes dentro de la escala diatónica. Dicho en otras palabras, la distancia entre las notas es de un tono o semitono y es notada según el pentagrama. Existen dos tipos de trinos, los ascendentes y los descendentes, su denominación refiere a la altura de la segunda nota con respecto a la primera. El trino es relativamente corto, con duración total de una negra o fracción. En el marco de esta tesis se considera el trino como una alternancia con duración de fusa para la codificación en representación intermedia. Por otro lado, las obras exigen tener en cuenta notas de gracia. Estas son notadas con un símbolo de menor tamaño con respecto al resto de las notas. También se las codifica como fusas que roban su duración de la nota siguiente. Este último criterio de las notas de gracia, puede ser discutible desde el punto de vista de notación musical, pero para la resolución del problema aquí planteado es suficiente.

6.2. Codificación en representación intermedia

El bloque de codificación tiene como entrada notación simbólica y a la salida arroja una matriz. Ésta se denomina como RI de tamaño (n, m), que codifica en

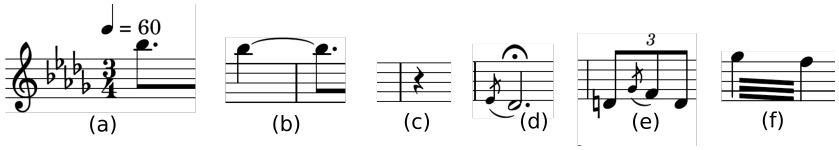


Figura 6.3: Elementos de notación simbólica considerados en el marco de la tesis. (a) Representación del tempo de la obra en BPM, se utiliza la negra para definir el valor. (b) Se observa la ligadura así como el puntillo, la primera representada como la línea curva que une las dos figuras y la segunda como el punto que sigue a la cabeza de la segunda nota. (c) Se observa un silencio de negra. (d) Arriba el símbolo correspondiente al calderón. (e) Se observa en primer lugar el símbolo para el tresillo, notado como el número 3 por arriba de las notas, además se observa un ejemplo de nota de gracia notada una corchea tachada más pequeña. (f) Símbolo para el trino.

Redonda	**Blanca**	**Negra**	**Corchea**	**Semicorchea**	**Fusa**	**Semifusa**
						
1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$

Tabla 6.1: Tabla de duraciones de las notas musicales. Se ejemplifican las utilizadas en el marco de la tesis, aunque el totalidad de las posibilidades no queda representado.

el eje vertical las alturas mientras que en el horizontal los períodos en que estas notas viven a lo largo del tiempo. Este formato la hace comparable con la salida del bloque de extracción de contenido musical, que de forma análoga codifica alturas y duraciones en los ejes vertical y horizontal respectivamente. De esta forma el dominio de la representación intermedia se convierte en el espacio donde partitura y audio pueden ser comparables.

El valor temporal de la figura, como se vio en la sección 6.1, es expresado de forma relativa a la redonda. De esta forma para obtener la duración de una nota musical se tiene en cuenta el valor de la redonda en segundos. Para el cálculo se debe tener en cuenta el tempo sugerido por el compositor, de esta forma transformar un valor relativo a absoluto. Como se muestra en la ecuación 6.1 mediante el producto entre el valor de la figura y el valor en segundos de una redonda se obtiene la duración en segundos,

$$duracion = \left(\frac{4 * 60}{tempo}\right) * figura \tag{6.1}$$

donde el tempo esta expresado en BPM (sigla por su denominación en inglés de *Beats per minute*) y la figura es el valor relativo a la redonda.

En coordinación con lo anterior, además de definir la duración y así el eje horizontal de la matriz denominada RI, se debe determinar la posición en el eje vertical en función de la altura musical notada. Para esto se tienen en cuenta dos

Nota	Midi	Hz	RICA	RIAA	Nota	Midi	Hz	RICA	RIAA	Nota	Midi	Hz	CA	RIAA
B3	59	246.9	11	0	C5	72	523.3	0	13	**D#6**	87	1244.5	3	28
C4	60	261.6	0	1	C#5	73	554.4	1	14	E6	88	1318.5	4	29
C#4	61	277.2	1	2	D5	74	587.3	2	15	F6	89	1396.9	5	30
D4	62	293.7	2	3	D#5	75	622.3	3	16	F#6	90	1480.0	6	31
D#4	63	311.1	3	4	E5	76	659.3	4	17	G6	91	1568.0	7	32
E4	64	329.6	4	5	F5	77	698.5	5	18	G#6	92	1661.2	8	33
F4	65	349.2	5	6	F#5	78	740.0	6	19	A6	93	1760.0	9	34
F#4	66	370.0	6	7	G5	79	784.0	7	20	A#6	94	1864.7	10	35
G4	67	392.0	7	8	G#5	80	830.6	8	21	B6	95	1975.5	11	36
G#4	68	415.3	8	9	A5	81	880.0	9	22	C7	96	2093.0	0	37
A4	69	440.0	9	10	A#5	82	932.3	10	23	C#7	97	2217.4	1	38
A#4	70	466.1	10	11	B5	83	987.8	11	24	D7	98	2349.3	2	39
B4	71	493.9	11	12	**C6**	84	1046.5	0	25	D#7	99	2489.0	3	40

Tabla 6.2: Tabla representativa de la correspondencia entre la nota musical expresada en notación Midi y Americana, la frecuencia fundamental en Hz y el valor correspondiente para las representación intermedia en clases de altura (RICA) y altura absoluta (RIAA). El rango de frecuencias es de B3-D#7 con 12 *bins* por octava. Se omiten C#6 y D6 por facilidad para dar formato a la tabla.

aspectos independientes, por un lado la forma de representar las alturas musicales (específicamente como altura absoluta, o clase de altura) y por otro lado la cantidad de *bins* (parámetro del sistema) que completan una octava musical. De esta forma, se tienen dos ecuaciones para el cálculo de la posición de la nota en el eje vertical de la matriz. En primer lugar para el caso de altura absoluta, la ecuación es de la siguiente forma:

$$RIAA(altura) = (midi(altura) - midi(B3)) * \left(\frac{resolucion}{12}\right) \tag{6.2}$$

donde la función midi devuelve para una nota el valor correspondiente en midi (ver tabla 6.2 donde se detallan los valores en midi relevantes) y resolución es la cantidad *bins* por octava. Por otro lado, para el cálculo de la posición de la nota en el eje vertical en el caso de clase de altura es de la siguiente forma:

$$RICA(altura) = resto[midi(altura), 12] * \left(\frac{resolucion}{12}\right) \tag{6.3}$$

donde la función resto devuelve un número entero en el intervalo $[0, 11]$, correspondiente a la clase de altura musical y la resolución, como en el caso anterior, es la cantidad *bins* por octava.

Por otro lado, para todas las notas de la partitura la intensidad es codificada con el valor unidad, siendo el máximo ya que se trabaja con valores normalizados. Esta decisión se desprende del hecho que las dinámicas no se tienen en cuenta para la codificación.

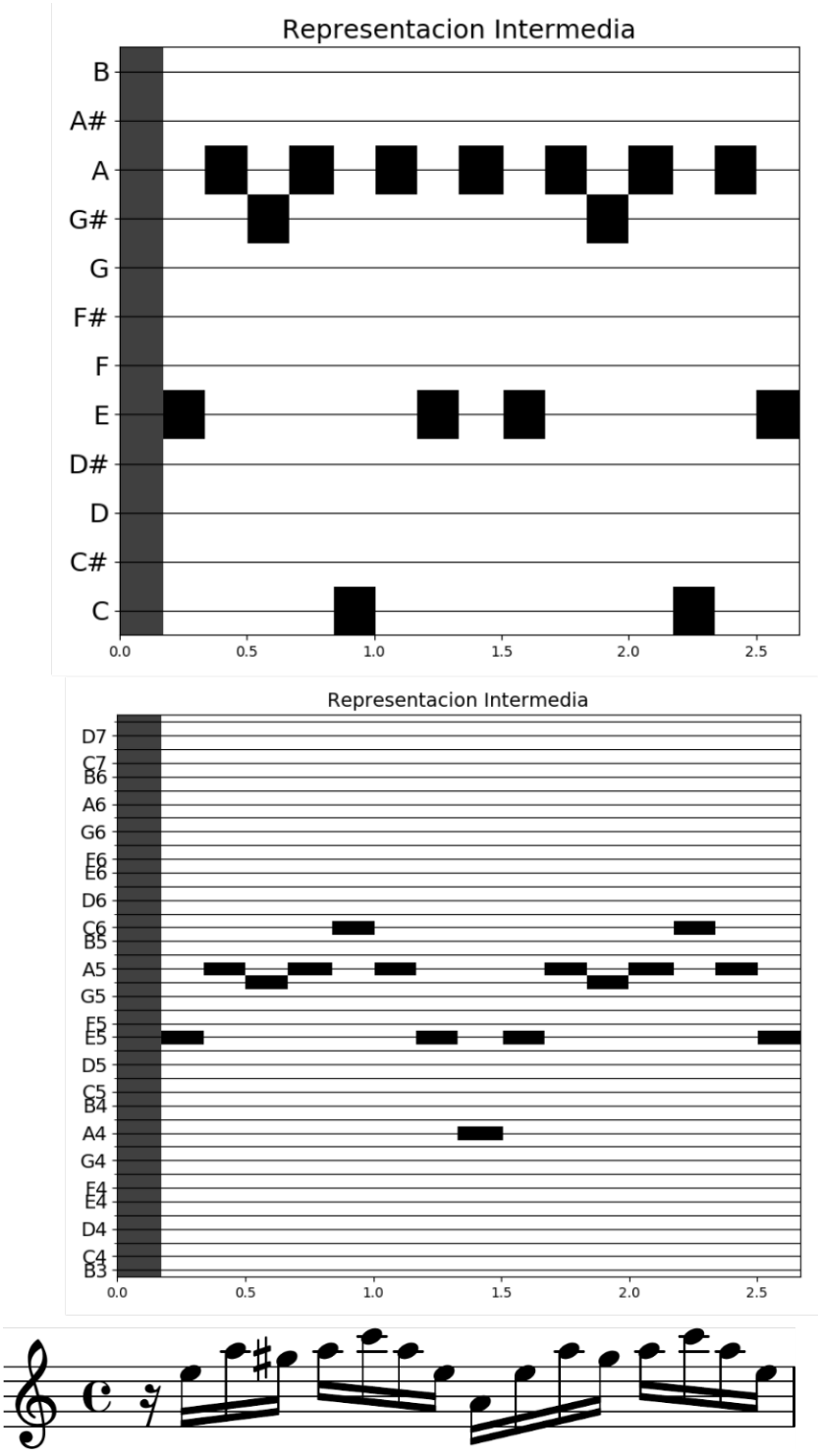


Figura 6.4: Se observa el resultado de codificación de la partitura en representación intermedia con ambas organizaciones de altura. Arriba clases de altura y abajo alturas absolutas, para el primer compás de BWV 1013 de J.S. Bach.

Capítulo 7

Experimentos

Este capítulo está dedicado a la evaluación de desempeño de la solución planteada en la base de datos de flauta traversa. Se comienza evaluando la influencia de los parámetros de representación intermedia en el resultado de alineación. Se evalúan además distintas restricciones en el camino óptimo de alineación de DTW. A continuación, se hace una comparación de la estrategia de codificación de notación simbólica propuesta aquí, frente a la síntesis como paso intermedio. Para finalizar se hace una comparación con un algoritmo desarrollado por terceros. Todos los experimentos son realizados en el presente capítulo se realizan con la base de datos construida en el marco de la tesis.

7.1. Medidas de desempeño

Para la evaluación de desempeño como se recomienda en la publicación [OLS03] se utilizan dos medidas, la tasa de aciertos y la precisión[1]. La primera cuantifica la cantidad de notas bien identificadas, como porcentaje del total. Por otro lado, la precisión es el promedio del desfasaje de las notas bien identificadas con respecto al ground truth.

A la salida de la etapa de alineación se obtiene una lista como la representada en la Figura 7.1. Ésta es comparada con el ground truth correspondiente. Siendo $u(t_u)$ la altura del resultado de la alineación (i.e. frecuencia representada en midi) y $v(t_v)$ la del ground truth, en los tiempos t_v y t_u respectivamente, se definen los aciertos como los puntos que cumplen $|t_v - t_u| < tol$, si $u(t_u) = v(t_v)$. Por otro lado, la precisión matemáticamente se define como $\frac{\sum |t_u - t_v|}{N}$ siendo N el largo de v (i.e. la cantidad de notas en el ground truth). Para los cálculos del presente capítulo se definió la tolerancia $tol = 200ms$ como se sugiere en la publicación [OS01].

[1]Definida en este caso como una medida de desfasaje temporal entre las anotaciones y el resultado de la alineación.

tiempo(s)	frecuencia (midi)	duración (s)
0.000	0.000	0.150
0.150	76.000	0.242
0.391	81.000	0.150
0.541	80.000	0.178
0.719	81.000	0.150
0.869	84.000	0.207
1.076	81.000	0.150
1.226	76.000	0.150
1.375	69.000	0.207

Figura 7.1: Ejemplo del resultado de etapa de alineación. Se observa la serie temporal representada como comienzo, altura y duración de las notas musicales.

Organización de las Alturas	Alturas Absolutas - Clases de Altura
Resolución Temporal (en milisegundos)	1.4 - 2.9 - 5.8 - 11.6 - 23.2 - 46.4
Bins por octava	12 - 24 - 36
Armónicos	0 - 1 - 2 - 3 - 4 - 5 - 6
β (Beta)	0.01 - 0.04 - 0.07 - 0.1 - 0.4 - 0.7 - 1.0
Sparsity	0 - 0.2 - 0.3 - 0.4 - 0.5 - 0.6 - 0.8 - 0.9

Tabla 7.1: Tabla con el detalle de los rangos de valores considerados para el ajuste, en cada parámetro de la solución.

7.2. Ajuste de parámetros en representación intermedia

Para comenzar, se evalúa la influencia en el desempeño del sistema de los parámetros asociados a la representación intermedia. Es decir, los bloques de codificación de la notación simbólica y de extracción de contenido musical. Algunos de los parámetros influyen en ambos bloques en simultáneo. Por el contrario, existen además parámetros propios de cada bloque que no son compartidos.

En la tabla 7.1 se especifican los valores que se utilizan para el ajuste en cada uno de los parámetros. En el caso de la resolución temporal se tuvo en cuenta que la duración de la semifusa más rápida es de aproximadamente de $50ms$ (en Sequenza con tempo de aproximadamente 70BPM). Esta duración se puede pensar como una cota superior en la resolución temporal, ya que saltos mas grandes tienen riesgo de perder notas en el análisis. De esta forma, se decide finalizar el barrido e $46{,}4ms$. En el caso de los bins por octava, el valor mínimo esta asociado a la cantidad de notas de la escala cromática. Además, se hace la evaluación del funcionamiento para 24 y 36 bins por octava. Por último la cantidad armónicos se acotó en 6 teniendo en cuenta para las notas altas del registro de la flauta, este armónico se encuentra en el límite del ancho de banda de la representación tiempo frecuencia. Por último para Sparsity y β no hubo una decisión asociada ya que se barren prácticamente todos los valores posibles.

7.2.1. Resultados

En esta sección se detallan los resultados del ajuste de parámetros. Se comienza con un grid search[2] en los parámetros de resolución temporal, cantidad de bins por octava y número de armónicos en la codificación de la partitura. Para esto, se fijan el resto de los parámetros de forma de minimizar los grados de libertad. Es así que en esta etapa, se computa la alineación con los parámetros sparsity y β en valores arbitrarios. Por otro lado, basado en la decisión conveniente de los tres iniciales (resolución temporal, bins por octava y número de armónicos) se barren los otros dos de forma independiente. Esta sección se divide en dos, en función de los resultados para la organización de sonidos en alturas absolutas y clases de altura.

En las Figuras 7.3, 7.2 y 7.4 se observan la tasa de aciertos y precisión del sistema para la organización en alturas absolutas, obtenidos con distintas combinaciones de los parámetros en etapa de ajuste. A continuación se en listan algunas observaciones que se desprenden del análisis de los resultados:

- Existe clara superioridad en el desempeño con la codificación de la partitura con un armónico.
- El aumento en la resolución temporal viene acompañado de una mejor tasa de aciertos al igual que un aumento en la precisión. El segundo, teniendo en cuenta que la distancia entre muestras disminuye, es inherente al aumento de la resolución temporal. Por contrapartida, mayor resolución provoca un aumento en el largo de las series temporales, desembocando en mayor cantidad de operaciones para el algoritmo de alineación.
- Existe una notoria superioridad cuando se emplean 12 bins por octava, en correspondencia directa a la escala cromática de la música occidental. El deterioro en el desempeño con resoluciones en frecuencia mayores se debe en parte a desvíos en la afinación. Si bien la flauta es un instrumento cromático, es decir existe una posición de llaves para cada nota de la escala, la embocadura y la velocidad del aire de soplido pueden fluctuar la afinación. Por otro lado, se observa que un mejor desempeño trae también mayor precisión.
- Para el ajuste del parámetro β se utilizan 12 bins por octava, 2.9 milisegundos de resolución temporal y codificación de la partitura con un armónico. Se observa que no tiene influencia en el total del desempeño.
- Para el ajuste del parámetro sparsity también se utilizan 12 bins por octava, 2.9 milisegundos de resolución temporal y con la codificación de la partitura con un armónico. Al igual que para el parámetro β el desempeño del algoritmo no se ve notoriamente modificado con la variación de sparsity.

[2]Término proveniente del inglés, utilizado para la estrategia de ajuste de parámetros donde se hacen los cálculos para todos los puntos de una grilla, que surge de la combinación de dos o más parámetros.

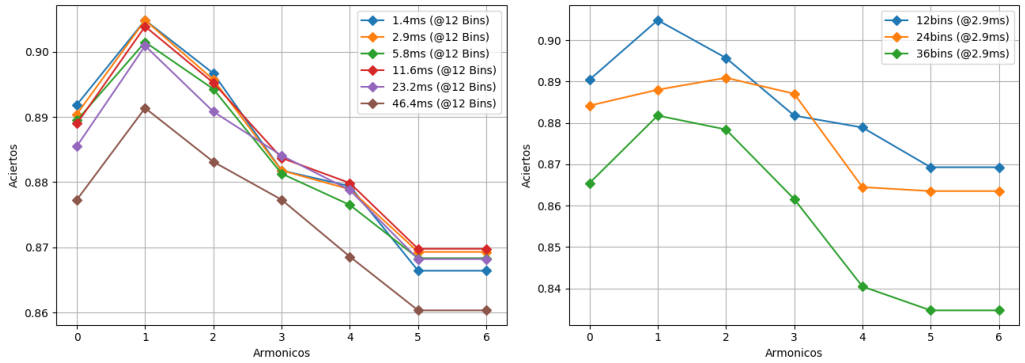


Figura 7.2: Aciertos en función de la cantidad de armónicos en la codificación, con organización en alturas absolutas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 Bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

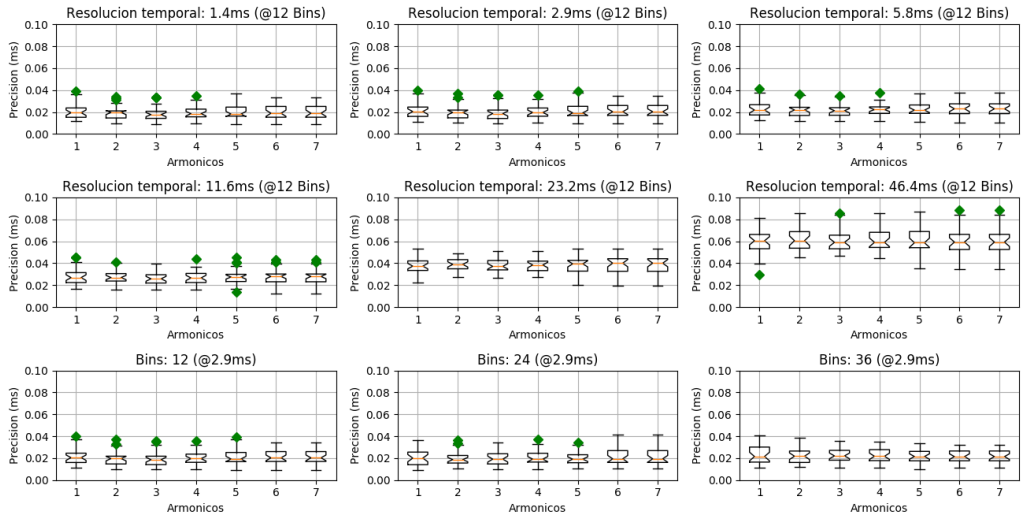


Figura 7.3: Precisión en función de la cantidad de armónicos en la codificación, con organización en alturas absolutas. Se observa en las filas 1 y 2 los resultados con resolución espectral fija en 12 Bins por octava variando la resolución temporal en cada cuadro. En la fila 3 se observa el ajuste resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

En las Figuras 7.6, 7.5 y 7.7 se observan la tasa de aciertos y precisión del sistema para la organización en clases de altura, obtenidos con distintas combinaciones de los parámetros en etapa de ajuste. A continuación se detallan algunas observaciones que se desprenden del análisis de los resultados:

- El mejor desempeño se da utilizando únicamente la fundamental en la codificación de la partitura.

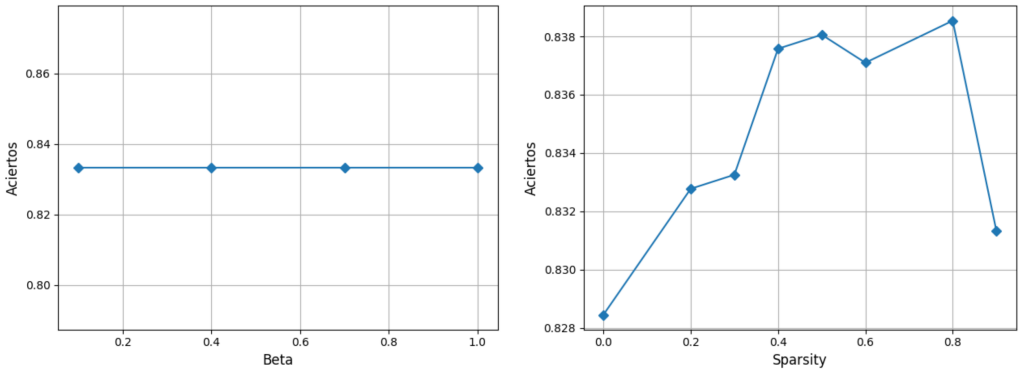


Figura 7.4: A la izquierda porcentaje de aciertos en función del parámetro β para 12 bins por octava, 2,9ms de resolución temporal y codificación de la partitura únicamente con la fundamental. A la derecha porcentaje de aciertos en función del parámetro sparsity para 12 bins por octava, 2,9ms de resolución temporal y codificación de la partitura únicamente con la fundamental.

- La resolución temporal en este caso no tiene gran influencia, ya que salvo para 46.4 milisegundos, el resto tienen desempeño similar. En cuanto a la precisión los resultados infieren lo ya dicho para alturas absolutas.
- Existe clara superioridad para 12 bins por octava al igual que en alturas absolutas.
- Para el ajuste del parámetro β se utilizan 12 bins por octava, 2.9 milisegundos de resolución temporal y únicamente la fundamental en la codificación de la partitura. El parámetro β muestra una notoria superioridad para 0.1.
- Para el ajuste del parámetro sparsity se utilizan 12 bins por octava, 2.9 milisegundos de resolución temporal y únicamente la fundamental en la codificación de la partitura. El parámetro sparsity presenta una meseta que se extiende desde 0.2 a 0.8. Para valores mayores a 0.8 se obtiene un deterioro en el rendimiento de los algoritmos, asociado al descarte de información relevante para la alineación.

En líneas generales se observa que el sistema es robusto frente a la variación en los parámetros. Por otro lado, 12 bins por octava parece ser la mejor resolución en frecuencia para ambas organizaciones de altura. Además desde el punto de vista del compromiso tiempo-frecuencia, tiene las ventanas de análisis más pequeñas siendo la resolución en frecuencia que menos compromete la estacionareidad en la extracción de contenido musical. En cuanto a la codificación de la partitura, se observa que para alturas absolutas la codificación con un armónico mientras que para clase de altura la codificación con único componente la fundamental, parecen ser las mejores opciones. Los parámetros sparsity y β no generan grandes variaciones en el desempeño, salvo para el caso de β en clases de altura donde 0.1

es de notoria superioridad. Por lo anterior se cree que $sparsity = 0{,}5$ y $\beta = 0{,}1$ son una elección razonable.

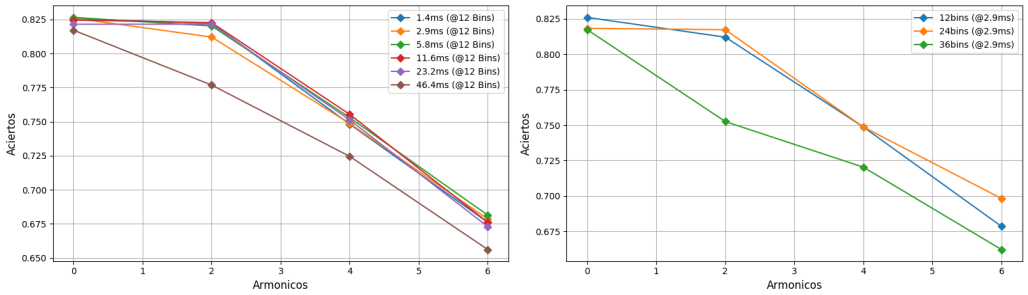


Figura 7.5: Aciertos en función de la cantidad de armónicos en la codificación, con organización en clase de alturas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 Bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

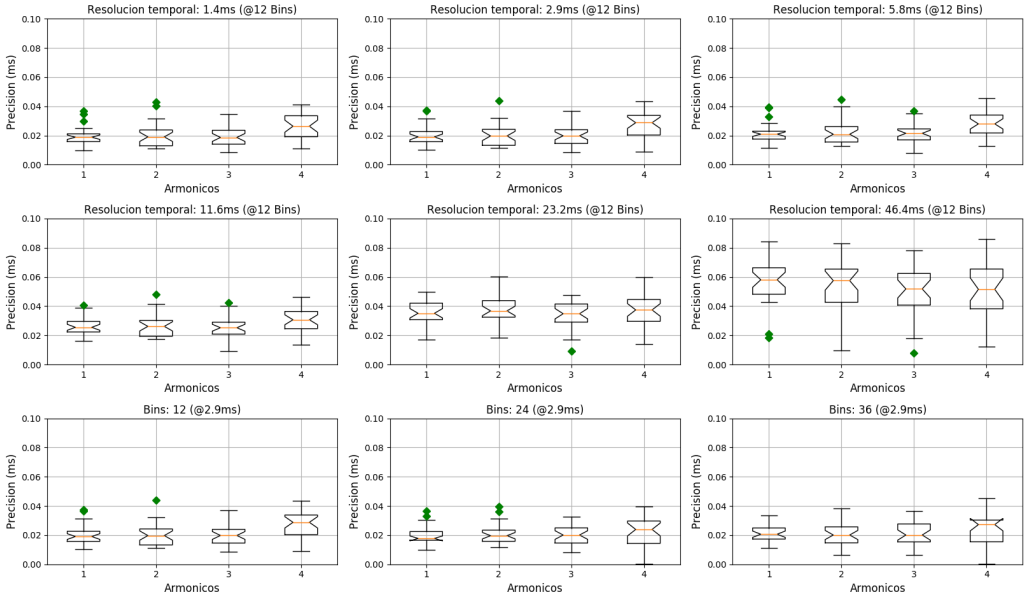


Figura 7.6: Precisión en función de la cantidad de armónicos en la codificación, con organización en clase de alturas. Se observa en las filas 1 y 2 los resultados con resolución espectral fija en 12 Bins por octava variando la resolución temporal en cada cuadro. En la fila 3 se observa el ajuste resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

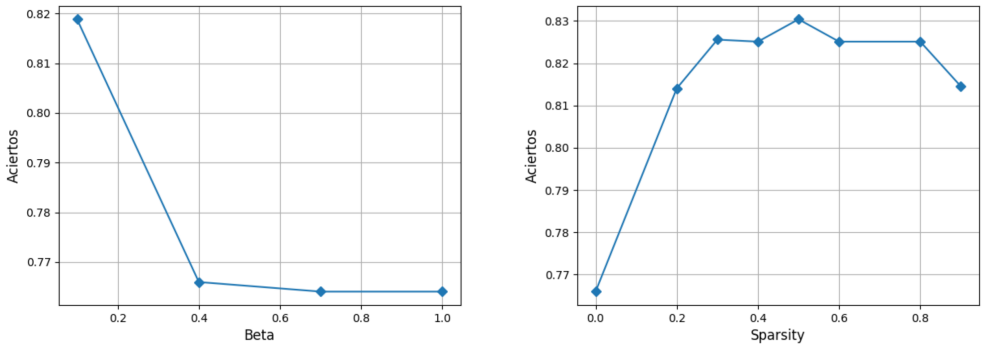


Figura 7.7: A la izquierda porcentaje de aciertos en función del parámetro β para 12 bins por octava, $2{,}9ms$ de resolución temporal y codificación de la partitura únicamente con la fundamental. A la derecha porcentaje de aciertos en función del parámetro sparsity para 12 bins por octava, $2{,}9ms$ de resolución temporal y codificación de la partitura únicamente con la fundamental.

7.3. Ajuste de parámetros en alineación

En la sección 4.2.1 se detalla sobre las restricciones que se pueden aplicar a el camino de mínimo costo para hacer más eficiente el cómputo de la alineación. En esta sección se detalla sobre la evaluación de desempeño de las diferentes estrategias anteriormente presentadas. En la Figura 7.8 se observa la tasa de aciertos en función de la tolerancia, donde *slope* refiere a pendiente y *radius* al tamaño de la ventana de ajuste (utilizando la nomenclatura de la sección 7.3). A continuación el detalle utilizando la nomenclatura de la Figura 7.8:

- **Slope P0 & No Radius:** Se utiliza la pendiente $P = 0$ y ninguna restricción en la ventana de ajuste.
- **Slope P05 & No Radius:** Se utiliza la pendiente $P = 0{,}5$ y ninguna restricción en la ventana de ajuste.
- **Slope P1 & No Radius:** Se utiliza la pendiente $P = 1$ y ninguna restricción en la ventana de ajuste.
- **Slope P2 & No Radius:** Se utiliza la pendiente $P = 2$ y ninguna restricción en la ventana de ajuste.
- **Slope P0 & Radius 50:** Se utiliza la pendiente $P = 0$ y una ventana de ajuste con radio 50.
- **Slope P0 & Radius 200:** Se utiliza la pendiente $P = 0$ y una ventana de ajuste con radio 200.
- **Slope P0 & Radius 500:** Se utiliza la pendiente $P = 0$ y una ventana de ajuste con radio 500.

Se observa que el mejor desempeño se da sin ventana de ajuste y pendiente $P = 0$. Además, se puede ver que con un radio de 500 en la ventana de ajuste para una tolerancia de $200ms$ el desempeño es similar. Teniendo en cuenta que las ventanas de ajuste determinan una disminución de la cantidad de operaciones en el cómputo de la alineación, en lo que sigue se trabaja con la estrategia **Slope P0 & Radius 500**.

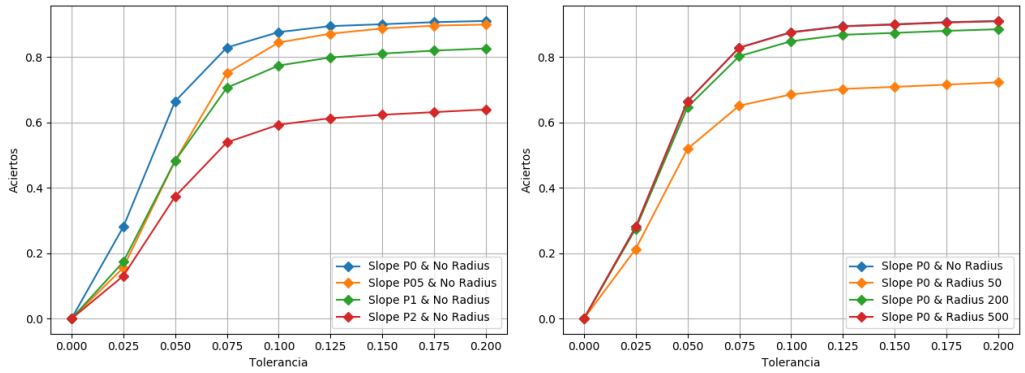


Figura 7.8: Comparación del desempeño en función de los parámetros de DTW. Se grafica tasa de aciertos contra tolerancia. A la izquierda se observa la variaciones en la pendiente, a la derecha la variación en el radio de la ventaja de ajuste.

7.4. Codificación Vs. Síntesis

En la elección de la transformación de la partitura a representación intermedia, se proponen básicamente dos alternativas en la literatura de alineación entre audio y notación simbólica. La primera, resulta en la codificación de la partitura de forma directa, en otras palabras desde notación simbólica se genera la representación intermedia sin un paso extra [OS01]. La segunda, recurre a la síntesis de audio a partir de la partitura, de esa forma mediante la extracción de contenido musical de la señal sintética se obtiene la representación intermedia [DR06].

En esta sección, se detallan los resultados de un experimento de comparación entre ambas estrategias. Para esto, se sintetiza la notación simbólica de cada fragmento utilizando el *toolbox* presentado en la publicación [RLJ09]. Posteriormente, la representación intermedia se genera utilizando el bloque de extracción de contenido musical con los mismos parámetros. Estos fueron seleccionados según lo visto en la sección 7.2.1 (i.e. 12 bins por octava, $2{,}9ms$ de resolución temporal y 1 ó 0 armónicos en la codificación según la organización de altura). Además la síntesis se hizo con distintos valores de reverberación para evaluar la influencia de esta elección. La reverberación en el *toolbox* se define con dos parámetros *delay* y *porcentaje de mezcla*. El primero se define como el tiempo en segundos, que demora en caer $60dB$ la señal (T60). El segundo el porcentaje de la señal de reverb que se

mezcla con la original (como su nombre lo indica). Se hizo la síntesis con cuatro combinaciones de estos valores (se presentan como Delay-Porcentaje de mezcla): $0-0$, $1-0{,}2$, $1-0{,}6$ y $1-1{,}0$.

En las Figuras 7.9 y 7.10 se observan los resultados la organización en alturas absolutas y clases de altura respectivamente. Se observa que el mejor desempeño se obtiene con la síntesis de valores $1-0{,}2$. Por otro lado, el caso de parámetros $0-0$ que no presenta reverberación, se lo puede pensar como una codificación de la notación simbólica similar a la que se propone en la presente tesis. Con la diferencia de ser más sofisticada en el contenido espectral (i.e. presenta armónicos que decaen en amplitud al aumentar la frecuencia y además el modelado la variación con el transcurrir temporal). Se observa en los resultados, que a pesar de los sofisticado de la síntesis, la codificación propuesta en la presente tesis obtiene mejores resultados de desempeño.

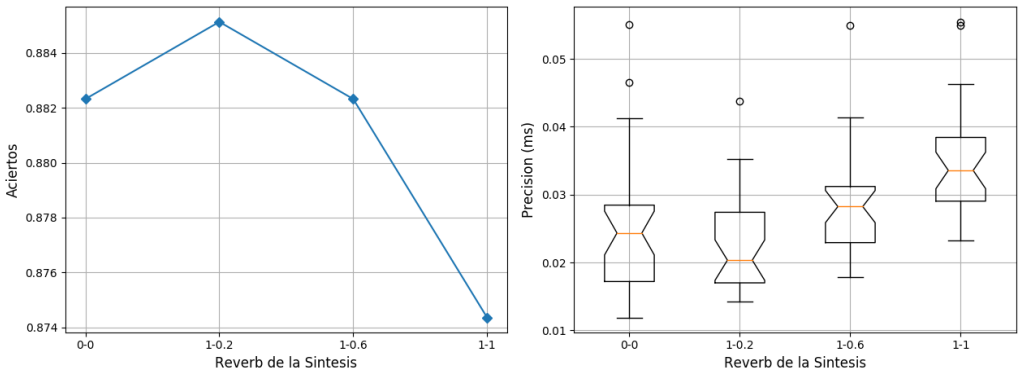


Figura 7.9: Medidas de desempeño para representación intermedia de notación simbólica realizada con síntesis y organización en alturas absolutas.

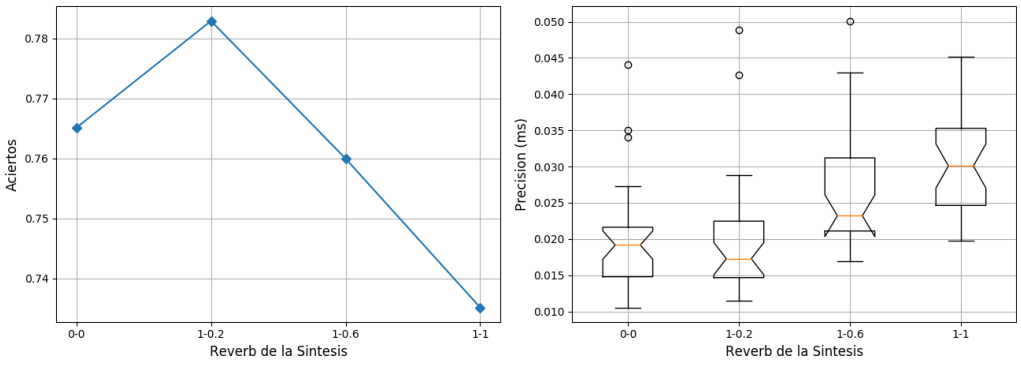


Figura 7.10: Medidas de desempeño para representación intermedia de notación simbólica realizada con síntesis y organización en clases de altura.

7.5. Evaluación de desempeño por obra

A continuación se presentan los resultados de desempeño por cada obra de forma independiente. Para esto se seleccionaron los parámetros más convenientes según lo visto en la etapa de ajuste de parámetros. Los resultados son presentados por independiente para ambas organizaciones de altura, y calculados en tres resoluciones temporales a modo comparativo. En las tablas 7.2 y 7.3 se observan la tasa de aciertos y el promedio de la precisión. A continuación se en listan algunas observaciones:

- Se observa que el desempeño es muy similar para ambas organizaciones de altura, siendo ambas estrategias aceptables para la resolución del problema. En cuanto a la utilización de organización en clases de altura tiene como punto fuerte que en etapa de alineación la dimensión del vector es menor, en contrapartida requiere la operación extra necesaria para colapsar la CQT en Chromagrama.

- Por otro lado, se observa que el parámetro de resolución temporal no tiene mayor influencia sobre la tasa de aciertos, por el contrario si se observan diferencias en la precisión.

- Vale resaltar que si bien los resultados frente a la variación de armónicos en la codificación de la notación simbólica son contundentes (i.e. ningún armónico para organización en clases de altura y solamente el primer armónico para alturas absolutas), estos resultados pueden variar si se ajustara la codificación con un decaimiento en la amplitud de los armónicos emulando de esa forma la naturaleza sonora de la flauta. A pesar de esto, en el experimento de comparación entre síntesis y codificación, el desempeño fue superior con el segundo.

- Dado que la flauta es un instrumento cromático se podría suponer previo a los experimentos que el parámetro de bins por octava que mejor se ajustaría a la solución del problema es el de 12. Esto se corrobora con los experimentos presentados en la tesis. La división del espectro en 12 bandas por octava es robusta frente a desviaciones en la afinación, por lo que resulta en la de mejor desempeño. Como contrapartida, se pierde potencialmente la capacidad de identificar ejecuciones de vibratos, glissandos, entre otras técnicas, que según la aplicación podrían ser de utilidad.

Por otro lado en la Figura 7.11 se observa un gráfico de barras para la tasa de aciertos con organización en alturas absolutas. Existe clara superioridad en el movimiento Allemande de BWV 1013 de J.S Bach. Esto es razonable si se tiene en cuenta que es la obra más simple en estructura rítmica. Además, de forma anecdótica sucede que estas medidas se deterioran progresivamente con el momento histórico de cada obra (en orden cronológico: Allemande - Syrinx - Density - Sequenza). La peor tasa de aciertos se da para Sequenza I de L. Berio en correspondencia con su complejidad rítmica.

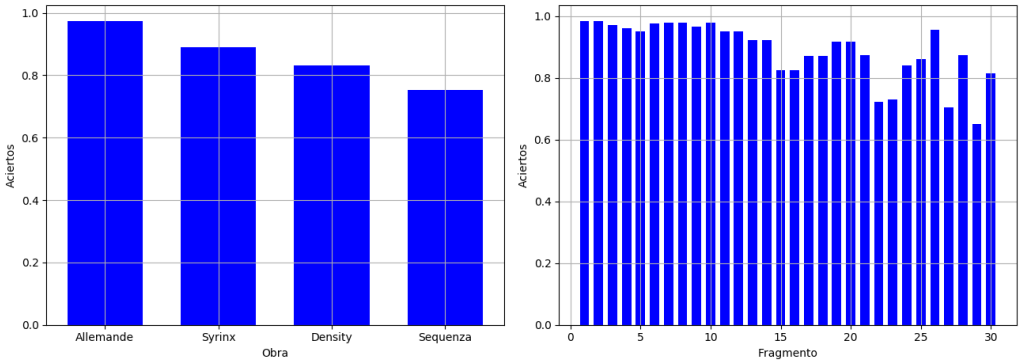


Figura 7.11: Tasa de aciertos para organización en alturas absolutas con 12 bins por octava, $2{,}9ms$ de resolución temporal y 1 armónico en la codificación de la partitura. A la izquierda se observa la tasa de aciertos por obra y a la derecha por fragmento (los primeros 10 corresponden a Allemande, los siguientes 10 a Syrinx, los siguientes 6 a Density y los últimos 4 a Sequenza)

	Allemande	Syrinx	Density	Sequenza
2.9ms	97 %/16ms	86 %/16ms	82 %/20ms	74 %/26ms
5.8ms	96 %/15ms	86 %/17ms	81 %/19ms	70 %/28ms
11.6ms	96 %/16ms	85 %/18ms	81 %/21ms	69 %/28ms

Tabla 7.2: Tabla con tasa de aciertos y promedio de la precisión, para representación intermedia en alturas absolutas. Se detalla las tres mejores combinaciones de parámetros obtenidas en etapa de ajuste. Las resoluciones temporales se detallan en la primer columna, el resto de los parámetros son 1 Armónico, 12 Bins. $\beta = 0.1$ y Sparsity $= 0.3$

7.6. Ajuste de parámetros en Alignmidi de Dan Ellis

En esta sección se detalla el ajuste de parámetros con el algoritmo de alineación entre audio y partitura desarrollado por Dan Ellis denominado como Alignmidi[3]. El objetivo final es el de comparar el algoritmo desarrollado en la presente tesis con otro implementado por terceros de forma objetiva. Teniendo en cuenta además

[3]D. P. W. Ellis (2014). 'Aligning MIDI files to music audio', web resource. http://www.ee.columbia.edu/ dpwe/resources/matlab/alignmidi/

	Allemande	Syrinx	Density	Sequenza
2.9ms	96 %/15ms	86 %/13ms	80 %/20ms	70 %/29ms
5.8ms	96 %/16ms	86 %/15ms	81 %/20ms	70 %/30ms
11.6ms	96 %/18ms	85 %/17ms	77 %/21ms	68 %/30ms

Tabla 7.3: Tabla con tasa de aciertos y promedio de la precisión, para representación intermedia en clase de alturas. Se detalla las tres mejores combinaciones de parámetros obtenidas en etapa de ajuste. Las resoluciones temporales se detallan en la primer columna, el resto de los parámetros son 1 Armónico, 12 Bins. $\beta = 0.1$ y Sparsity $= 0.3$

que la base de datos fue compilada en el marco de la presente tesis.

El algoritmo de Ellis presenta dos métodos de programación dinámica para encontrar el camino óptimo en la matriz de similaridad, por un lado DTW y por otro Viterbi. El mejor desempeño, sobre la base de datos de flauta traversa, se obtiene con el primero. Por lo que todos los resultados que se muestran a continuación son con DTW. Además, para este método de alineación se pueden variar dos parámetros adicionales. En primer lugar el desfasaje permitido entre comienzos y finales de audio y partitura (llamado como *Gulley* en el *toolbox*). Teniendo en cuenta que en el caso de la base de datos de flauta todos los fragmentos y sus partituras comienzan y terminan en lugares correspondientes, se define como cero. En segundo lugar, se puede variar el peso de los pasos horizontales y verticales en el camino óptimo de alineación (llamado como *Horizwt* en el *toolbox*). En la sección 4.2 fueron definidos cómo: w_h y w_v. En la implementación de Ellis, se varían por igual siendo $w_h = w_v$ en todo momento.

A continuación se presenta la tasa de aciertos en función del parámetro *horizwt* en etapa de ajuste. Se observa entonces, en la Figura 7.12 los resultados para cada caso. Es clara la superioridad cuando todas las direcciones tienen el mismo peso, matemáticamente: $w_h = w_v = w_d = 1$.

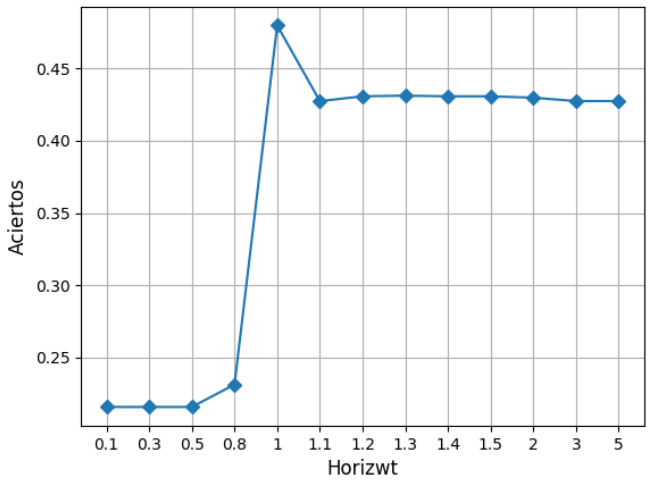


Figura 7.12: Tasa de aciertos en función del parámetro *horizwt*.

7.7. Comparación de todas las estrategias

El objetivo de la presente sección es realizar una comparación de todos los métodos presentados en el marco de este trabajo. Se detalla la tasa de aciertos en función de la tolerancia (fue definida en la sección 7.1). Para esto se hace un

barrido de $0ms$ a $200ms$. Por otro lado, la precisión detallada corresponde en todos los casos a $tol = 200ms$.

Vale resaltar que se presentan las estrategias que fueron detalladas a lo largo del presente capítulo, y además se agrega como variante, el cálculo de la matriz de similaridad con distancia euclidiana (por más detalle ir a la sección **??**). Para esto se computa la representación intermedia en organización en alturas absolutas con los mejores parámetros de etapa de ajuste (12 bins por octava, 2,9ms de resolución temporal y un armónico en la codificación de la partitura). En resumen se hace la comparación de:

- **AA & Cosine**: Organización en alturas absolutas con los mejores parámetros de la etapa de ajuste (12 bins por octava, 2,9ms de resolución temporal y un armónico en la codificación de la partitura) y distancia coseno para el cómputo de la matriz de similaridad.
- **AA & Euclidean**: Organización en alturas absolutas con los mejores parámetros de la etapa de ajuste (12 bins por octava, 2,9ms de resolución temporal y un armónico en la codificación de la partitura) y distancia euclidiana para el cómputo de la matriz de similaridad.
- **CA & Cosine**: Organización en clases de altura con los mejores parámetros de la etapa de ajuste (12 bins por octava, 2,9ms de resolución temporal y un armónico en la codificación de la partitura) y distancia coseno para el cómputo de la matriz de similaridad.
- **D. Ellis Alignmidi**: El algoritmo implementado por Dan Ellis detallado en la sección 7.6.
- **AA & Síntesis**: Organización en alturas absolutas y representación intermedia de la notación simbólica realizada mediante la síntesis. Se utilizan los mejores parámetros más detalle en la sección 7.4.
- **CA & Síntesis**: Organización en clases de altura y representación intermedia de la notación simbólica realizada mediante la síntesis. Se utilizan los mejores parámetros más detalle en la sección 7.4.

Es claro que el mejor desempeño está dado por las estrategias *AA & Cosine* y *CA & Cosine*. La siguen las estrategias basadas en la síntesis para representación intermedia de la notación simbólica. Se ve que al cambiar la distancia para el cómputo de la matriz de similaridad de distancia coseno a euclidiana hay un deterioro del desempeño. Por último el desempeño obtenido con Alignmidi de D. Ellis fue el más bajo.

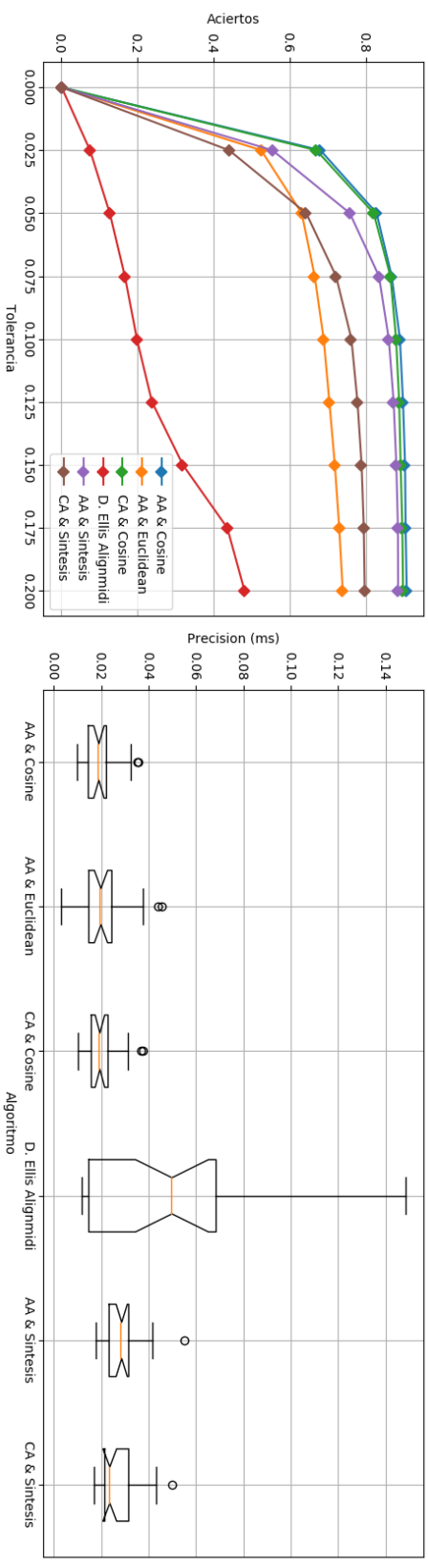


Figura 7.13: A la izquierda la tasa de aciertos en función de la tolerancia. A la derecha la precisión en forma de *boxplot* para el caso $tol = 200ms$. Aclaración: AA refiere a Alturas Absolutas y CA a Clases de Altura.

Capítulo 8

Caso de estudio en la flauta contemporánea: Aliento/Arrugas

Muchos compositores académicos contemporáneos, se han dedicado a explorar las capacidades sónicas de los instrumentos. Tal es el caso de la flauta traversa que hoy en día cuenta con técnicas modernas, claramente definidas y reproducibles a las que se denominan extendidas. Diversos recursos técnicos forman parte del repertorio actual de la flauta y los intérpretes deben expandir sus habilidades con el instrumento. El repertorio contemporáneo define un área del *MIR* (por su sigla en inglés *Music Information Retrieval*) desafiante y de espectro más amplio, donde existe tierra fértil para explorar y lograr contribuciones.

Para la alineación entre audio y partitura en este tipo de obras no es suficiente con la representación basada en alturas y duraciones, a diferencia de lo que sucede con el lenguaje tradicional de la flauta. Se vuelve necesario entonces explorar otras representaciones intermedias de las que fueron planteadas en capítulos anteriores (i.e. CQT y Chromagrama) para la resolución del problema de alineación entre audio y partitura en obras del repertorio contemporáneo.

El cometido de la presente sección es la evaluación de diferentes representaciones de audio de uso extendido, para el material sonoro del repertorio contemporáneo. Se plantea con este fin el objetivo de la extracción automática de embocadura a partir de grabaciones de la obra contemporánea Aliento/Arrugas como caso de estudio. En lo que sigue, Sección 8.0.1, se define con más detalle el término embocadura, y se hace un análisis de sus principales implicaciones en el sonido de la flauta. Además en la Sección 8.0.2 se describe en mayor profundidad la pieza musical Aliento/Arrugas. Luego en la Sección 8.1 se define el problema y la estrategia de resolución, se detalla la base de datos y se da un breve marco teórico de las características a utilizar. La Sección 8.2 esta dedicada a comentar los resultados y algunas consideraciones, para finalmente concluir con la exploración en el Sección 8.3.

8.0.1. Embocadura

El término embocadura refiere al aparato de producción de la excitación de la columna de aire, en conjunto a la técnica de soplido [Pis55]. Por ejemplo en el caso particular de la flauta ejecutada con técnica tradicional, los labios dirigen el flujo de aire directamente al bisel en el hueco del instrumento. De esta forma la turbulencia producida por la colisión, genera una excitación periódica en la columna de aire que provoca la resonancia del instrumento y un sonido tonal.

La embocadura es un elemento determinante del material sonoro ejecutado, siendo perceptible de forma auditiva a través de variaciones en la dinámica, altura y timbre [Dic75]. Las características sonoras quedan determinadas por los siguientes parámetros físicos de la ejecución del instrumento:

- **Ángulo de la flauta:** Por un lado afecta la altura de la ejecución. Al girar hacia el intérprete la altura baja, por el contrario sube al girar en el otro sentido. Por otro lado genera cambios en el timbre del material sonoro. Al girar hacia afuera (sentido opuesto al intérprete) mas allá del ángulo normal de ejecución, el sonido se vuelve primero más brillante y luego aumenta la prominencia del componente de ruido, en inglés se lo define como *Breathy*, se lo puede traducir al español como *Respirado*. Sin embargo al girar hacia el intérprete aumenta la energía de los parciales altos y disminuye la fundamental generando un sonido que se puede definir metafóricamente como *Filoso* o *Edgy* por su denominación en inglés.

- **Apertura de los labios:** La apertura de los labios determina la dispersión del flujo de aire. Aperturas pequeñas producen flujos puntuales, disminuyendo la dinámica y clarificando el sonido. Del otro lado aperturas mayores aumentan la intensidad y la naturaleza ruidosa.

- **Posición de los labios:** Una posición correcta de los labios genera que la embocadura tenga gran control del sonido. Si bien la posición de los labios, y los movimientos de los mismos en la ejecución es un aspecto personal del ejecutante, existen dos tipos básicos. Alturas bajas y/o dinámicas intensas aumentan con el movimiento de los bordes de los labios hacia afuera generando casi una sonrisa en el interprete. En la segunda posición de los labios, los bordes se mueven hacia abajo en vez de hacia afuera, teniendo un efecto similar al mencionado anteriormente.

- **Presión de aire:** La presión de aire es controlada por el diafragma. Determina el nivel dinámico de la ejecución. La intensidad del aire es proporcional a la intensidad de la ejecución. Además afecta la altura del material sonoro, presiones de aire altas tienden a elevar la nota, mientras que presiones menores la disminuyen.

8.0.2. Aliento/Arrugas de Marcelo Toledo

Aliento/Arrugas es una obra para flauta traversa solista, compuesta por el argentino Marcelo Toledo. Incluye una cantidad de sonoridades exóticas mediante la ejecución del instrumento a través de técnicas extendidas. Según el compositor la intención detrás es la exploración sonora del instrumento utilizando la respiración del intérprete como elemento de expresión orgánica [Str05].

El compositor utiliza como recurso expresivo tres tipos de embocadura para ejecución del instrumento. Se diferencian por cambios en la posición de los labios y el ángulo de la flauta (ver 8.0.1), en otras palabras el ángulo entre el flujo de aire frente al bisel de la embocadura. Se en lista a continuación los nombres, manteniendo su denominación en Inglés (idioma utilizado en la partitura de la obra). Además en la Figura 8.1 se observa la notación utilizada por el compositor en la partitura de Aliento/Arrugas.

- *Normal Embouchure*: Embocadura clásica de la flauta, donde el flujo de aire frente al bisel de la embocadura genera la excitación con pulsos periódicos de la columna de aire.
- *Blow Hole Covert*: El flujo de aire ingresa directo al tubo de la flauta, sin generar turbulencia contra el bisel de la embocadura. Los labios cubren el agüjero del instrumento.
- *Breathy Embouchure*: La flauta se encuentra rotada hacia el lado contrario del intérprete, tomando como referencia la embocadura normal. Genera sonidos con orientación tonal pero con un gran componente ruidoso.

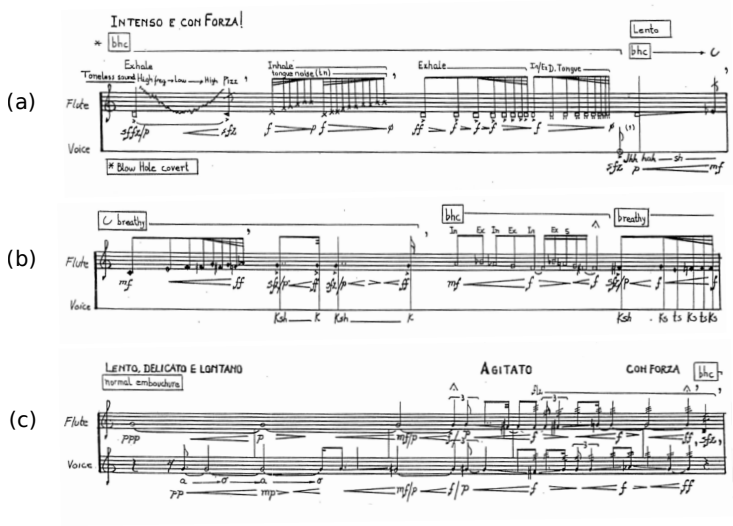


Figura 8.1: Notación de las embocaduras se observa en la parte superior de los sistemas. (a) *Blow Hole Covert*. (b) *Breathy Embouchure*. (c) *Normal Embouchure*. Fragmentos extraídos de la partitura de Aliento/Arrugas.

8.1. Definición del Problema

Teniendo en cuenta que la embocadura es un elemento determinante del material sonoro ejecutado perceptible de forma auditiva (Sección 8.0.1), se propone la extracción automática del tipo de embocadura a través del análisis computacional de grabaciones de la obra.

8.1.1. Estrategia de resolución

Se propone la resolución del problema con un enfoque de *Machine Learning*. Para esto, se procesa el audio como un *Bag of Frames* a partir del computo de descriptores numéricos, siendo estos insumo para el entrenamiento de clasificadores con estrategia de validación cruzada. Los clasificadores, siendo *Support Vector Machines*, *Random Forest* y *K-Nearest Neighbors*, se utilizan con parámetros por defecto. Debido a que el principal desafío y cometido del caso de estudio es encontrar los descriptores que extraigan las diferencias en la naturaleza sonora y permitan la separación de las embocaduras en el espacio de características. Siendo de esta forma una buena representación matemática del material sonoro con potencial uso en algoritmos de alineación entre audio y partitura.

8.1.2. Conjunto de Datos

Se cuenta con 5 grabaciones de diferentes intérpretes de la obra Aliento/Arrugas. Los intérpretes son: Pablo Somma, Emma Resmini, Claire Chase, Juan Pablo Quin-

teros y Ulla Suokko. Los archivos de audio se etiquetaron utilizando el software *Sonic Visualiser* [CLS10] dividiendo los archivos de audio en 5 clases:

- Silencio.
- Silencio con respiración del intérprete.
- Sonido generado con *Blow Hole Covert*.
- Sonido generado con *Breathy Embouchure*.
- Sonido generado con *Normal Embouchure*.

Las grabaciones de Claire Chase y Juan Pablo Quinteros que se obtuvieron para el presente trabajo sufrieron un proceso de compresión con pérdida, por lo que estos datos reciben un tratamiento distinto. No se utilizan para entrenar los algoritmos de clasificación, solo se utilizan como datos de test. Por lo que folds son de la siguiente forma:

- Cuando la grabación de test es la de Ulla Suokko, Pablo Somma o Emma Resmini, se entrena con las otras dos restantes. Metodología de *Leave One Out* por su denominación en Inglés.
- Por otro lado cuando la grabación de test es de Claire Chase o Juan Pablo Quinteros, el conjunto de entrenamiento esta compuesto por las tres grabaciones sin pérdida (i.e. la de Ulla Suokko, Pablo Somma y Emma Resmini).

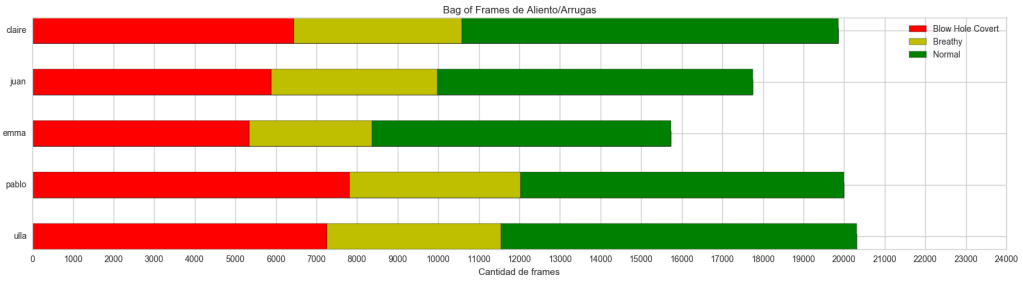


Figura 8.2: Detalle de la composición del *bag of frames* de embocaduras.

En la Figura 8.2 se observa en detalle las cantidades de frames de embocaduras en la base de datos. Se puede observar según el intérprete la proporción de las clases. Se observa que la clase mayoritaria es la *Normal embouchure* de numero comparable *Blow Hole Covert* y sensiblemente menor es la dada por la embocadura *Breathy*.

En lo que sigue se utilizan únicamente las clases asociadas a cada una de las embocaduras. Queda por fuera del alcance de este trabajo, una etapa de preprocesamiento para la segmentación del audio en fragmentos de actividad de la flauta y silencios (este problema se conocido como *Activity Detection* por su denominación en Inglés).

8.1.3. Extracción de características

Se en listan a continuación las características que se evalúan en la extracción automática de embocadura. Además se describe brevemente sus principales atributos.

Mel-Frequency Cepstral Coefficients (MFCC)

Los *Coeficientes Cepstrales de Frecuencia-Mel* fueron introducidos por [DM80] en la resolución del problema de reconocimiento del hablante a partir de señales de voz (*Speaker Recognition* su denominación en Inglés). Estos coeficientes como características de un sistema de reconocimiento automático del hablante han demostrado tener de los mejores desempeños [Qua02, Capítulo 14]. A partir de ahí han sido utilizados en diversas problemáticas de clasificación que no involucran señales de voz hablada, con buenos resultados también como es el caso de reconocimiento de instrumentos [KD07, Capítulo 6]. Su fortaleza radica en la incorporación del modelado psicoacústico de la audición humana mediante un banco de filtros basados en la escala Mel [SVN37] y la decorrelación que presentan los datos en el dominio de las *quefrencys*, dado por la aplicación de la Transformada Coseno. Son un buen descriptor para la extracción de aspectos tímbricos de la señal.

El cómputo de estas características cuenta con las etapas que se en listan a continuación de manera conceptual:

1. División de la señal en fragmentos mediante enventanado.
2. Cálculo de la magnitud de la Transformada discreta de Fourier de tiempo corto (STFT).
3. Filtrado de la señal con banco de filtros Mel.
4. Cálculo de la energía para cada filtro del banco.
5. Logaritmo de las energías.
6. Transformada Coseno de los valores a la salida del Logaritmo.
7. *Liftrado* de la señal resultante en el dominio de las *quefrencys*, luego de la Transformada Coseno. Determina la cantidad de coeficientes, o en otras palabras la dimensión del espacio de características.

El cálculo de los coeficientes MFCC tiene los siguientes parámetros determinantes de su desempeño: En primer lugar el largo de las ventanas, que define el compromiso entre resolución temporal y espectral. En segundo lugar la cantidad de filtros del banco de filtros Mel, que se puede pensar como un submuestreo de la resolución espectral ya determinada por el largo del enventanado. Por último el *liftrado* de la señal a la salida de la Transformada Coseno que determina la cantidad de coeficientes efectivos previos al clasificador.

Linear Prediction Coefficients (LPC)

La técnica de análisis de señales de tiempo discreto por predicción lineal, tiene su aplicación en diversas áreas del conocimiento. Es parte de un problema más general denominado *identificación de sistemas* desarrollado en el área de control para el análisis de sistemas dinámicos. Supone que la señal de análisis es la salida $s[n]$ de un sistema lineal con entrada $u[n]$. Su fortaleza y versatilidad radica en la estimación de parámetros del sistema lineal que define el problema.

Su enunciado más general modela la señal de análisis como un proceso *Auto-Regresivo de Media Móvil (ARMA)* [Mak75]. En otras palabras, supone que la muestra actual de la señal de análisis puede ser expresada como una combinación lineal de las muestras pasadas de la salida, y la muestra actual y pasadas de la entrada:

$$s[n] = -\sum_{k=1}^{p} a_k s[n-k] + G\sum_{l=0}^{p} b_l u[n-l] \tag{8.1}$$

Ha sido utilizado para la resolución de problemas con señales de audio, en particular existe mucha literatura al respecto con la voz humana. Para voz hablada es de los métodos más poderosos, con diversas aplicaciones. La importancia de este método se basa tanto en la precisión de la estimación de parámetros del modelo de mecanismo de producción de voz, como en su relativo bajo costo computacional [RS78, Capítulos 3 y 9].

Alineado con la utilización de LPC en problemas de señales de voz, se supone que es suficiente un modelo todo-polos para la extracción de características en el presente trabajo. De forma matemática a partir de la Ecuación 8.1 se escribe como $b_l = 0$ con $l = 1...p$. Los parámetros relevantes en el computo del descriptor son entonces, en primer lugar p asociado a la cantidad de polos del modelo AR y por otro lado el largo de las ventanas de análisis. Componiendo el vector de características por los coeficientes a_k con $k = 1...p$ (ver Ecuación 8.1).

Conjunto de características Espectrales y Armónicas

Se genera un vector compuesto por 5 características acústicas uni-dimensionales, para evaluación del desempeño en la extracción de embocadura. Entre los descriptores se optaron por 4 medidas espectrales y una medida de armónica de la señal de análisis, según la taxonomía de *features* acústicos propuesta en el libro de [KD07]. Las características son:

- Voicing: Es una medida de periodicidad de la señal. Es el *feature* armónico del conjunto. Generalmente el Voicing se encuentra embebido en los algoritmos de extracción de pitch. En particular para el presente trabajo se computa como en la referencia: [DCK02].

- Zero-Crossing Rate: Mide la cantidad de cruces por cero de la señal. Si bien es calculado en el dominio del tiempo, es una medida del contenido de alta frecuencia.
- Roll-off: Es el valor de frecuencia para el que la energía espectral acumulada supera una fracción denominada λ. En general λ se elige 95 % o 85 %.
- Centroid: Es el promedio en los bins de frecuencia ponderado por los valores de magnitud del espectro. Se puede pensar como el centro de masa en el espectro.
- Bandwidth: Es una medida de la dispersión espectral con respecto al centroide.

En todas las medidas acústicas recién mencionada es de relevancia la elección del largo de la ventana análisis, que define el compromiso entre estacionareidad de la señal y resolución en frecuencia.

Octave-based Spectral Contrast (SC)

El *Contraste Espectral por Octavas* fue desarrollado en el trabajo publicado por [JLZ+02]. Tiene como cometido ser una medida de las características relativas del espectro de la señal de análisis. Extrae la diferencia entre la prominencia de los picos en el espectro y los valles en cada octava de análisis por separado. Ha tenido buenos resultados en el problema de clasificación de estilo musical.

El cómputo de estas características tiene las siguientes etapas, que se enuncian de forma conceptual:

1. División de la señal en fragmentos mediante enventanado.
2. Cálculo de la magnitud de la Transformada discreta de Fourier de tiempo corto (STFT).
3. Filtrado de la señal con banco de filtros por octava.
4. Cálculo de la diferencia entre la energía en un entorno de los picos y de los valles en cada una de las octavas.
5. Logaritmo de las diferencias del paso anterior.
6. Transformada *Karhunen-Loeve* para representación de las características en base ortonormal y decorrelación entre las dimensiones.

A diferencia de *MFCC* y *LPC* que realizan un promediado de la información espectral, estos descriptores extraen la información relativa, mediante la comparación de picos y valles por octava. Los parámetros relevantes son en primer lugar el largo de la ventana de análisis, el entorno de los picos y valles denominado α en la literatura, y por último el número de octavas.

8.2. Experimentos

Se evalúa la capacidad de los descriptores presentados en la Sección 8.1.3 en la separación de embocaduras. Para esto se utilizan tres clasificadores distintos para minimizar el bías que pueda existir entre los datos y un algoritmo en particular. Se trabaja con los algoritmos: *Random Forest (trees=10)*, *Support Vector Machine (kernel lineal)* y *K-Nearest Neighbors (k=10)*. En todos los casos se utilizan los parámetros por defecto ya que no es objetivo de este trabajo encontrar los valores óptimos de clasificación. La implementación se realiza mediante el módulo de *Python* llamado *Scikit Learn* [PVG+11]. En todos los casos los datos son preprocesados de manera de centrar en cero y escalar la varianza a uno, previo al clasificador.

Todos los experimentos se realizan con *5-fold cross validation* donde los folds son las diferentes interpretaciones de la pieza musical, como se detalla en la Sección 8.1.2. De esta forma se asegura que frames provenientes de la misma grabación no sean usados para train y test en un mismo experimento. Además los *features*: *MFCC*, *SC*, *Roll-off*, *Centroid*, *ZCR* y *Bandwith* se calculan utilizando el módulo de *Python* llamado *Librosa* [MRL+15].

8.2.1. Primer experimento: Mejor descriptor para extracción de embocadura

El propósito es cuantificar el poder de separación de las características y determinar cual tiene el mejor desempeño. Para tener una noción general del comportamiento de los descriptores, en todos los casos se utiliza más de una combinación de parámetros. Se eligieron de forma que sea suficiente para descartar los de menor desempeño. A continuación se en listan los parámetros utilizados en cada caso.

- Características Espectrales y Armónicas: Se varían los largos de ventana y saltos de la siguiente forma (se detallan respectivamente): (a) $11ms$ y 50 % de salto (256-128 muestras), (b) $23ms$ y 50 % (1024-512 muestras) y por último (c) $46ms$ y 50 % (2048-1024 muestras). En todos los casos anteriores se utilizó:
  - Voicing: Número de retardos: 250 muestras.
  - Roll-off: $\lambda = 85\,\%$
  - Centroid, Bandwith y Zero-Crossing Rate: quedan definidos por el largo de la ventana de análisis y el salto.
- MFCC: Se computan con ventana de análisis de $23ms$ y salto del 50 %, 40 bandas Mel y se liftra la señal para obtener: (a) 20 coeficientes, (b) 30 coeficientes y (c) 40 coeficientes.
- LPC: Se computan con ventana de análisis de $23ms$ y salto del 50 % y numero de polos: (a) 10, (b) 20 y (c) 40.

- SC: Se computan con ventana de análisis de $23ms$ y salto del $50\,\%$ y numero de bandas: (a) 3 y (b) 6.

8.2.2. Resultados

En lo que sigue se muestra el resultado del desempeño de los descriptores detallados en la Sección 8.1.3 para la extracción del tipo de embocadura.

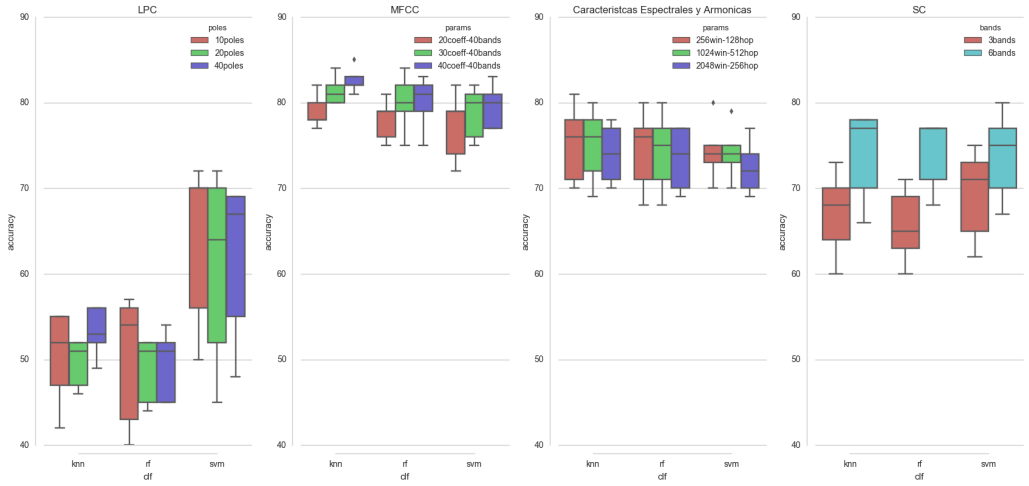


Figura 8.3: Boxplot para el accuracy de los algoritmos de clasificación. Se muestra los resultados de forma independiente por algoritmo de clasificación y según los parámetros de las características.

Se observa en la Figura 8.3 el comportamiento de las características en la separación de las clases del problema. Como medida de desempeño se utiliza la razón entre los frames bien clasificados y el total, denominada en Inglés como *accuracy*. Además en el eje horizontal se detalla el algoritmo de clasificación utilizado y en colores las distintas combinaciones de parámetros en la extracción de características.

En la Figura 8.3 se puede observar claramente que *LPC* es el que presenta peores resultados. Además de forma cualitativa se puede decir que la variación del número de polos del modelo, no afecta considerablemente el desempeño. Como resultado anecdótico a los fines del presente trabajo, se tiene que existe una mejora sustancial en el rendimiento de estos *features* con *SVM*.

Por otra parte el *feature SC* es de desempeño intermedio junto con *Características Espectrales y Armónicas* en el experimento. Para el caso de *SC* existe una mejora notoria al variar los parámetros. En otras palabras la comparación entre picos y valles tiene un poder descriptivo mayor de el problema, al dividir el espectro en 6 bandas con respecto a 3 bandas.

Del otro lado vemos que la variación de la ventana de análisis en *Características Espectrales y Armónicas* no generan un mejor rendimiento. Vale decir que su desempeño similar al de *SC* pesar de tener dimensiones correlacionadas como es el caso de *ZCR* y *Centroid*. Por lo que alguna estrategia de decorrelación previo al clasificador podría mejorar el desempeño de este conjunto de medidas.

También se observa que *MFCC* es el *feature* de mejor desempeño para la resolución del problema. Como contra partida, frente a *SC* y *Características Espectrales y Armónicas* la dimensión del espacio de características es mayor, resultando en un costo computacional superior.

Otro análisis relevante del experimento esta dado por las matrices de confusión. Para detallar los resultados se deja de lado *LPC* de pobre rendimiento, y se computan las matrices de confusión con la predicción realizada con *KNN* a la grabación de Emma Resmini como conjunto de test. En la Figura 8.4 se observa las matrices de confusión respectivas, todos los valores son en porcentaje, relativos a la cantidad de frames en la clase.

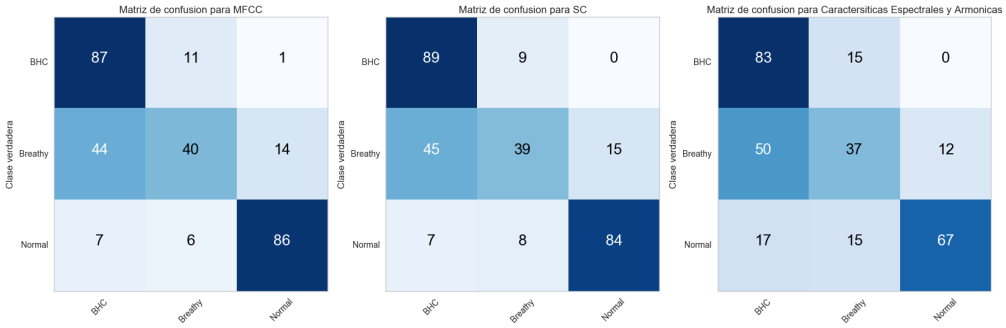


Figura 8.4: Matrices de confusión para las características *MFCC*, *SC*, y *Características Espectrales y Armónicas* de izquierda a derecha respectivamente. Para todos los casos el algoritmo de clasificación es KNN.

Los resultados demuestran, independientemente de los *features*, que las clases *Blow Hole Covert* y *Normal Embouchure* se separan frente al resto. No es el caso de *Breathy* que principalmente se confunde con *Blow Hole Covert*. Es razonable ya que se puede pensar como el caso intermedio desde el punto de vista acústico, entre las tres clases. Queda planteado entonces el punto débil en la extracción de embocadura de las características propuestas. De ahora en más se trabaja con el *feature MFCC* de mejor desempeño.

8.2.3. Segundo experimento: Blow Hole Covert Vs. Breathy

En lo que sigue se evalúa nuevamente el desempeño de *MFCC* en la separación de las clases pero con una versión reducida del problema, teniendo en cuenta

solamente las dos clases problemáticas: *Blow Hole Covered* y *Breathy Embouchure*.

Por simpleza se trabaja solamente con *MFCC* computado con 40 bandas Mel y 20 coeficientes. En la Figura 8.5 se observa el *accuracy* para los distintos algoritmos de clasificación.

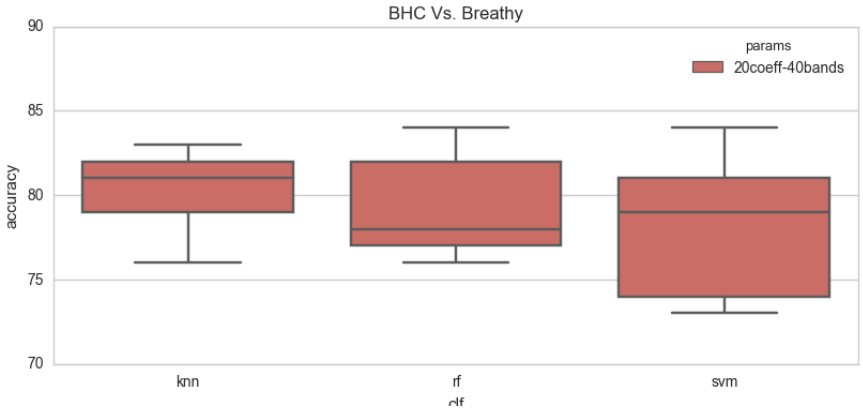


Figura 8.5: Matrices de confusión para las clases *BHC* Vs. *Breathy* generadas con características MFCC y el clasificador KNN.

El desempeño es similar al problema completo de tres clases y no existen grandes diferencias en el rendimiento según el algoritmo de clasificación. En la Figura 8.6 se observa la matriz de confusión para la predicción realizada con KNN a la grabación de Emma Resmini como conjunto de test, además se normaliza los resultados del experimento de la Sección 8.2.1 para realizar la comparación.

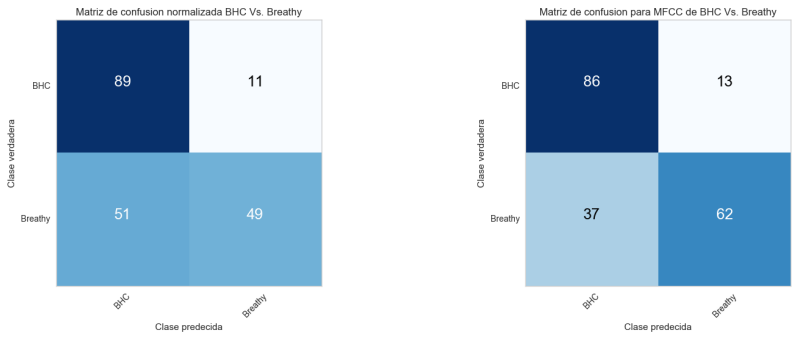


Figura 8.6: Matrices de confusión para las clases *BHC* Vs. *Breathy* generadas con características MFCC y el clasificador KNN.

Si bien sigue siendo considerable la confusión con un 37 % de elementos de *Breathy* clasificados como *Blow Hole Covert*, hay una mejora en comparación al experimento anterior, acertando ahora en la mayoría de los casos.

Una estrategia de dos etapas de clasificación en cascada mejoraría los resultados con respecto al Experimento 8.2.1 a cambio de mayor costo computacional.

8.2.4. Refinamiento de la extracción de características basada en MFCC

En lo que sigue se buscan los parámetros de *MFCC* que logran el resultado óptimo para el problema dado. Recordando la matriz de confusión de la Figura 8.4, para el mejor caso (*MFCC* + *KNN*) existe un 13 % de frames ejecutados con *Normal Embouchure* que fueron mal clasificados.

Mientras *MFCC* es una buena medida del aspecto tímbrico del material sonoro por extraer la envolvente espectral, a priori no contiene información de la periodicidad de la señal de análisis. Por lo que se propone agregar el computo de *Voicing* como una dimensión más del vector de características y evaluar si existe disminución en la confusión de la clase *Normal Embrochure*.

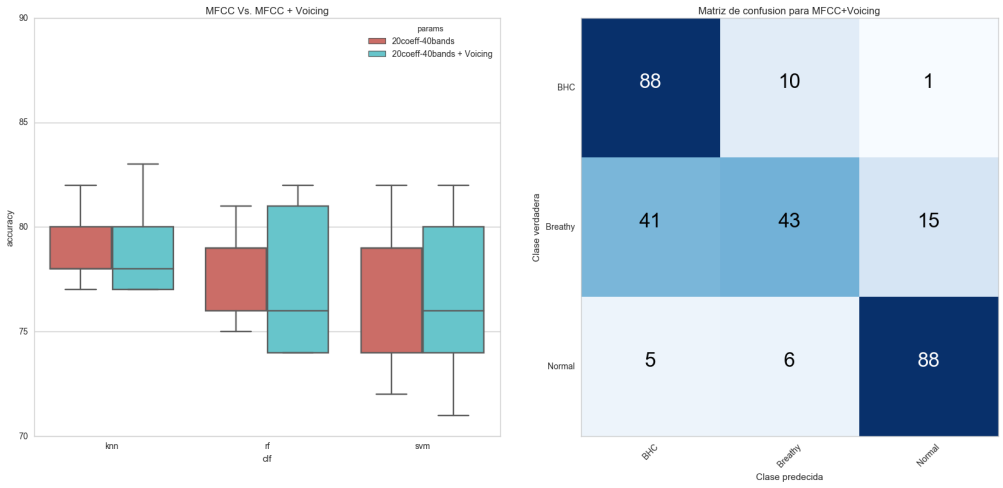


Figura 8.7: *Accuracy* y matriz de confusión para la evaluación comparativa del agregado de *Voicing* a las características *MFCC*

Se observa en la Figura 8.7 que tanto el *accuracy*, como la confusión de la clase *Normal Embouchure*, no denotan cambios relevantes de forma cualitativamente.

En lo que sigue se evalúa el desempeño de *MFCC* variando el largo de la ventana de análisis en los valores $11ms$, $23ms$ y $46ms$. En todos los casos el salto entre ventanas es del 50 %.

Como se observa en la Figura 8.8 no existen cambios relevantes en el desempeño al variar el largo de la ventana. Se puede decir que para estos valores la variación tanto en la resolución temporal como en la espectral no es suficiente como para denotar diferencias en el desempeño.

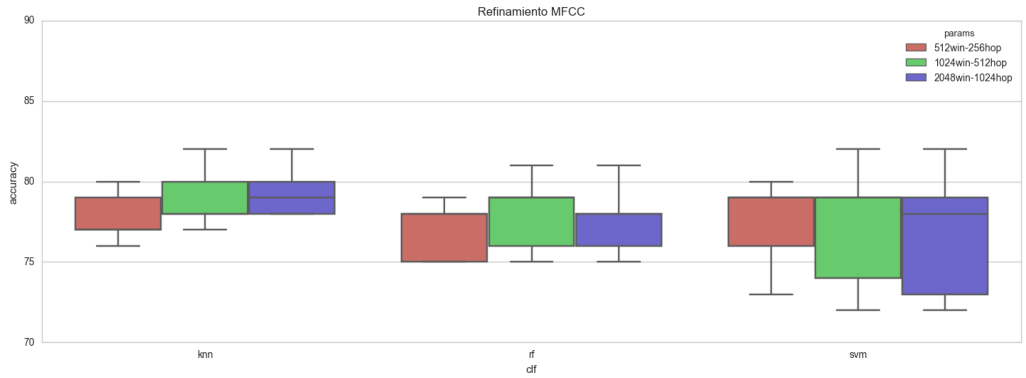


Figura 8.8: Accuracy de *MFCC* al variar el largo de ventana de análisis.

Por ultimo decir que la mejor combinación de parámetros para el cómputo de *MFCC* esta dada por una ventana de $23ms$, 40 bandas Mel y sin liftrado en el dominio de las quefrencys, resultando en un vector 40 dimensiones, siendo alta con respecto a las que se han manejado en el presente trabajo (ver Figura 8.3). Vale resaltar que usualmente se liftra la señal ya que la información tímbrica no se encuentra en las altas quefrencys pero los resultados para estos datos sugieren lo contrario.

8.3. Conclusiones

Las características evaluadas en el presente trabajo no fueron adecuadas para la separación de las clases *Blow Hole Covert* y *Breathy*. Con un estudio mas minucioso sobre la naturaleza acústica de estos dos embocaduras se podría diseñar algún *feature* que desambigüe la decisión.

Si bien se concluye que se debe mejorar la separación entre las embocaduras *Blow Hole Covert* y *Breathy*, existen pasajes de transición, inicios y finales de frases musicales, en los que se vuelve ambigua su naturaleza acústica, y se deben tener en cuenta al momento del etiquetado de las embocaduras.

Desde otro punto de vista la ambigüedad entre las clases sugiere que el enfoque basado en el aspecto tímbrico estimado como la envolvente espectral, no es suficiente. Si bien la embocadura varia el material sonoro y, en un sentido amplio, la composición tímbrica, hay que tener en cuenta que el instrumento físico no cambia. Por lo que el resonador es estacionario y mínima la variación de la estimación de la envolvente espectral a lo largo el tiempo. Esto explicaría por un lado porque los *features LPC* tuvieron un pobre desempeño y por otro porque el accuracy óptimo se logra con *MFCC's* sin liftrar. Queda planteada la hipótesis de que la información relevante está en la excitación generada por el intérprete y no por las características del resonador dadas por la envolvente espectral.

Por último vale mencionar que la estrategia de *bag of frames* es exigente, ya que se descarta la información temporal de la señal de audio, dejando de lado toda la información a priori. Es mucha la información relevante dada por ser grabaciones de audio de una interpretación musical, de un estilo definido y por si fuera poco con partitura disponible.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 9

Conclusiones

En la tesis aquí descrita se construyó una solución completa al problema de alineación entre audio y partitura para señales de flauta traversa. Para esto, se utilizaron diferentes herramientas de representación tiempo frecuencia y se desarrollaron algoritmos sobre esas representaciones. En adición a lo anterior, se desarrolló una base de datos generada a partir de obras de referencia en el repertorio de la flauta traversa que fue publicada como recurso web con fines académicos.

Por otro lado, se hizo una evaluación objetiva del sistema desarrollado mediante el análisis de la influencia de los distintos parámetros sobre el desempeño final. El análisis tuvo en cuenta además aspectos musicológicos de las obras seleccionadas. Se determinó de esta forma, que el sistema desarrollado es robusto frente a la variación en sus parámetros. Por otro lado, se presentó una comparación final de todas las estrategias implementadas, así como el desempeño de un algoritmo de terceros a modo comparativo. En base a los resultados obtenidos se considera que se resolvió satisfactoriamente el problema planteado inicialmente.

Por último se presentaron los desafíos del repertorio contemporáneo de la flauta traversa para la alineación entre audio y partitura. Además, mediante un caso de estudio se hizo la evaluación de algunas características clásicas en la literatura, para la representación del material sonoro ejecutado con técnicas extendidas.

9.1. Trabajo a futuro

Cómo trabajo a futuro queda planteado el uso de DTW en forma online para la implementación de un sistema de acompañamiento automático de flauta traversa con técnicas tradicionales. En cuanto al repertorio contemporáneo el desafío radica en encontrar representaciones matemáticas del material sonoro que puedan ser utilizados como representación intermedia, en sistemas de alineación audio partitura para flauta con técnicas extendidas.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice A

Fragmentos seleccionados

Todos los fragmentos de las obras musicales fueron seleccionados de forma que la unidad mínima fuere una frase musical. Además, solamente se nota los parámetros de altura y duración, descartando indicaciones de dinámica, articulación y variaciones expresivas de tempo. La presente sección tiene como objetivo detallar los fragmentos seleccionados en notación musical. Todas las partituras aquí presentadas fueron transcritas desde su versión en papel en el marco de esta tesis, utilizando el lenguaje de notación musical Lilypond [NN03].

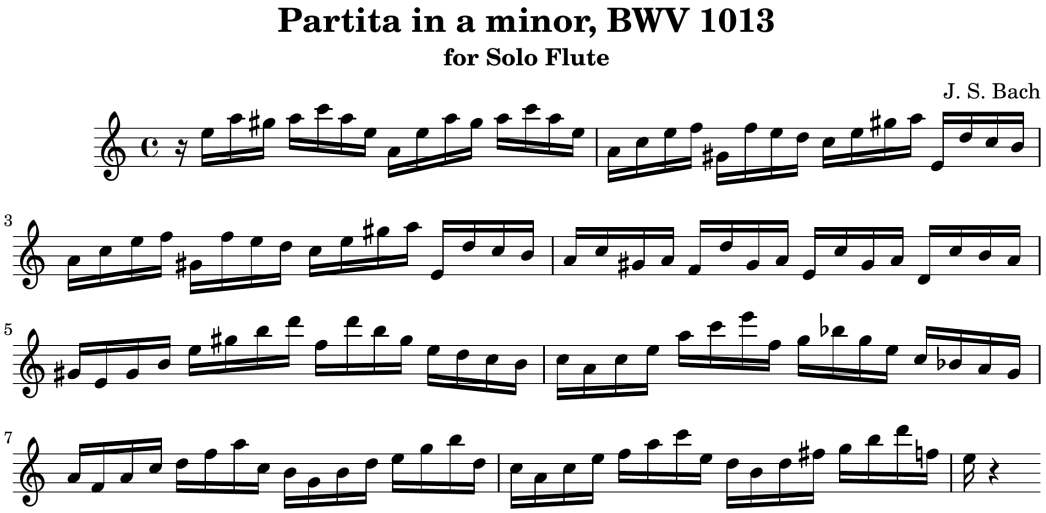


Figura A.1: Primer fragmento seleccionado del movimiento Allemande de la Partita en La menor BWV 1013, de J.S. Bach.

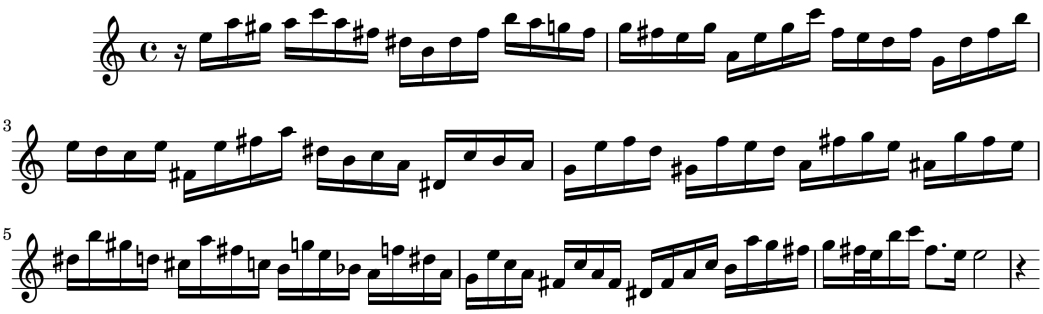


Figura A.2: Segundo fragmento seleccionado del movimiento Allemande de la Partita en La menor BWV 1013, de J.S. Bach.

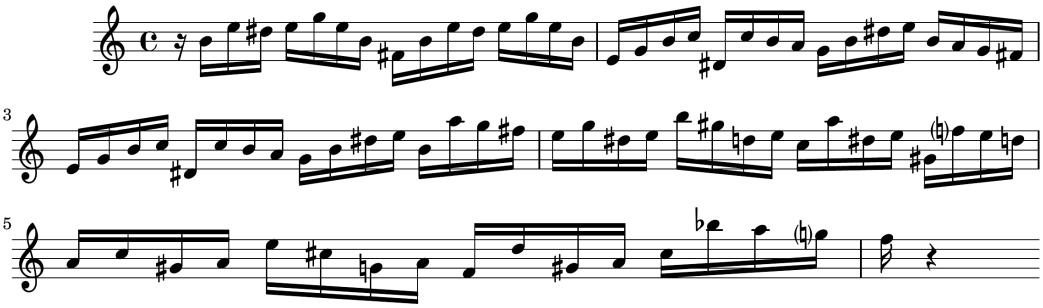


Figura A.3: Tercer fragmento seleccionado del movimiento Allemande de la Partita en La menor BWV 1013, de J.S. Bach.

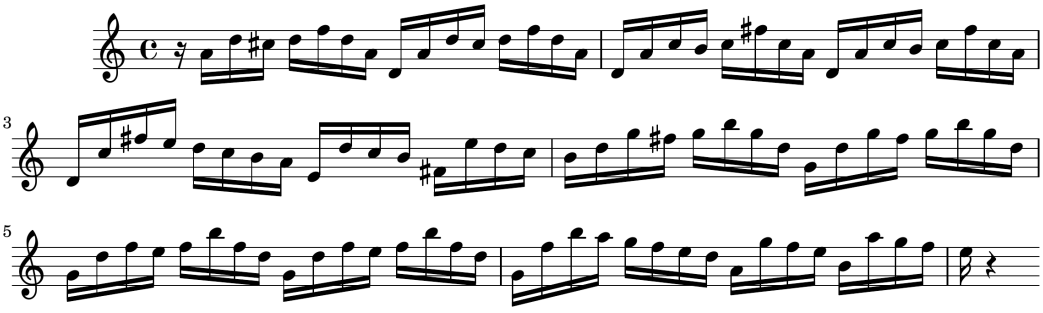


Figura A.4: Cuarto fragmento seleccionado del movimiento Allemande de la Partita en La menor BWV 1013, de J.S. Bach.

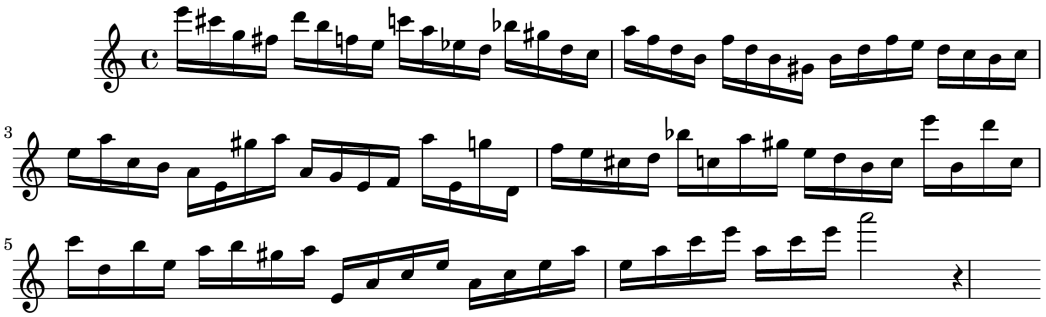


Figura A.5: Quinto fragmento seleccionado del movimiento Allemande de la Partita en La menor BWV 1013, de J.S. Bach.

Syrinx

1913

Claude Debussy

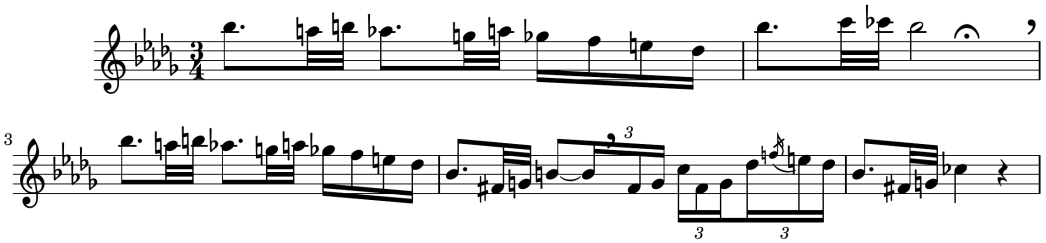


Figura A.6: Primer fragmento seleccionado de la obra Syrinx, de C. Debussy.

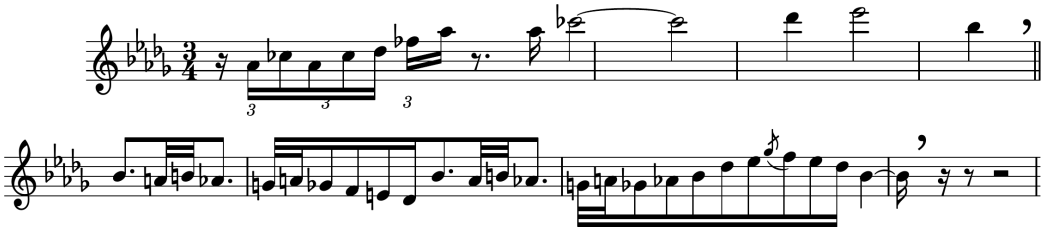


Figura A.7: Segundo fragmento seleccionado de la obra Syrinx, de C. Debussy.

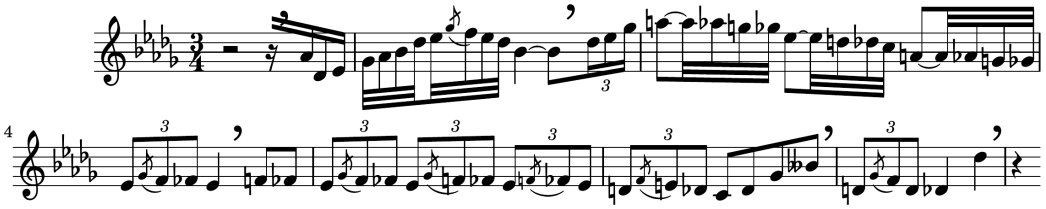


Figura A.8: Tercer fragmento seleccionado de la obra Syrinx, de C. Debussy.

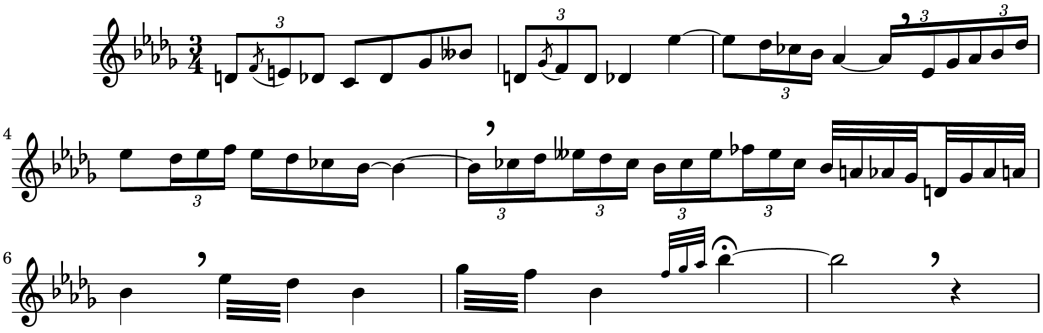


Figura A.9: Cuarto fragmento seleccionado de la obra Syrinx, de C. Debussy.

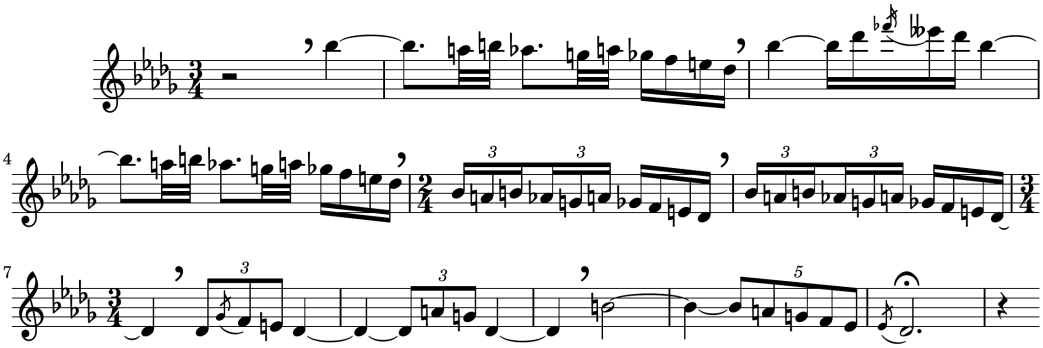


Figura A.10: Quinto fragmento seleccionado de la obra Syrinx, de C. Debussy.

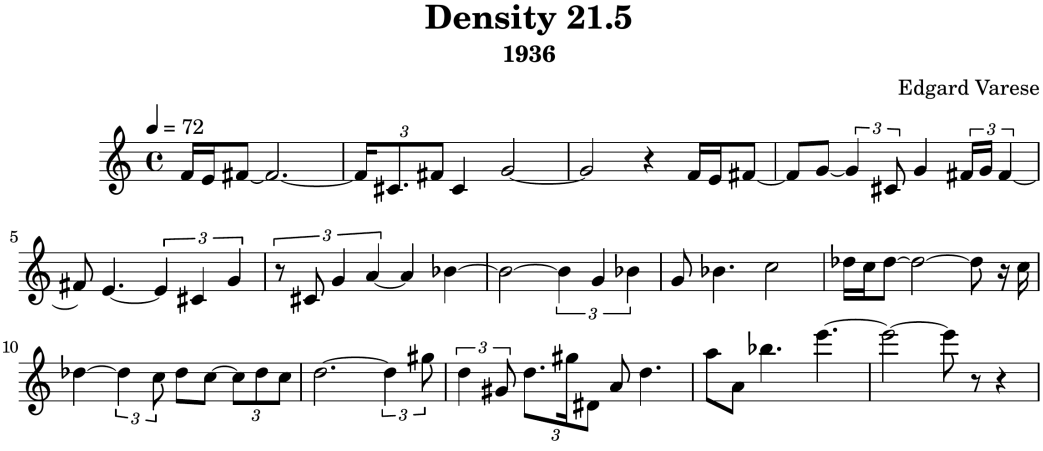


Figura A.11: Primer fragmento seleccionado de la obra Density 21.5, de E. Varese.

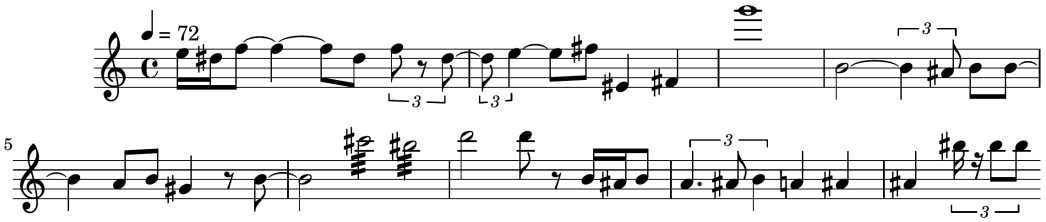


Figura A.12: Segundo fragmento seleccionado de la obra Density 21.5, de E. Varese.

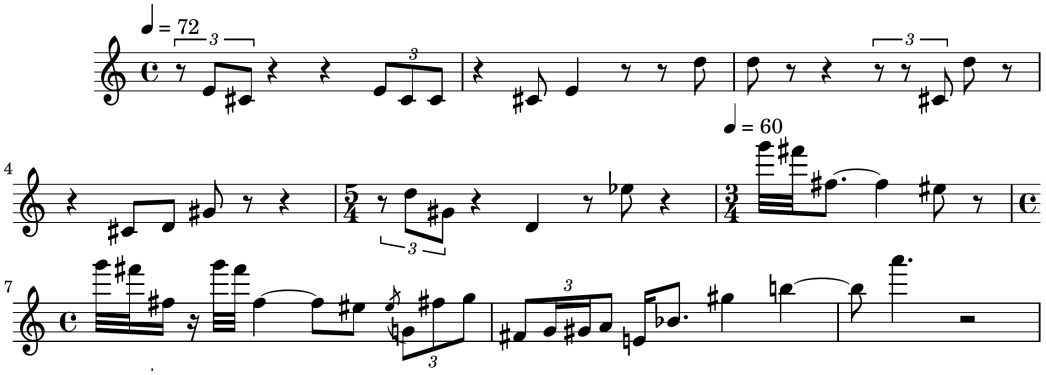


Figura A.13: Tercer fragmento seleccionado de la obra Density 21.5, de E. Varese.

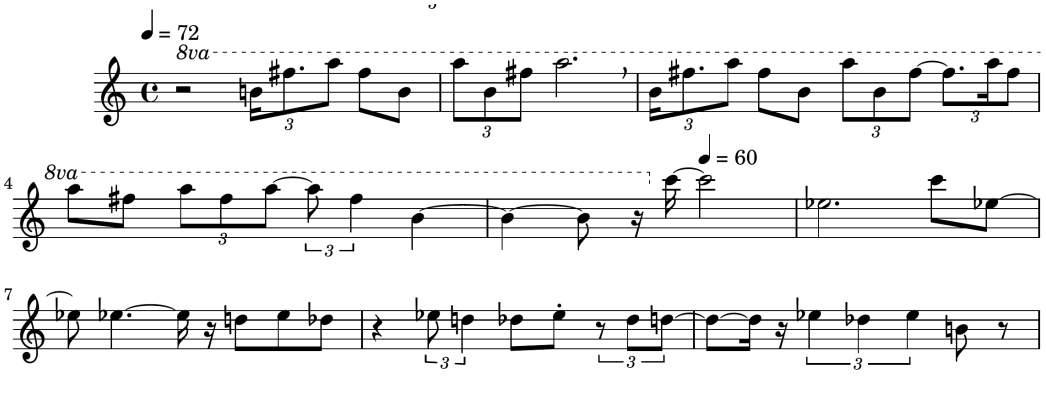


Figura A.14: Cuarto fragmento seleccionado de la obra Density 21.5, de E. Varese.

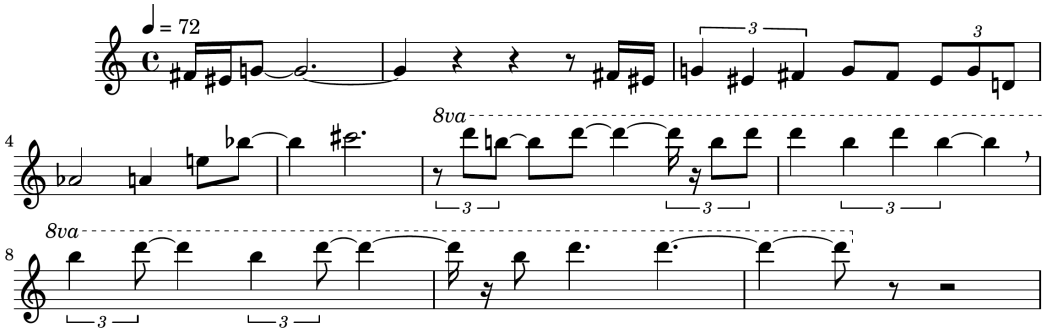


Figura A.15: Quinto fragmento seleccionado de la obra Density 21.5, de E. Varese.

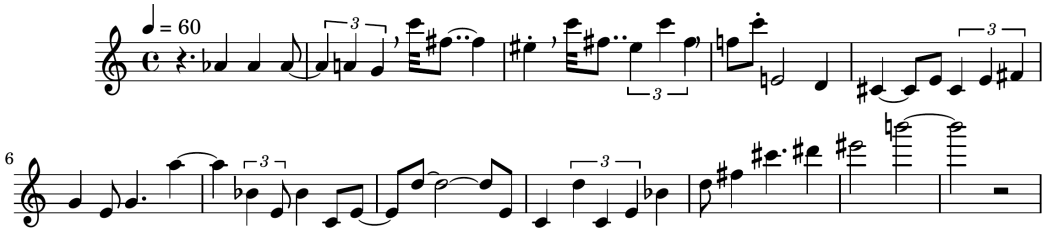


Figura A.16: Sexto fragmento seleccionado de la obra Density 21.5, de E. Varese.

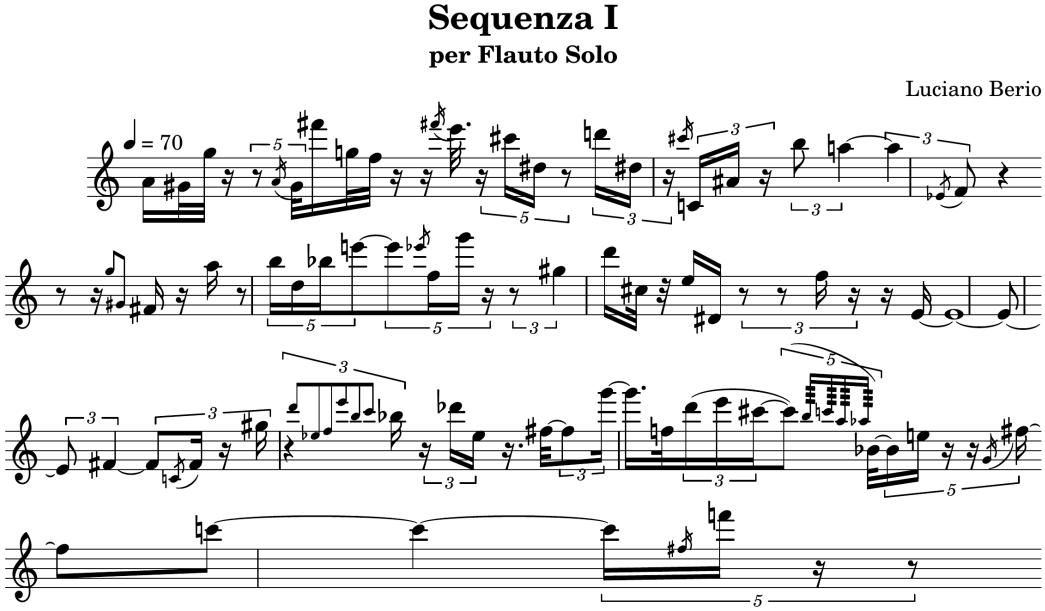


Figura A.17: Primer fragmento seleccionado de la obra Sequenza I, de L. Berio.

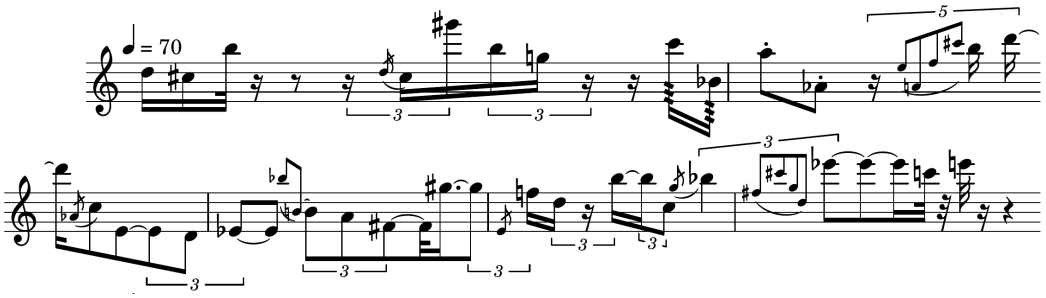


Figura A.18: Segundo fragmento seleccionado de la obra Sequenza I, de L. Berio.

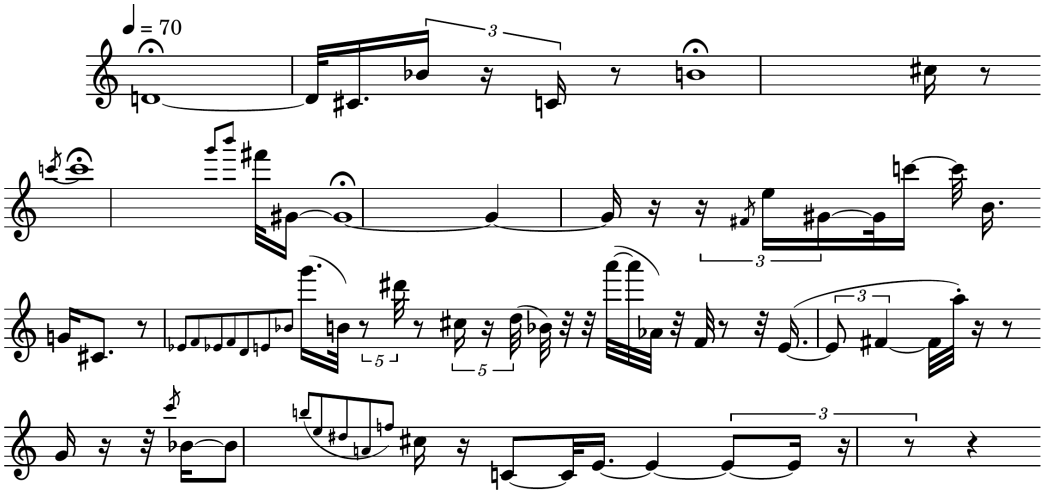


Figura A.19: Tercer fragmento seleccionado de la obra Sequenza I, de L. Berio.

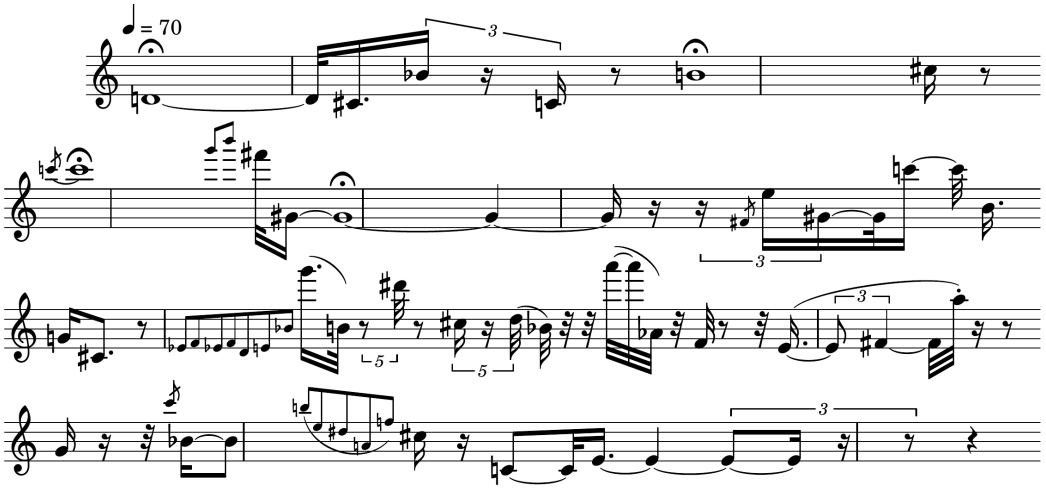
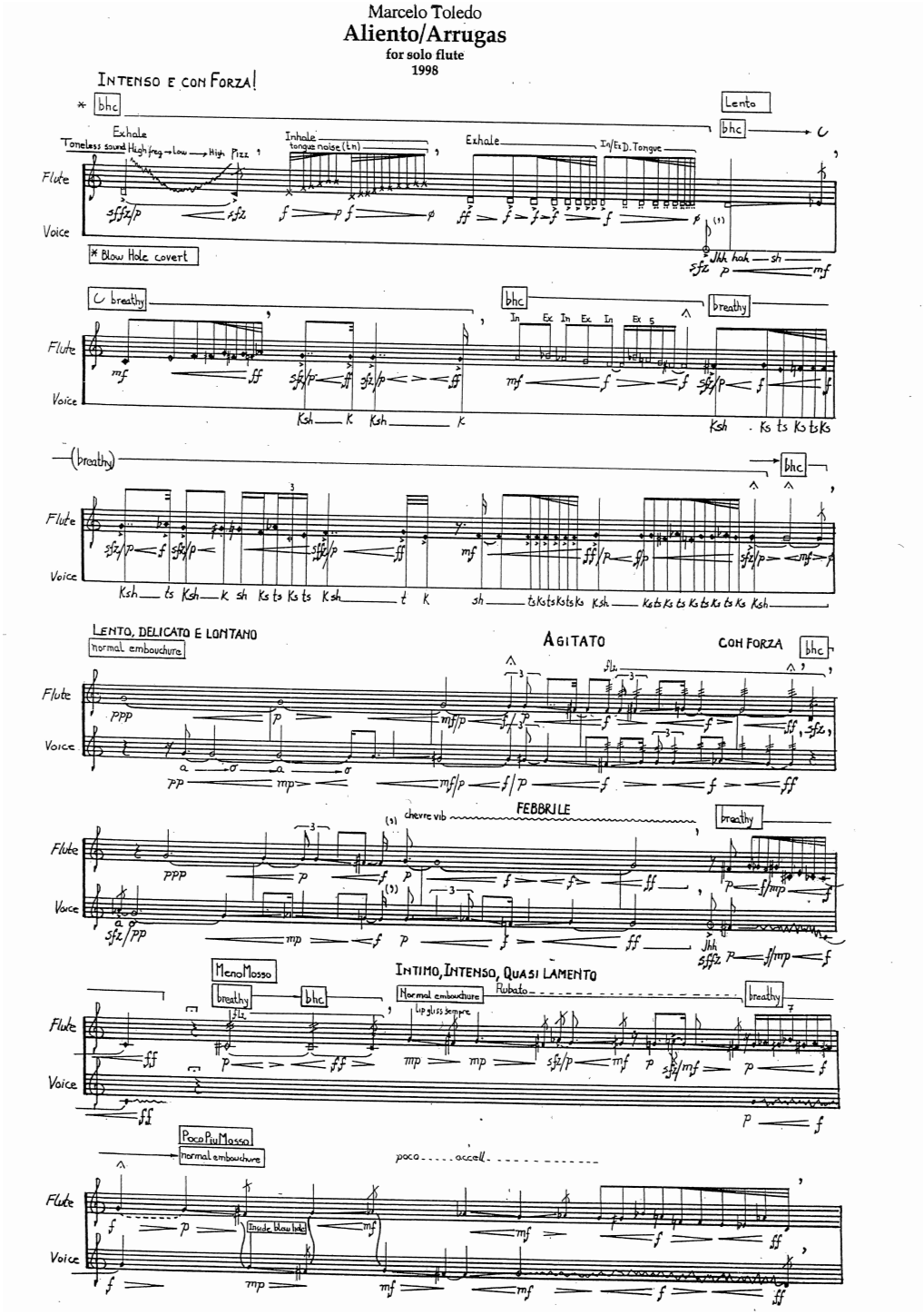


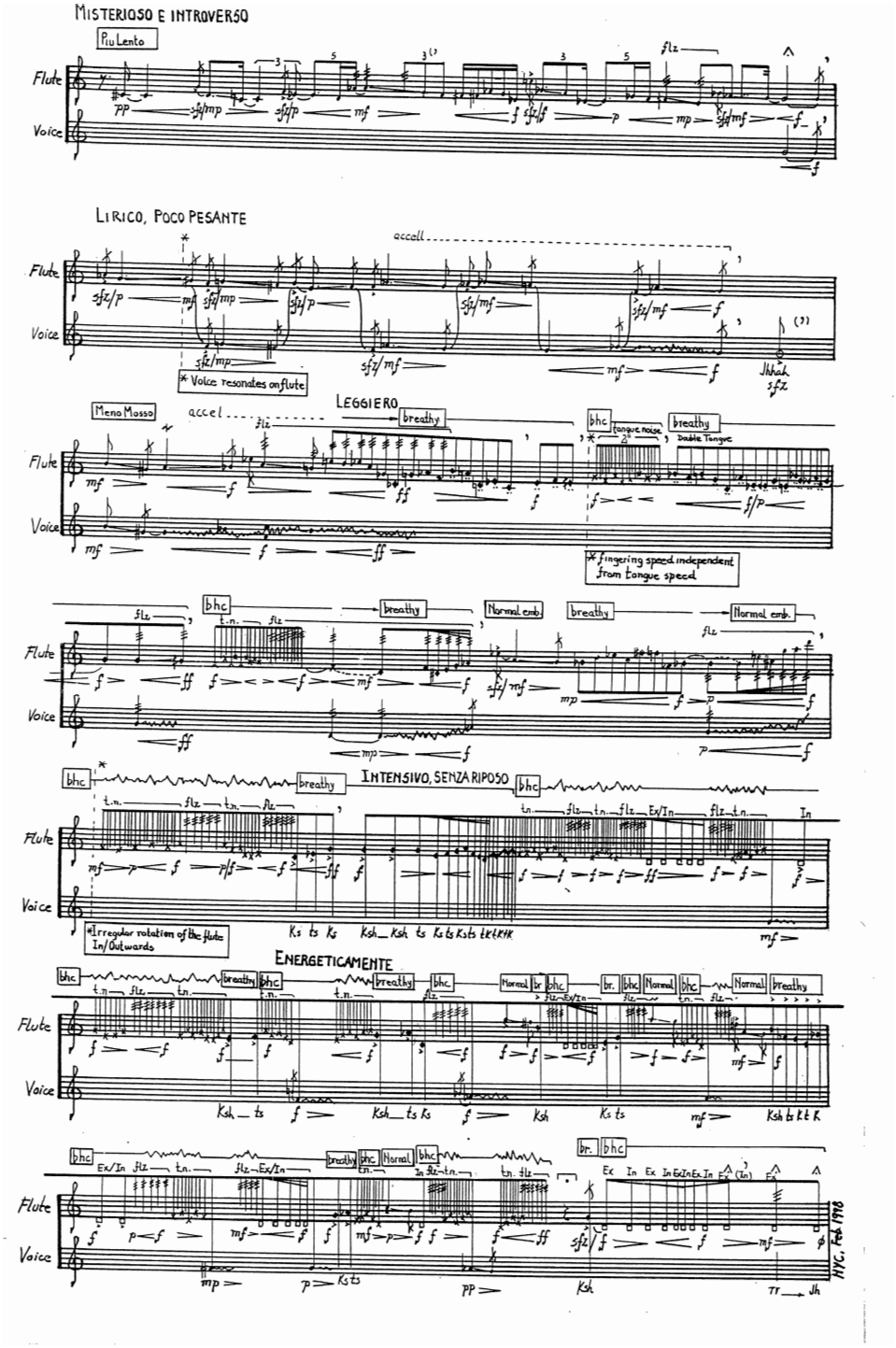
Figura A.20: Cuarto fragmento seleccionado de la obra Sequenza I, de L. Berio.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice B

Partitura de Aliento/Arrugas de Marcelo Toledo


Marcelo Toledo
Aliento/Arrugas
for solo flute
1998
INTENSO E CON FORZA!
Lento
Exhale
Toneless sound
High freq → low → High
Inhale
tongue noise (tn)
Exhale
In/Ex D. Tongue
Flute
Voice
* Blow Hole covert
breathy
LENTO, DELICATO E LONTANO
normal embouchure
AGITATO
CON FORZA
FEBBRILE
chevre vib
Meno Mosso
INTIMO, INTENSO, QUASI LAMENTO
Rubato
lip gliss sempre
Poco Più Mosso
poco ..... accell.
Inside blow hole


MISTERIOSO E INTROVERSO
Piu Lento
Flute
Voice
LIRICO, POCO PESANTE
accell
* Voice resonates on flute
Meno Mosso
accel
LEGGIERO
breathy
bhc
tongue noise
Double Tongue
* fingering speed independent from tongue speed
Normal emb.
INTENSIVO, SENZA RIPOSO
* Irregular rotation of the flute In/Outwards
ENERGETICAMENTE
Normal
br.
NYC, Feb 1978

Esta página ha sido intencionalmente dejada en blanco.

Referencias

[Arz08] Andreas Arzt. *Score following with dynamic time warping: An automatic page-turner*. na, 2008.

[BO74] Carlo Braccini and A Oppenheim. Unequal bandwidth spectral analysis using digital frequency warping. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(4):236–244, 1974.

[BP92] Judith C Brown and Miller S Puckette. An efficient algorithm for the calculation of a constant q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, 1992.

[Bro91] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[CLS10] Chris Cannam, Christian Landone, and Mark Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1467–1468. ACM, 2010.

[Con08] Arshia Cont. Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. In *International Computer Music Conference (ICMC)*, pages 33–40, 2008.

[CORSVC+15] Julio José Carabias-Orti, Francisco J Rodríguez-Serrano, Pedro Vera-Candeas, Nicolás Ruiz-Reyes, and Francisco J Cañadas-Quesada. An audio to score alignment framework using spectral factorization and dynamic time warping. In *ISMIR*, pages 742–748, 2015.

[CSSR07] Arshia Cont, Diemo Schwarz, Norbert Schnell, and Christopher Raphael. Evaluation of real-time audio-to-score alignment. In *International Symposium on Music Information Retrieval (ISMIR)*, 2007.

[D+07] Paul A Dickens et al. Flute acoustics: measurement, modelling and design by. 2007.

[Dan84] Roger B Dannenberg. An on-line algorithm for real-time accompaniment. In *ICMC*, volume 84, pages 193–198, 1984.

[Dan07] Roger B Dannenberg. An intelligent multi-track audio editor. In *ICMC*, pages 89–94, 2007.

[DCK02] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[DGW02] Simon Dixon, Werner Goebl, and Gerhard Widmer. The performance worm: Real time visualisation of expression based on langner's tempo-loudness animation. In *ICMC*, 2002.

[Dic75] Robert Dick. *The other flute: a performance manual of contemporary techniques*. Oxford University Press, 1975.

[Dix01] Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

[Dix05] Simon Dixon. Live tracking of musical performances using online time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects*, pages 92–97. Citeseer, 2005.

[DM80] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[DR06] Roger B Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM*, 49(8):38–43, 2006.

[GLB07] Bruno Gagnon, Roch Lefebvre, and Charles-Antoine Brunet. A high level musical score alignment technique based on fuzzy logic and dtw. In *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[Gro04] G Grove. Dictionary of music and musicians, ed. by fuller-maitland, 1904.

[Har76] Fredric J Harris. High-resolution spectral analysis with arbitrary spectral centers and arbitrary spectral resolutions. *Computers & Electrical Engineering*, 3(2):171–191, 1976.

[Hel76] H Helms. Power spectra obtained from exponentially increasing spacings of sampling positions and frequencies. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):63–71, 1976.

[JLZ+02] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 113–116. IEEE, 2002.

[KD07] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[Ler12] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.

[Mak75] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[MO09] Nicola Montecchio and Nicola Orio. A discrete filter bank approach to audio to score matching for polyphonic music. In *ISMIR*, pages 495–500, 2009.

[MRL+15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[Mül07] Meinard Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[NN03] Han-Wen Nienhuys and Jan Nieuwenhuizen. Lilypond, a system for automated music engraving. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, volume 1, pages 167–171, 2003.

[OD01] Nicola Orio and François Déchelle. Score following using spectral analysis and hidden markov models. In *ICMC: International Computer Music Conference*, pages 1–1, 2001.

[OJS71] Alan Oppenheim, Don Johnson, and Kenneth Steiglitz. Computation of spectra with unequal resolution using the fast fourier transform. *Proceedings of the IEEE*, 59(2):299–301, 1971.

[OLS03] Nicola Orio, Serge Lemouton, and Diemo Schwarz. Score following: State of the art and new developments. In *Proceedings of the 2003 conference on New interfaces for musical expression*, pages 36–41. National University of Singapore, 2003.

[Opp75] Alan V Oppenheim. Rw schafer digital signal processing. *Prentice-Hall, Englewood Cliffs, New Jersey*, 6:125–136, 1975.

[OS01] Nicola Orio and Diemo Schwarz. Alignment of monophonic and polyphonic music to a score. In *International Computer Music Conference (ICMC)*, pages 1–1, 2001.

[PAS82] Kuldip K Paliwal, Anant Agarwal, and Sarvajit S Sinha. A modification over sakoe and chiba's dynamic time warping algorithm for isolated word recognition. *Signal Processing*, 4(4):329–333, 1982.

[PGHK13] Matthew Prockup, David Grunberg, Alex Hrybyk, and Youngmoo E Kim. Orchestral performance companion: Using real-time audio to score alignment. *IEEE MultiMedia*, 20(2):52–60, 2013.

[Pis55] W. Piston. *Orchestration*. Norton, 1955.

[Puc90] Miller Puckette. Explode: a user interface for sequencing and score following. In *ICMC*, 1990.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[Qua02] T.F. Quatieri. *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice-Hall signal processing series. Prentice Hall PTR, 2002.

[Rab89] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Rap99] Christopher Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE transactions on pattern analysis and machine intelligence*, 21(4):360–370, 1999.

[RE16] Colin Raffel and Daniel PW Ellis. Extracting ground-truth information from midi files: A midifesto. In *ISMIR*, pages 796–802, 2016.

[RJ93] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. 1993.

[RLJ09] Martin Rocamora, Ernesto Lopez, and Luis Jure. Wind instruments synthesis toolbox for generation of music audio signals with labeled partials. In *SBCM09: Proceedings of 2009 Brazilian Symposium on Computer Music*, volume 2, pages 2–4, 2009.

[RR64] Gardner Read and Gardner Read. Music notation: a manual of modern practice. Technical report, 1964.

[RRL78] Lawrence Rabiner, A Rosenberg, and S Levinson. Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(6):575–582, 1978.

[RS78] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[RSCOVCMM17] Francisco Jose Rodriguez-Serrano, Julio Jose Carabias-Orti, Pedro Vera-Candeas, and Damian Martinez-Munoz. Tempo driven audio-to-score alignment using spectral decomposition and online dynamic time warping. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):22, 2017.

[RSMCV+15] FJ Rodríguez-Serrano, J Menéndez-Canal, A Vidal, FJ Cañadas-Quesada, and R Cortina. A dtw based score following method for score-informed sound source separation. In *Proceedings of the 12th sound and music computing conference*, pages 491–496, 2015.

[Sam02] Adler Samuel. The study of orchestration, 2002.

[SC78] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[SC07] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[She64] Roger N Shepard. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12):2346–2353, 1964.

[SK10] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64, 2010.

[Str05] Mariana Stratta. Argentine music for flute with the employment of extended techniques: an analysis of selected works by Eduardo Bertola and Marcelo Toledo. 2005.

[SVN37] Stanley Smith Stevens, John Volkmann, and Edwin 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.

[Ver84] Barry Vercoe. The synthetic performer in the context of live performance. In *Proc. ICMC*, pages 199–200, 1984.

[VP85] Barry Vercoe and Miller Puckette. Synthetic rehearsal: Training the synthetic performer. In *ICMC*, volume 85, pages 275–289, 1985.

[Wak99a] Gregory H Wakefield. Chromagram visualization of the singing voice. In *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 1999.

[Wak99b] Gregory H Wakefield. Mathematical representation of joint time-chroma distributions. In *International Symposium on Optical Science, Engineering, and Instrumentation, SPIE*, volume 99, pages 18–23, 1999.

[YB78] J Youngberg and S Boll. Constant-q signal analysis and synthesis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78.*, volume 3, pages 375–378. IEEE, 1978.

Índice de tablas

Índice de figuras

Esta es la última página.
Compilado el lunes 26 noviembre, 2018.
http://iie.fing.edu.uy/