Sociedad de Ingeniería de Audio

Artículo de Congreso

Congreso Latinoamericano de la AES 2018
24 a 26 de Septiembre de 2018
Montevideo, Uruguay

Alineación audio partitura en la flauta traversa

Juan P. Braga Brum,[1] L. W. P. Biscainho[2] y Pablo Cancela[1]

[1] Universidad de la República (UdelaR), Facultad de Ingeniería (FIng), Instituto de Ingeniería Eléctrica (IIE)
Montevideo, 11300, Uruguay

[2] Universidade Federal do Rio de Janeiro (UFRJ), Escola Politécnica (Poli), Departamento de Engenharia Eletrônica e de Computação (DEL)
Rio de Janeiro, RJ, 21941-972, Brasil

juanbragabrum@gmail.com, wagner@smt.ufrj.br, pcancela@fing.edu.uy

RESUMEN

Un sistema de alineación entre audio y partitura para señales de flauta traversa es presentado. Para eso, un abordaje desde el material sonoro generado de la flauta traversa es realizado. La comparación de desempeño de varias estrategias en la base de datos de flauta tradicional es realizada. La base de datos queda disponible para futuros trabajos en el área.

0. INTRODUCCIÓN

La alineación entre audio y partitura 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 [1]. 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. 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, herramien-

[1] Refiere a análisis en tiempo real

tas para el análisis automático de parámetros expresivos como son las dinámicas, variaciones de tempo, articulaciones, entre otros [2].

Figura 1: Esquema conceptual de la alineación entre audio y notación simbólica.

En cambio, 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 [3]. Tiene directa implicancia en la música electroacústica de medios mixtos, donde se combina el material sonoro generado por una computadora con la ejecución de instrumentos musicales.

Por otro lado, la flauta traversa es elegida por muchos compositores para la creación de música para medios mixtos[2] (i.e. Música Electroacústica). Teniendo en cuenta que sistemas basados en algoritmos de alineación audio partitura tienen directa implicancia en esta corriente musical, el alcance del presente trabajo se acota a señales de flauta traversa. Dando lugar a la creación de una base de datos de flauta traversa para evaluación de algoritmos de alineación audio partitura que se hace pública para su uso con fines académicos.

Base de datos, Experimentos,

0.1. Flauta traversa

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[3]. 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 [**?**, **?**, **?**]. 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).

0.2. 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.

Para poner en oscilación al instrumento, el flautista debe soplar superando en el interior de su boca la presión atmosférica. El trabajo[4] 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.

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

[3] 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

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

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[5], 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 **??** 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[6] 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[7] tiene una longitud de dos veces la distancia entre los nodos de presión (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). La producción del sonido a partir de $A6$ se vuelve dificultosa, siendo posible para flautistas expertos [**?**].

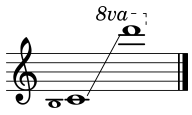

Figura 2: Registro de la flauta traversa. Se observa la cota inferior del registro para flauta con *pie en* B y *pie en* C.

0.3. Técnicas extendidas

Con el afán de extender el lenguaje musical, los compositores contemporáneos[8] se han dedicado a explorar las capacidades sónicas de los instrumentos musicales. Para esto, se siguen procedimientos como la intervención[9] 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[**?**].

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 enlistan 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

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

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

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

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

[9] 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).

línea, la ejecución de sonidos de banda angosta y banda ancha son muchas veces elementos buscados desde la composición. En el caso particular de la flauta, este nuevo material sonoro se ejecuta mediante el control de la embocadura[10] así como la presión de aire. En el Capítulo **??** 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 caracaterísticas 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).

0.4. Alcance y desarrollo de la publicacion

1. Alineación Audio-Partitura

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. De donde se define la correspondencia punto a punto mediante algún algoritmo de alineación y una distancia dada.

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 [4] 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 [3] 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 [5]. Varias piezas musicales para medios mixtos[11] 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.

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

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

1.1. Dynamic Time Warping

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:

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

Este problema de minimización, con algunas restricciones sobre las secuencias p y q, es resolube 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[**?**, **?**]. También es el caso, en el problema de alineación entre audio y partitura como se vió en la sección **??** 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

$$D[i, j] = d(\vec{X}[i], \vec{Y}[j]) \qquad (2)$$

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),

$$C[i, j] = min \begin{cases} C[i, j-1] + w_h \cdot D[i, j] \\ C[i-1, j] + w_v \cdot D[i, j] \\ C[i-1, j-1] + w_d \cdot D[i, j] \end{cases} \qquad (3)$$

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 $\vec{w} = (w_h, w_v, w_d)$[12] 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 [**?**], se utiliza $\vec{w} = (1, 1, 2)$ sin penalizar ninguna dirección.

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 comineza desde $C[M, N]$ evaluando todas las celdas vecinas buscando el mínimo, éste se agrega al comienzo del camino y de forma sucesiva el procediemitno finaliza al llegar a $C[1, 1]$. En la Figura **??** se observa a modo de ejemplo una matrices C y D.

Por otro lado en las ecuaciones 4 y 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.

$$d_{coseno}(\vec{X}, \vec{Y}) = 1 - \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\|_2 \|\vec{Y}\|_2} \tag{4}$$

$$d_{euclideana}(\vec{X}, \vec{Y}) = \|\vec{X} - \vec{Y}\|_2 \tag{5}$$

1.2. Restricciones sobre el cámino 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 de M. Muller [6].

En el caso partícular 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 cominezo 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 ventana de ajuste. Existen varias formulaciones en la literatura, en la presente tesis se trabaja con la denominada de Paliwal en honor a su autor [**?**]. 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 cámino puede contener fragmetos puramente horizontales o verticales, permitiendo saltearse 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 cámino, los expermientos realizados en la presente tesis utilizan las propuestas en [**?**]. A continuación se las define matemáticamente:

1.3. Estado del arte

Existen diversas implementaciones de sistemas de alineación audio partitura con DTW, en la publicación [7] 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 [2] 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 [8] 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 [9] el camino optimo de alineación es calculado a partir de

[12] Notar que los subindices refieren respectivamente a dirección horizontal, vertical y diagonal

información de alto nivel como es chroma y una estimación de la duración y ritmo local a partir de la señal de análisis.

Tabla 1: Resultados de la competencia Mirex en Real-time Audio to Score Alignment.

Autor (Año)	**Resultado**
Francisco J. Bris Peñalver (2017)	94 %
Francisco J. Rodriguez Serrano (2016)	97 %
Francisco J. Rodriguez Serrano (2015)	95 %
Chunta Chen* (2014)	91 %
Julio J. Carabias Orti (2013)	86 %
Julio J. Carabias Orti (2012)	83 %
Kosuke Suzuki (2011)	67 %

Un salto cualitativo en los resultados del MIREX (observar tabla 1) fue logrado por el algoritmo implemetado por J. Carbias y detallado en la publicación[10]. 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[11], 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 alineacíon se hace con DTW incorprorando información del tempo de la interpretación, mejorando notoriamente los resultados.

2. METODOLOGÍA

En la Figura 3 se observa el diagrama de la metodología de resolución. Por un lado, el bloque de extracción de contenido musical, tiene el objetivo de transformar muestras de audio en representación intermedia. Del otro lado, el bloque de codificación de notación simbólica, lleva la partitura a la misma representación.

Dado que las señales de flauta de la base de datos son ejecutadas con técnicas tradicionales el material sonoro se encuentra 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 dos formas: por alturas absolutas o por clases de altura. Es por eso que para la extracción de contenido musical se opta respectivamente por la *Constant Q Transform*[12] (CQT) y el *Chromagrama* (computado a partir del cálculo de la CQT).

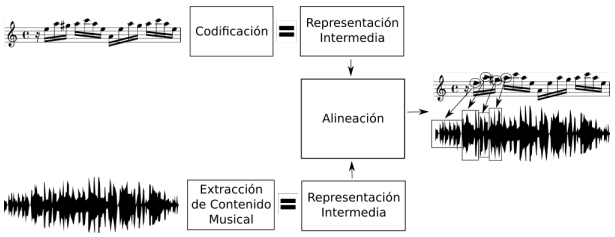

Figura 3: Esquema general de la solución del problema de alineación entre audio y partitura utilizada en el presente trabajo.

2.1. Extracción de contenido musical

La escala de la música occidental tiene 12 sonidos por octava, donde las frecuencias fundamentales se ditribuyen de forma geométricamente espaciada. Suponiendo afinación éstandar de $440Hz$ matemáticamente se escribe como:

$$F_k = 440Hz \times 2^{\,k/12} \text{ con } k \in [-50, 40] \quad (6)$$

Como se detalla en la publicación [13] la representación espectral CQT fue diseñada con el propósito de adaptarse a la organización en alturas de los sonidos musicales. Puede ser directamente calculada mediante una evaluación conveniente de la DFT. El *k-ésimo* componente se escribe como:

$$X^{CQT}[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w[n,k]x[n]e^{-j2\pi Qn/N[k]} \quad (7)$$

donde el factor de normalización $1/N[k]$ aparece para compensar la dependencia con k del número de términos en la sumatoria. Para el problema que aquí se quiere resolver, la mínima resolución queda 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.

El cálculo de la CQT se hace en el rango de frecuencias: $B3$ a $B8$. Con el criterio de abarcar el registro más amplio posible, teniendo en cuenta por un lado el límite inferior del registro de la flaut, y por otro, que las señales están muestreadas a $44100Hz$. Además, teniendo en cuenta que el Chromagrama se cálcula colapsado la CQT se calcula en octavas completas. Por lo que se obtiene para la CQT un vector de dimensión 72 y para el Chromagrama de 12.

2.2. Codificación de la partitura

El bloque de codificación tiene como entrada notación simbólica y a la salida genera una matriz que co-

difica en el eje vertical las alturas mientras que en el horizontal las duraciones (de forma análoga a la CQT y Chromagrama). La hipotesis central radica en el modelado de la notación exclusivamente bajo los parámetros de altura y duración, dejando de lado las dinámicas, articulaciónes y otros parámetros expresivos.

La duración en notación musical es expresada de forma relativa a la redonda. Para obtener la duración entonces se debe tener en cuenta el tempo sugerido por el compositor, de esta forma transformar un valor relativo de la redonda a segundos.

Por otro lado, la posición en el eje vertical queda determinada por la altura musical notada. Para esto se tienen en cuenta dos aspectos independientes, por un lado la forma de representar las alturas musicales (específicamente como alturas absolutas o clases de altura) y por otro lado la cantidad de *bins* (parámetro del sistema) que completan una octava musical. Además, la intensidad es codificada con el valor unidad (se trabaja con valores normalizados). Esta decisión se desprende del hecho que las dinámicas no se tienen en cuenta para la codificación.

Dos aspectos importantes se tienen encuenta para la codificación de la partitura. Por un lado la representación de los silencios musicales y la cántidad de armónicos en los momentos de notas. El primero se representa como un valor constante β en todo el registro. Del otro lado, en función de la naturaleza sonora de la flauta los armónicos. La decisión de los mejores parámetros es determinada en función de los resultados con la base de datos y se verá mas adelante. En la Figura 4 se observa la varición en el resultado según la cantidad de armónicos.

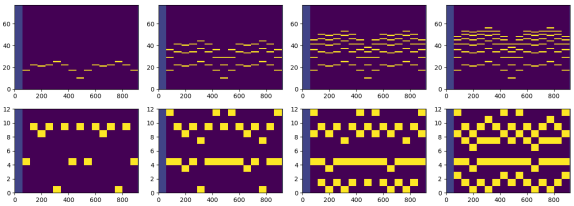

Figura 4: Cantidad de armónicos en la codificación. Arriba en alturas absolutas, abajo en clases de altura. De izquierda a derecha los armónicos elegidos son 0, 2, 4 y 6. Generado con el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

3. EXPERIMENTOS

La sección esta dividida en tres partes. Se comienza evaluando la influencia de los parámetros de representación intermedia en el resultado de alineación para determinar la mejor combinación. Posteriormente, se evalúan distintas restricciones en el cámino óptimo de alineación de DTW. Luego, 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 construída en el marco de la tesis.

3.1. Medidas de desempeño

Para la evaluación de desempeño como se recomienda en la publicación [1] se utilizan dos medidas, la tasa de aciertos y la precisión[13]. La primera cuantifíca 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 6. 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 [7].

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

3.2. Base de datos de flauta tradicional

La flauta traversa cuenta con un repertorio vasto de obras musicales asociado a su larga historia, diversos compositores han compuesto para este instrumento. La base de datos esta compuesta por fragmentos de cuatro piezas musicales ejecutadas con técnicas tradicionales. La elección de las obras determina variaciones sustanciales en los estilos musicales que la componen. Las obras seleccionadas son:

- Allemande, BWV 1013 de J.S. Bach
- Syrinx de C. Debussy
- Density 21.5 de E. Varese

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

- Sequenza I de L. Berio

De las cuatro obras músicales se tomaron grabaciones de distintos intérpretes para lograr variación en aspectos expresivos. De estas grabaciones se generaron fragmentos de forma que la unidad mínima fuere una frase musical[14] y genearon archivos de anotaciones manuales como ground truth. Resultando en un total de 30 fragmentos de audio con un archivo de texto con notación simbólica y con anotaciones manuales. Las notas que aparecen en la base van desde $C4$ a $D7$. En total existen 2245 eventos, entre notas y silencios. La base se encuentra accesible para su uso con fines académicos en: https://www.kaggle.com/jbraga/traditional-flute-dataset.

3.3. Resultados

Comparación

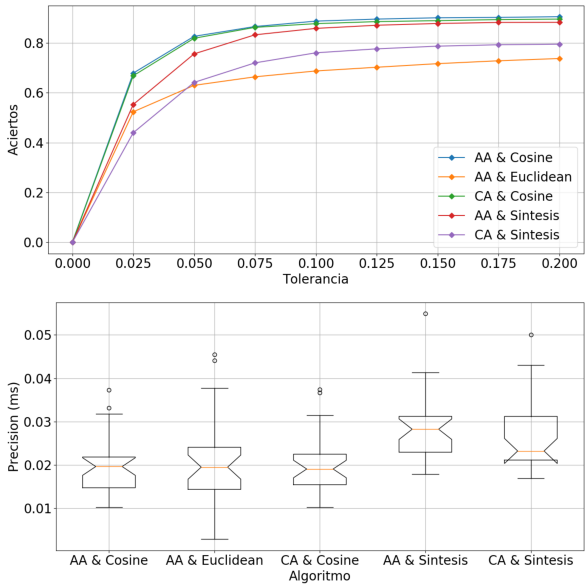

Figura 6: Arriba, tasa de aciertos en función de la tolerancia. Abajo, precisión en forma de *boxplot* para el caso $tol = 200ms$. Aclaración: AA refiere a Alturas Absolutas y CA a Clases de Altura.

Por Obra

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.

[14]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[14].

Tabla 2: Tasa de aciertos por obra.

Obra	Resultado
Allemande	96 %
Syrinx	86 %
Density 21.5	80 %
Sequenza I	70 %

REFERENCIAS

[1] 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*. National University of Singapore, 2003, pp. 36–41.

[2] Roger B Dannenberg and Christopher Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[3] Barry Vercoe, "The synthetic performer in the context of live performance," in *Proc. ICMC*, 1984, pp. 199–200.

[4] Roger B Dannenberg, "An on-line algorithm for real-time accompaniment," in *ICMC*, 1984, vol. 84, pp. 193–198.

[5] Miller Puckette, "Explode: a user interface for sequencing and score following.," in *ICMC*, 1990.

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

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

[8] Simon Dixon, "Live tracking of musical performances using on-line time warping," in *Proceedings of the 8th International Conference on Digital Audio Effects*. Citeseer, 2005, pp. 92–97.

[9] 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.

[10] 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*, 2015, pp. 742–748.

[11] 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)*, vol. 8, no. 2, pp. 22, 2017.

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

[13] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[14] Eric Blom and Denis Stevens, *Grove's dictionary of music and musicians*, St. Martin's Press, 1955.