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 traversa. Con ese fin, se hace un estudio del estado del arte en el área, así como una descripción de la naturaleza de las señales de flauta en correspondencia con las técnicas de ejecución del instrumento. Se plantea una solución al problema para señales de flauta ejecutadas con técnicas tradicionales y su evaluación de desempeño en una base de datos desarrollada para el 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 una señal de audio digital con una discripción simbólica de la misma pieza musical. Determinando para cada elemento de notación simbólica la correpondencia temporal en la señal de audio (i.e. comienzo y duración), en la Figura 1 se observa un diagrama conceptual. El problema se puede dividir en dos 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. 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

[1] Refiere a análisis en tiempo real

expresivos como son las dinámicas, variaciones de tempo, articulaciones, entre otros [1].

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 [2] fue la motivación que dió comienzo a esta línea de investigación. Otras aplicaciones como un pasador de páginas automático [3], o el despliegue de información sincronizada en un concierto de orquesta [4] han sido presentadas recientemente.

Por otro lado, extenso es el repertorio de la flauta traversa en la música electroacústica para medios mixtos[2]. Donde se combina el material sonoro generado por medios electrónicos con la ejecución de instrumentos musicales. Con el rol central que cumple hoy en día el computador en esta corriente musical, los algoritmos de alineación entre audio y partitura amplían las posibilidades. Tal es el caso de los sistemas de acompañamiento automático, motivación detrás del presente trabajo. Por esta razón, se decidió acotar la resolución del problema a música ejecutada con flauta traversa. A estos efectos, se creó una base de datos a partir de obras de referencia del repertorio, que se hace pública para ser utilizada con fines académicos.

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, con los que se determina la correspondencia punto a punto mediante un algoritmo de alineación.

El enfoque de resolución que se implementa en el presente trabajo está basado en *Dynamic Time Warping (DTW)*[3]. Esto se debe a que ha logrado los mayores desempeños reportados en los últimos 7 años de la competencia más importante de alineación entre audio y partitura. 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*).

1.1. Estado del arte

Existen diversas implementaciones de sistemas de alineación audio partitura con DTW, por ejemplo en la publicación [5] 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. Esta metodología es aplicable a señales polifónicas logrando mejores resultados y mayor robustez que las técnicas basadas en extracción de pitch. Por otro lado en [1] se propone la utilización de DTW con extracción de características basadas en la representación tiempo-frecuencia denominada como Chromagrama. Dixon en la publicación [6] es el primero en proponer una variante de DTW para la resolución del problema en tiempo real con la información disponible a cada instante. Además, en [7] el camino óptimo de alineación es calculado a partir de información de alto nivel simbólica como es el chroma y una estimación del ritmo local a partir de la señal de análisis.

Tabla 1: Resultados de la competencia Mirex en Real-time Audio to Score Alignment. La medida de desempeño refiere a *Over-all precision rate*, en la sección 4.1 se detalla como tasa de aciertos.

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*	

Anualmente se lleva adelante una competencia de algoritmos con aplicaciones en música, denominanada en inglés *Music Information Retrieval Evaluation eXchange* MIREX[4]. El problema de alineación entre audio y partitura, es una tarea dentro de esta competencia. Como se observa en la Tabla 1 un salto cualitativo fue

[2] Algunos ejemplos: 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] Esta técnica además, se ha aplicado con éxito en la resolución de problemas de *Speech Recognition*.

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

logrado por el algoritmo implemetado por J. Carbias y detallado en la publicación [8]. 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. [9], 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:

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

Este problema de minimización, con algunas restricciones sobre las secuencias p y q, es resoluble con DTW. Donde, 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:

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

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:

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

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 [10] 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.

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

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

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 para los experimentos de la tesis, para más detalle se recomienda consultar [11].

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**: $p[1] = q[1] = 0$ 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**: $p[i + 1] \geq p[i]$ 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.

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

- **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 profundizar en la naturaleza de estas señales. Los compositores son los que definen las características del material sonoro, mediante la elección de las técnicas para ejecución del instrumento. Éstas, se pueden dividir en dos grandes grupos: las denominadas técnicas tradicionales y las técnicas extendidas. Cada grupo con un resultado sonoro característico, comparten el universo de música para flauta traversa y determinan el tipo de técnicas computacionales utilizables para el abordaje del problema de alineación audio-partitura.

2.1. Técnicas tradicionales

Las técnicas tradicionales de la flauta son aquellas en las que el material sonoro ejecutado es definible mediante los parámetros de altura y duración. Como su nombre lo indica, las mismas refieren a los mecanismos tradicionales de producción de sonido en la flauta traversa. La naturaleza de estas emisiones es tonal generando señales monofónicas y escencialmente periódicas. En este campo, aplicaciones musicales han sido ampliamente abordadas en la literatura, y existe gran conocimiento.

Con las técnicas tradicionales, las frecuencias fundamentales se encuentran en la grilla definida por la escala cromática de la música occidental (salvo para los glissandos y vibratos). Más aún, limites bien definidos por el registro[6] del instrumento acotan el universo de posibilidades. Se tiene una cota inferior determinada por el pie de la flauta: para el caso de *pie en C* el límite es el $C4$ (262 Hz), y para el *pie en B* es el $B3$ (247 Hz). Del otro lado, en la parte alta la flauta moderna alcanza a $D7$ (2349 Hz) [12].

2.2. Técnicas extendidas

Con el afán de extender el lenguaje musical, los compositores contemporáneos[7] se han dedicado a explorar las capacidades sónicas de los instrumentos musicales. Para esto, se siguen procedimientos como la intervención mecánica[8] de instrumentos o la definición técnicas no ortodoxas de ejecución. Es así, que en el caso particular de la flauta existe un diccionario bien definido de técnicas reproducibles, denominadas extendidas [13]. Algunas de las más conocidas son (por sus denominaciones en inglés): *Flutter Tonguing:* 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) y *Cantar y tocar a la vez:* ejecución de dos alturas a la vez mediante el canto y el instrumento, entre otros.

En esta exploración de la música contemporánea el material sonoro notado por el compositor ya no se puede representar exclusivamente con alturas y duraciones. Por el contrario sutilezas tímbricas son utilizadas como recurso compositivo, resultando en señales no necesariamente monofónicas y de mayor complejidad. Por lo que, representaciones matemáticas basadas en estimación de frecuencia fundamental y bancos de filtros centrados en la escala cromática ya no son suficientes para describirlas. Se vuelve necesario, explorar otras representaciones intermedias para la resolución del problema de alineación entre audio y partitura en obras del repertorio contemporáneo.

3. METODOLOGÍA

En la Figura 2 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 se puede caracterizar a partir de las técnicas de ejecución, definiendo el tipo de representación intermedia acorde. En lo que sigue se acota el problema a las técnicas tradicionales de la flauta traversa.

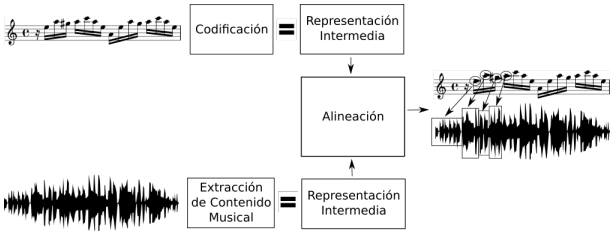

Figura 2: Esquema general de la solución del problema de alineación entre audio y partitura utilizada en el presente trabajo.

3.1. Extracción de contenido musical

Las técnicas tradicionales generan material sonoro asociado a la escala cromática de la música occidental. Estos sonidos, por su naturaleza se organizan dos formas: por alturas absolutas o por clases de altura. Por este motivo, se opta respectivamente para la extracción de

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

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

[8] 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).

contenido musical por la *Constant Q Transform* (CQT) [14] y el *Chromagrama* (computado a partir del cálculo de la CQT). Vale resaltar que si bien técnicas de extracción de pitch o estimación de frecuencia fundamental logran buenos desempeños en señales monofónicas, con el objetivo de extender a las técnicas extendidas se opta por representaciones que abarquen polifonías como las de tiempo-frecuencia.

Las frecuencias fundamentales en la escala cromática se distribuyen de forma geométricamente espaciada. Suponiendo afinación estándar de 440 Hz matemáticamente se escribe como:

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

La representación espectral CQT fue diseñada con el propósito de adaptarse a esta escala y puede ser directamente calculada mediante una evaluación conveniente de la DFT. El *k-ésimo* componente se escribe como:

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

Para el problema que aquí se quiere abordar, la mínima resolución espectral 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 $\Delta_{f_k} = f_{k+1} - f_k = f_k(2^{\frac{1}{B}} - 1)$ que para el caso de $B = 12$ se cumple que $Q = 1/(2^{\frac{1}{12}} - 1) \approx 17$. De esta forma queda determinado el tamaño de las ventanas de análisis, siendo la cantidad de bins por octava el parámetro que define el compromiso tiempo-frecuencia.

Con el criterio de abarcar el rango espectral más amplio posible, el cálculo de la CQT se hace en el rango de frecuencias: 247 Hz a 15804 Hz ($B3$ a $B9$ en notación musical). Para esta decisión se tiene en cuenta por un lado el límite inferior del registro de la flauta, y por otro, que las señales tienen frecuencia de muestreo 44100 Hz. Vale resaltar que el rango seleccionado coincide con 6 octavas completas, ya que el Chromagrama se calcula colapsando el resultado de CQT a una octava. Se obtiene entonces, para CQT y Chromagrama un vector de dimensión 72 y 12 respectivamente.

3.2. Codificación de la partitura

El bloque de codificación transforma notación simbólica de música en una serie temporal de vectores (i.e. representación intermedia). La estrategia más extendida utiliza la síntesis de audio como paso intermedio para luego, mediante la extracción de contenido musical generar la representación intermedia de la partitura. En el presente trabajo se propone por el contrario, una estrategia de codificación directa en el entendido de la naturaleza espectral de la ejecución de alturas y duraciones. Para esto se dejan de lado parámetros expresivos tipo dinámicas, articulaciones, variaciones de tempo, entre otros, modelando la ejecución como inexpresiva.

En concreto, las alturas y duraciones se codifican en los ejes vertical y horizontal respectivamente (en correspondencia con CQT y Chromagrama). La duración en notación musical es notada de forma relativa al tempo, usualmente expresado como la duración de una negra[9]. Por lo tanto, se debe tener en cuenta el tempo sugerido por el compositor para generar duraciones en segundos desde la partitura.

Por otro lado, las entradas del eje vertical tienen correspondencia directa con las alturas musicales en función de aspectos independientes: la organización de las alturas musicales (i.e. alturas absolutas o clases de altura) y la cantidad de *bins* en una octava musical (i.e. resolución espectral). Para todos los casos la intensidad es codificada con el valor unidad, en concordancia con el modelado inexpresivo de la ejecución. La cantidad de armónicos se presenta como parámetro del sistema y determina la superposición de otras alturas como se observa en la Figura 3. Por último, los silencios musicales son representados como un valor constante entre $(0, 1]$ a lo largo de todo eje vertical para diferenciarse de los momentos de nota musical. La constante es denominada como β (beta) y al igual que para el número de armónicos es paramétrico.

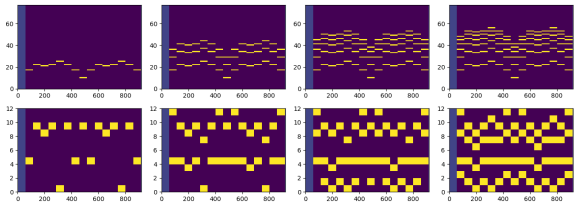

Figura 3: Representación intermedia a partir de la codificación de notación simbólica. En las figuras de arriba se observa en organización en alturas absolutas, y abajo en clases de altura. De izquierda a derecha, aumenta la cantidad de armónicos. Generado con el primer compás del movimiento Allemande de BWV 1013 de J.S. Bach.

4. EXPERIMENTOS

Esta sección esta dividida en tres partes. En primer lugar se hace un ajuste de los parámetros de representación intermedia con el objetivo de encontrar el mejor desempeño del sistema, utilizando la base creada como conjunto de entrenamiento. En segundo lugar, se presenta una comparación de las estrategias implementadas incluyendo las dos distancias en etapa de alineación, y además de la codificación directa de la partitura el agregado a la síntesis como paso intermedio. Por último se presenta el desempeño discriminado por obra musical con fines analíticos.

[9] Duración relativa asociada a una cuarta parte en una división cuaternaria

4.1. Medidas de desempeño

Para la evaluación de desempeño como se recomienda en la publicación [15] se utilizan dos medidas, la tasa de aciertos y la precisión[10]. La primera cuantifica la cantidad de notas bien identificadas como porcentaje del total, en toda la base de datos. Por otro lado, la precisión es el promedio del desfajase de las notas bien identificadas con respecto al ground truth.

Siendo $u(t_u)$ la altura del resultado de la alineación 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 [5].

4.2. Base de datos

La flauta traversa cuenta con un repertorio vasto de obras musicales asociado a su larga historia. Diversos compositores han trabajado con este instrumento en todas las epocas musicales. La base de datos está compuesta por fragmentos de cuatro piezas musicales ejecutadas con técnicas tradicionales. La elección de las obras determina variaciones sustanciales en los estilos musicales, con un orden creciente de complejidad compositiva de forma cronológica. Las obras seleccionadas son (en orden cronológico): *Allemande, BWV 1013* (1725) de J.S. Bach; *Syrinx* (1913) de C. Debussy; *Density 21.5* (1936) de E. Varese; *Sequenza I* (1958) 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[11] y generaron archivos de anotaciones manuales como ground truth. En total se tienen 30 fragmentos de audio, asociados a un archivo de notación simbólica y otro de anotaciones manuales. Las notas que aparecen van desde $C4$ a $D7$, con un total de 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.

4.3. Ajuste de parámetros

En lo que sigue se presenta el ajuste de parámetros de representación intermedia. Para eso se utiliza la base de datos como conjunto de entrenamiento evaluándose el desempeño del sistema con los valores que se indican en la Tabla 2. En esta etapa se utiliza la distancia coseno para el cálculo de la matriz de similaridad.

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

[11]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.

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)	0.05 - 0.1 - 0.4 - 0.7 - 1.0

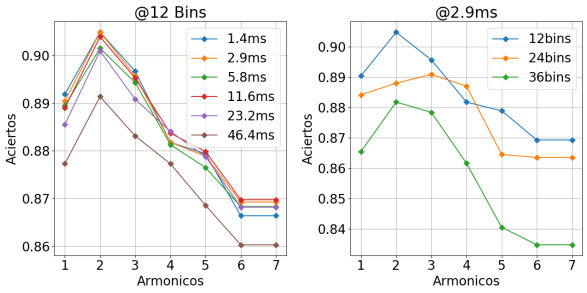

Figura 4: Tasa de aciertos en función de la cantidad de armónicos en la codificación, con organización en alturas absolutas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

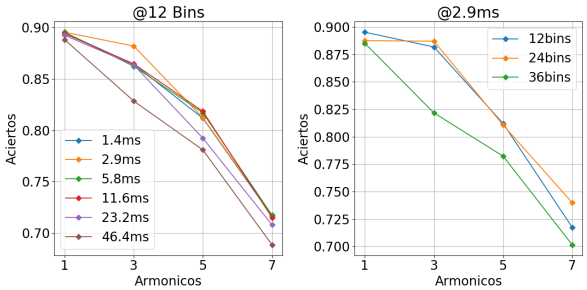

Figura 5: Tasa de aciertos en función de la cantidad de armónicos en la codificación, con organización en clase de alturas. Se observa a la izquierda el ajuste con resolución espectral fija en 12 bins por octava variando la resolución temporal. A la derecha el ajuste con resolución temporal fija en $2{,}9ms$ variando la cantidad de bins por octava.

En las figuras 4 y 5 se observan los resultados, donde para ambas organizaciones de altura 12 bins por octava y 2,9 ms son las mejores resoluciones en frecuencia y temporal respectivamente. En cuanto a la codificación de la partitura, se observa que para alturas absolutas la representación con dos armónicos obtiene los mejores resltados. Mientras que para clases de altura la fundamental como único componente, es la mejor op-

ción. Por otro lado en el modelado del silencio musical el parámetro $\beta = 0{,}1$ es el óptimo como se observa en la Tabla 3.

Tabla 3: Tasa de aciertos en función del parámetro β.

β	**Alturas Absolutas**	**Clases de Altura**
0.05	89 %	87 %
0.1	90 %	88 %
0.4	87 %	82 %
0.7	86 %	79 %
1.0	86 %	79 %

4.4. Comparación

En esta sección se presenta una comparación de todas las estrategias detalladas en el presente trabajo, teniendo en cuenta la mejor combinación de parámetros en etapa de ajuste (12 bins por octava, 2,9 ms). Para la síntesis de las partituras se utilza el *toolbox* presentado en [16]. Para claridad del lector se respasan las estrategias a continuación:

- **AA & Cosine**: Organización en alturas absolutas y dos armónicos en la codificación de la partitura. El cómputo de la matriz de similaridad se hace con distancia coseno.
- **AA & Euclidean**: Organización en alturas absolutas y dos armónicos en la codificación de la partitura. El cómputo de la matriz de similaridad se hace con distancia euclideana.
- **CA & Cosine**: Organización en clases de alturas y un armónico en la codificación de la partitura. El cómputo de la matriz de similaridad se hace con distancia coseno.
- **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.

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. En cuanto a las distancias se ve una amplio deterioro frente al resto en el caso de distancia euclideana.

4.5. Análisis por obra musical

Los resultados se observan en la Figura 7. Existe clara superioridad de desempeño sobre Allemande, siendo razonable por ser la obra más simple en estructura rítmica. Además, de forma anecdótica sucede que el deterioro es progresivo en correspondencia con el momento histórico de cada obra, asociado al aumento de la

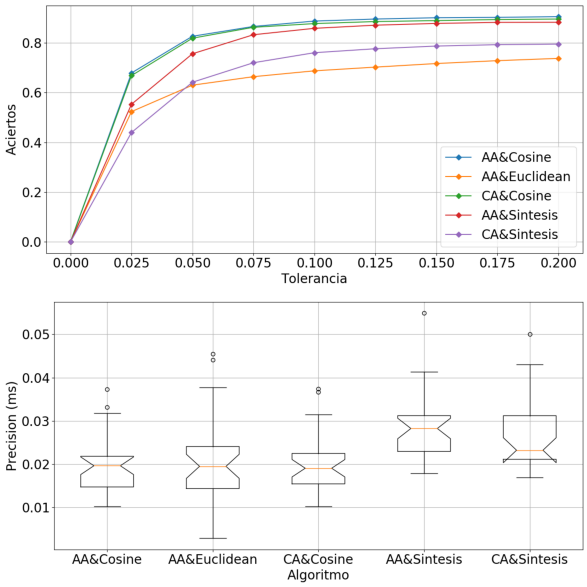

Figura 6: Arriba, tasa de aciertos en función de la tolerancia. Abajo, precisión en forma de *boxplot* para el caso $tol = 200$ ms.

complejidad compositiva. La peor tasa de aciertos se da para Sequenza I de L. Berio en correspondencia directa con su complejidad rítmica.

A modo de comparación cualitativa con el MIREX (la comparación directa no tiene sentido ya que se tratan de experimentos en bases distintas) se ve que el desempeño aquí obtenido es del orden de los resultados que se han obtenido en MIREX. Donde, para el caso particular de Allemande es comparable con los del estado del arte.

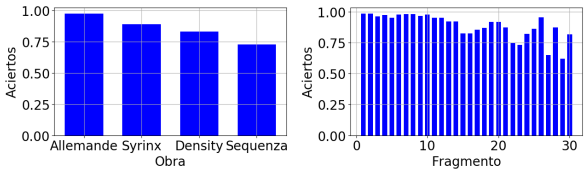

Figura 7: Tasa de aciertos discriminada por obra para la estrategia de mayor desempeño. A la izquierda agrupada por obra, a la derecha discriminada por fragmento (Allemanade: 0 a 9; Syrinx: 10 a 19; Density 21.5: 20 a 25; Sequenza I: 26 a 29).

5. Conclusiones

Se propuso una solución completa al problema de alineación entre audio y partitura para señales de flauta traversa con técnicas tradicionales. Además, la mejor de las estrategias propuestas presenta resultados en desempeño comparables a los que obtienen los mejores algoritmos de MIREX, siendo razón suficiente para

considerar como satisfactoria la estrategia implementada para la base de datos de flauta traversa. 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. Donde, la complejidad en la resolución del problema crece de forma cronológica con los años, asociado a los estilos compositivos propios de cada epoca musical. La base se hace disponible como recurso web con fines académicos.

5.1. Trabajo a futuro

En la resolución del problema de alineación entre audio y partitura para música ejecutada con flauta traversa, queda por un lado explorar la resolución de forma online mediante la modificación de DTW. Además, medir el desempeño de algoritmos del estado del arte en la base de flauta traversa con fines comparativos. Por otro, la extensión de la representación intermedia para incorporar el material sonoro del repertorio contemporáneo. En esta línea, ya fue realizado un trabajo preliminar en un caso de estudio. Donde la pieza musical involucrada contempla el control de aspectos tímbricos desde la embocadura del instrumento. Se encontró en este caso que los MFCC (*Mel Frequency Cepstral Coefficients*, de uso extendido en la literatura) tienen el mayor poder de discriminación en el material sonoro ejecutado por el flautista mediante el control de la embocadura. Por lo tanto, una buena alternativa para representación intermedia del material sonoro de esa obra.

REFERENCIAS

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

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

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

[16] 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*, 2009, vol. 2, pp. 2–4.