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 para música ejecutada con flauta traversa

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

[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, pcancela@fing.edu.uy, wagner@smt.ufrj.br

RESUMEN

En este trabajo se aborda el problema de la alineación entre audio y partitura para música ejecutada con flauta. Con ese fin se hace un estudio del estado del arte en el área, así como un abordaje musicológico desde las técnicas de ejecución del instrumento. Se plantea una solución al problema para señales de flauta ejecutadas con técnicas tradicionales y la evaluación cuantitativa de desempeño en una base de datos desarrollada con este propósito. En complemento, se plantean los desafíos que presenta el repertorio contemporáneo ejecutado con técnicas extendidas. Además, la base de datos se hace disponible para futuros trabajos en el área con fines académicos.

0. INTRODUCCIÓN

La alineación entre audio y partitura consiste en la sincronización de señales de audio digital y una discripción simbólica de la misma pieza musical. 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: *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 entre la señal de audio y partitura. En la Figura 1 se observa un

[1] Refiere a análisis en tiempo real

diagrama conceptual de alineación, donde cada elemento de notación simbólica (representada como partitura) tiene su correspondencia temporal mediante los limites (i.e. comienzo y duración) de la posición en la señal de audio. La resolución del problema *offline* tiene diversas aplicaciones de interés como por ejemplo: 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 [2].

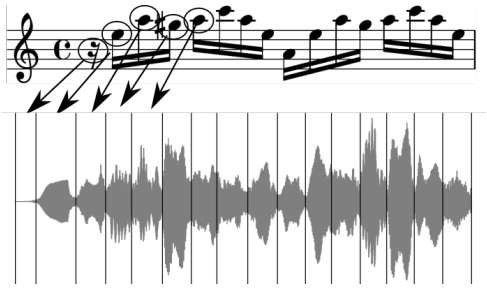

Figura 1: Esquema conceptual del problema de alineación audio y partitura.

En cambio, la resolución del problema en tiempo real cuenta con la información disponible a cada instante, determinando la alineación de forma causal únicamente con datos del pasado. Tiene como principal motivación lograr que la interacción entre computadora-humano en la ejecución de una pieza musical sea una experiencia bidireccional, simulando el comportamiento de una interpretación de un músico con otro. Denominado también como acompañamiento automático o músico sintético [3] fue la motivación que dió comienzo a esta línea de investigación. Otras aplicaciones como un pasador de páginas automático [4], o el despliegue de información sincronizada en un concierto de orquesta [5] han sido presentadas recientemente.

Por otro lado, extenso[2] es el repertorio de la flauta traversa en la música electroacústica para medios mixtos[3]. Esta corriente musical se ve favorecida por aplicaciones basadas en alineación entre audio y partitura cómo el acompañamiento automático, motivación detrás del presente trabajo que acota el problema a música ejecutada con flauta traversa. A los efectos de trabajar con señales de flauta traversa, se creó una base de datos, que se hace pública para ser utilizada con fines académicos. Esta base de datos se construyó mediante la compilación de obras de referencia del repertorio.

[2]Algunos ejemplos: Davidovsky, Mario. 1963. Synchronisms 1. flauta y sonidos electroacústicos sobre cinta analógica - Lanza, Alcides. 1977. Acufenos III. flauta, piano y sonidos electroacústicos sobre cinta analógica - Truax, Barry. 1981. East Wind. flauta y sonidos electroacústicos sobre cinta analógica - Dashow, James. 1987. Oro, argento & legno. flauta y sonidos electroacústicos sobre cinta analógica - Manoury, Philippe. 1987. Jupiter. flauta y computadora 4X Kessler, Thomas. 1988. Flute control. flauta electrónica en directo - Di Scipio, Agostino. 1990. Events. flauta bajo, clarinete bajo y electrónica en directo - Boulez, Pierre. 1991. ...Explosante-fixe... Flauta midi, 2 flautas, conjunto instrumental y computadora 4X.

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

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) son transformadas a un espacio de características donde puedan ser comparables matemáticamente, usualmente llamado cómo representación intermedia. Esta transformación genera a la salida dos series temporales de vectores con los que se determina la correspondencia punto a punto mediante algún algoritmo de alineación.

El enfoque de resolución que se implementa en el presente trabajo está basado en *Dynamic Time Warping (DTW)* debido a que los mayores desempeños reportados en los últimos 10 años lo utilizan (ver Tabla 1 resumen de la competencia MIREX[4] en alineación audio partitura en tiempo real). Esta técnica además, se ha aplicado con éxito en la resolución de problemas de *Speech Recognition*. Por este motivo se dejan de lado enfoques estadísticos de resolución, donde se destaca (por ser el más extendido) el basado en HMM (de su denominación en inglés *Hidden Markov Models*). Para más información se recomienda dirigirse a las publicaciones [6, 7, 8].

1.1. Estado del arte

Existen diversas implementaciones de sistemas de alineación audio partitura con DTW, por ejemplo en la publicación [9] 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 [10] 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 el desempeño del algoritmo. Además, en [11] 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.

Anualmente se lleva adelante una competencia de algoritmos de alineación entre audio y partitura dentro del MIREX. Como se observa en la Tabla 1 un salto cualitativo fue logrado por el algoritmo implemetado

[4]http://www.music-ir.org/mirex/wiki/MIREX_HOME

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 %

**Con un algoritmo offline*

por J. Carbias y detallado en la publicación [12]. 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. Esto resulta en una matriz de distorsión, que es utilizada como matriz de costo para el cómputo de DTW de forma online. La alternativa presentada por Rodriguez-Serrano et al. [13], que actualmente tiene el mejor resultado en la competencia, define el estado del arte. El algoritmo esta basado en el de Carabias donde el cómputo de la alineación se hace con DTW incorporando información del tempo de la interpretación, mejorando notoriamente los resultados.

1.2. 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 resoluble con el algoritmo denominado como DTW (*Dynamic Time Warping* por su denominación en inglés). Esta es una técnica clásica de programación dinámica para la alineación de series numéricas con fuerte correspondencia temporal.

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 $M \times N$ entradas que representan la distancia entre todos los pares de elementos 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:

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

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 $\vec{w} = (w_h, w_v, w_d)$[5] son factores de penalización, donde valores mayores que 1 desalientan movimientos en la dirección correspondiente. A efectos de los cálculos en el presente trabajo, se siguen las recomendaciones de [14] y se utiliza $\vec{w} = (1, 1, 2)$ para no 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 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 $C[1, 1]$.

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 los experimentos.

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

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

1.3. Restricciones

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

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

para los experimentos de la tesis, para más detalle se recomienda consultar [15].

En el caso particular de alineación entre audio y partitura existe una correspondencia directa entre notación simbólica y las grabaciones de audio. Esto es claro si se tiene en cuenta que el músico ejecuta la pieza mediante la lectura de la partitura. Haciendo posible 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**: $p[i+1] \leq p[i]+1$ 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.

2. FLAUTA TRAVERSA

Con el objetivo de resolver el problema de alineación entre audio y partitura para música ejecutada con flauta traversa se hace necesario categorizar el material sonoro producido con el instrumento. Las técnicas para la ejecución del instrumento se pueden dividir en dos grandes grupos: las denominadas técnicas tradicionales y las técnicas extendidas[6], cada grupo con un resultado sonoro característico se describen a continuación.

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 es necesario identificar por un lado la producción de la excitación periódica, y en complemento a lo anterior el largo de la columna de aire.

Producción del sonido

La excitación periódica es generada a partir de la colisión del flujo de aire con el filo de la embocadura. En otras palabras, la turbulencia provocada 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 y da la característica de sonido tonal (sonido de naturaleza periódica denominado como nota musical). En complemento, la frecuencia fundamental (i.e. altura) de la nota emitida depende estrictamente del largo de la columna de aire. Para el control de este parámetro, existen las llaves del instrumento que tapan o liberan los agujeros del tubo definiendo de esta forma el largo de la misma.

La cota inferior del registro[7] 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 $D7$ como se observa en la Figura 2. La producción del sonido a partir de $A6$ se vuelve dificultosa, siendo posible para flautistas expertos [16].

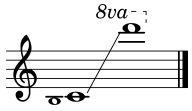

Figura 2: Registro de la flauta traversa. Se observa la cota inferior del registro para flauta con *pie en* B y *pie en* C.

Timbre

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. Éstas dan un sonido característico y se las denomina armónicos, definiendo el timbre del instrumento. De forma teórica, se pueden deducir los armónicos permitidos por la construcción del instrumento modelándolo como un tubo cilíndrico con sus dos extremos abiertos. Se deduce que la onda de mayor longitud que soporta las condiciones de borde[9] tiene dos veces la distancia entre los nodos de presión (en su forma matemática se escribe como $\lambda = 2L$). Además, que existen otras longitudes de onda permitidas asociadas a los armónicos, y se demuestra que cumplen $\lambda = 2L/k$, con $k \in N$. De donde la estructura armónica de la flauta se puede expresar como $f_i = (i+1)f_0$, con $i \in N$ y f_0 es la frecuencia fundamental en Hertz.

2.2. Técnicas extendidas

Con el afán de extender el lenguaje musical, los compositores contemporáneos[10] se han dedicado a ex-

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

[7] Esta característica asociada al instrumento determina el rango de frecuencias emitibles.

[8] Refiere a las ondas estacionarias que un medio de propagación permite.

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

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

plorar las capacidades sónicas de los instrumentos musicales. Para esto, se siguen procedimientos como la intervención[11] 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 [17].

No es objetivo de esta sección hacer una descripción exhaustiva del material sonoro generado con las distintas técnicas, para eso existe extensa bibliografía [16, 17]. 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).

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:** Inflexiones 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 elementos buscados desde la composición. En el caso particular de la flauta, este nuevo material sonoro se ejecuta mediante el control de la embocadura[12] así como la presión de aire. 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 basadas en la escala cromática para la resolución del problema de alineación entre audio y partitura en obras del repertorio contemporáneo.

[11] 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).

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

3. METODOLOGÍA

En la Figura 3 se esquematiza el método de resolución abordado. 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.

El material sonoro ejecutado con la flauta define el tipo de representación intermedia acorde. En lo que sigue se acota el problema a las técnicas tradicionales de la flauta. 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*[18] (CQT) y el *Chromagrama* (computado a partir del cálculo de la CQT).

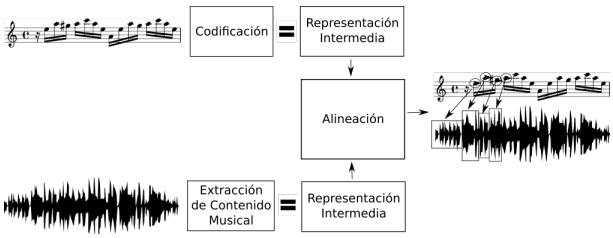

Figura 3: Esquema general de la solución del problema de alineación entre audio y partitura utilizada en el presente trabajo.

3.1. Extracción de contenido musical

La escala de la música occidental tiene 12 sonidos por octava, donde las frecuencias fundamentales se distribuyen de forma geométricamente espaciada. Suponiendo afinación estándar 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 [19] 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 flauta, y por otro, que las señales están muestreadas a 44100 Hz. Además, teniendo en cuenta que el Chromagrama se calcula colapsando el resultado de 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.

3.2. Codificación de la partitura

El bloque de codificación tiene como entrada notación simbólica y a la salida genera una matriz que codifica en el eje vertical las alturas mientras que en el horizontal las duraciones (de forma análoga a la CQT y Chromagrama). La hipótesis central radica en el modelado de la notación exclusivamente bajo los parámetros de altura y duración, dejando de lado las dinámicas, articulaciones 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 encuentra para la codificación de la partitura. Por un lado la representación de los silencios musicales y la cantidad 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 variación en el resultado según la cantidad de armónicos.

4. EXPERIMENTOS

Esta sección esta dividida en dos partes. En primer lugar se hace un ajuste de los parámetros de representación intermedia para buscar el mejor desempeño del sistema utilizando la base creada como conjunto de entrenamiento. Para finalizar se presenta una comparación de todas las estrategias implementadas: representación intermedia con organización en alturas absolutas ó con organización en clases de altura y codificación de la notación simbólica con la metodología propuesta aquí ó con la síntes como paso intermedio. Además se presenta el resultado con las dos distancias planetadas para el cálculo de la matriz de similaridad, y se evalúa el desempeño de un algoritmo implementado por terceros con la base de datos de flauta traversa compilada en el presente trabajo.

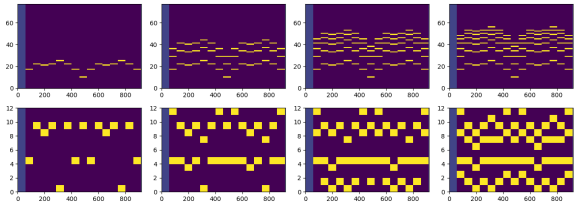

Figura 4: Representación intermedia de la codificación de la notación simbólica. Arriba en organización en alturas absolutas, abajo en clases de altura. De izquierda a derecha, los armónicos elegidos son 0, 2, 4 y 6. Generado con el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

4.1. 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 cuantifica la cantidad de notas bien identificadas, como porcentaje del total. Por otro lado, la precisión es el promedio del desfajase 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 5. 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 de la presente sección se define la tolerancia $tol = 200$ ms como se sugiere en la publicación [9].

4.2. 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 está compuesta por fragmentos de cuatro piezas musicales ejecutadas con técnicas tradicionales.

[13] Definida en este caso como una medida de desfajase 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 5: Ejemplo del resultado de etapa de alineación. Se observa la serie temporal representada como comienzo, altura y duración de las notas musicales.

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
- Sequenza I de L. Berio

De las cuatro obras musicales 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 generaron 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.

Tabla 2: Tabla con el detalle de los rangos de valores considerados para el ajuste de parámetros.

Parámetro	**Valores**
Organización	Alturas Absolutas - Clases de Altura
Resolución (ms)	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)	0.1 - 0.4 - 0.7 - 1.0

4.3. Ajuste de parámetros de la representación intermedia

En lo que sigue se presenta el ajuste de parámetros de representación intermedia. Para eso se evalúa el desempeño del sistema para los rangos de valores de cada parámetro que se indican en la Tabla 2. Con fines de ajuste de los parámetros se utiliza la distancia coseno para el cálculo de la matriz de similaridad.

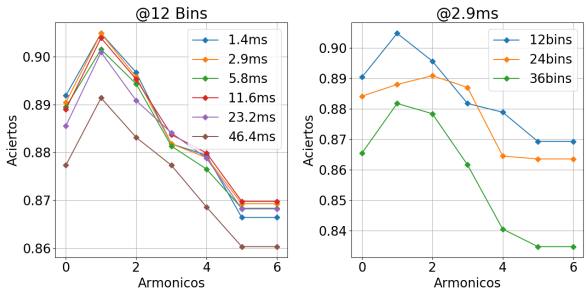

Figura 6: Tasa de aciertos en función de la cantidad de armónicos en la codificación, con organización en alturas absolutas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

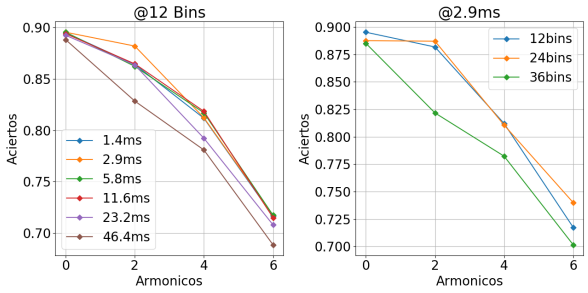

Figura 7: Tasa de aciertos en función de la cantidad de armónicos en la codificación, con organización en clase de alturas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

En las figuras 6 y 7 se observan los resultados. En líneas generales 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 obtiene los mejores resltados. Mientras que para clases de altura la codificación con único componente la fundamental, es

[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[20].

ka mejor opción. 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.

Tabla 3: Tabla con el desempeño en función del parámetro β.

β	**Alturas Absolutas**	**Clases de Altura**
0.1	82 %	90 %
0.4	77 %	87 %
0.7	76 %	86 %
1.0	76 %	86 %

4.4. Comparación

En esta sección se presenta una comparación de todas las estrategias evaluadas en el presente trabajo. Además se detalla el desempeño obtenido con Alignmidi [21] en la base de datos compilada en el marco del presente trabajo. A continuación se enlistan:

- **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 euclideana 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.
- **AA & Sintesis**: Organización en alturas absolutas y representación intermedia de la notación simbólica realizada mediante la síntesis.
- **CA & Sintesis**: Organización en clases de altura y representación intermedia de la notación simbólica realizada mediante la síntesis.
- **Alignmidi**: El algoritmo implementado por Dan Ellis.

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 eculideana hay un deterioro del desempeño.

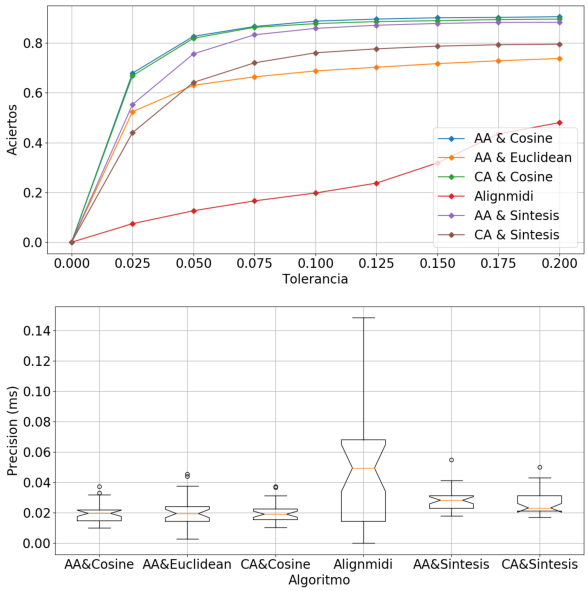

Figura 8: Arriba, tasa de aciertos en función de la tolerancia. Abajo, precisión en forma de *boxplot* para el caso $tol = 200ms$. Aclaración: AA refiere a Alturas Absolutas y CA a Clases de Altura.

5. Conclusiones

Se construyó una solución completa al problema de alineación entre audio y partitura para señales de flauta traversa con técnicas tradicionales. En adición a lo anterior, se desarrolló una base de datos compilada 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. Se determinó de esta forma, que el sistema desarrollado es robusto frente a la variación de los mismos. Además, de todas las estrategias implementadas la de representación intermedia con organización en alturas absolutas, en conjunto con la codificación de la partitura con un sólo armónico y distancia coseno obtuvo el mejor desempeño.

Por último, a partir de la evaluación de desempeño de un algoritmo de terceros en la base de datos de flauta traversa se observa que el algoritmo propuesto obtuvo mejores resultados.

5.1. Trabajo a futuro

Con el objetivo de la resolución del problema de alineación entre audio y partitura para música ejecutada con flauta traversa queda por un lado, el uso de DTW en forma online para la implementación de un sistema de acompañamiento automático. Por otro, la extensión de la representación intermedia para incorporar el material sonoro del repertorio contemporáneo de la flauta traversa.

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

[5] Matthew Prockup, David Grunberg, Alex Hrybyk, and Youngmoo E Kim, "Orchestral performance companion: Using real-time audio to score alignment," 2013, vol. 20, pp. 52–60, IEEE.

[6] Nicola Montecchio and Nicola Orio, "A discrete filter bank approach to audio to score matching for polyphonic music.," in *ISMIR*, 2009, pp. 495–500.

[7] Christopher Raphael, "Automatic segmentation of acoustic musical signals using hidden markov models," 1999, vol. 21, pp. 360–370, IEEE.

[8] Nicola Orio and François Déchelle, "Score following using spectral analysis and hidden markov models," in *ICMC: International Computer Music Conference*, 2001, pp. 1–1.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," 1978, vol. 26, pp. 43–49, IEEE.

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

[16] Adler Samuel, "The study of orchestration," 2002, WW Norton and Company, Inc, New York, London.

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

[18] 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.

[19] Judith C Brown, "Calculation of a constant q spectral transform," 1991, vol. 89, pp. 425–434, ASA.

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

[21] D. W. P. Ellis, "Aligning midi files to music audio, web resource," 2014, http://www.ee.columbia.edu/ dpwe/resources/.