Seguimiento en tiempo real de interpretaciones musicales

J. A. Camarena Ibarrola
Facultad de Ingeniería Eléctrica
DES Ingenierías Arquitectura, UMSNH.

Resumen: El seguimiento en tiempo real de interpretaciones musicales permite la implementación de aplicaciones tales como la enseñanza de la música mediante un maestro virtual; el acompañamiento automático de cantantes o la inclusión automática de efectos especiales en espectáculos musicales en vivo. Los enfoques tradicionales llevan a cabo un alineamiento del "audio objetivo" y la interpretación en turno, tal enfoque implica que los inevitables errores de alineamiento se acumulen. La propuesta expuesta en este trabajo consiste en llevar a cabo búsquedas de pequeños segmentos de audio (de un segundo de duración) de la interpretación en el audio objetivo utilizando para ello índices de proximidad, esta estrategia es mas robusta puesto que no se asume nada acerca del pasado, los errores no se acumulan y además se puede realizar el seguimiento de una pieza musical ya iniciada. Los experimentos fueron hechos con 62 parejas de interpretaciones musicales, se hizo el seguimiento de una interpretación usando a la otra como audio objetivo. Los resultados fueron excelentes, se incluyen gráficas comparativas.

Motivación

Seguimiento en tiempo real de una interpretación musical consiste en establecer la posición en cada instante de la interpretación en relación con una señal de audio objetivo a medida que esta se va capturando. La señal audio objetivo está normalmente codificada de alguna manera, la mas idealista de ellas son las partituras, sin embargo, resulta más practico para nuestro propósito hacer uso de otro tipo de codificación como lo es el formato MIDI (Musical Instrument Device Interface) o la "firma" o "huella digital" del audio que es lo que se usó en este trabajo de investigación.

Maestro virtual de música: Para la aplicación de maestro virtual de algún instrumento musical podemos usar como audio objetivo a la firma de audio de la música interpretada por algún músico bien entrenado mientras que la interpretación a la que se le pretende hacer seguimiento es aquella tocada por el estudiante. Un maestro de música normalmente enseña a un estudiante a la vez, a medida que su alumno toca una pieza el maestro lo va corrigiendo dándole indicaciones de acuerdo a lo que escucha, el maestro virtual de música debería hacer lo mismo. Por supuesto, la gran ventaja de un maestro virtual a diferencia de un maestro real es que puede multiplicarse tantas veces como sea necesario pues solo se necesita una computadora donde el sistema se instale.

Acompañamiento automático: Otra aplicación de este trabajo consiste en realizar acompañamiento automático de

cantantes o de músicos solistas, para esta aplicación, utilizaríamos como audio objetivo la grabación de una interpretación del solista sin acompañamiento, es decir, sin la orquesta, mientras que el audio al que se le daría seguimiento sería una interpretación en vivo. El sistema reproduciría la música de acompañamiento variando el ritmo e introduciendo retardos de acuerdo al seguimiento que paralelamente lleva a cabo de la misma manera que lo haría una orquesta o grupo de músicos que normalmente acompañan al cantante.

Efectos especiales automáticos: Para esta aplicación, el audio objetivo sería la señal de audio grabada en una sesión de práctica, mientras que el audio al que se le daría seguimiento sería la señal capturada en el evento en vivo. Ejemplos de acciones efectuadas por tal sistema de seguimiento son el encendido de luces, la reproducción de ciertos sonidos e incluso el lanzamiento de fuegos artificiales en momentos predefinidos

Trabajos relacionados

Hay dos aspectos que definen un sistema de seguimiento de audio; Las características que extraen de la señal de audio y la técnica que utilizan para alinear la interpretación musical con el audio objetivo. En [1] se usa como característica de la interpretación musical el compás dinámico. En [2], la energía, el régimen de cruces por cero y la frecuencia fundamental fueron las características elegidas para propósitos de seguimiento de audio. En [3-4], el tono global, los valores croma, el flujo cepstral y el espectro fueron las características seleccionadas para realizar seguimiento de interpretaciones en línea.

Una vez que el conjunto de características de la señal han sido seleccionadas se utiliza normalmente una técnica de alineamiento para relacionar cada instante de la interpretación musical con su correspondencia en el audio objetivo. El enfoque tradicional para alinear dos secuencias es el Doblado Dinámico en Tiempo (DTW por sus siglas en Inglés) [5]. DTW consiste en encontrar una función óptima que establezca en que forma una de las dos secuencias debería de ser doblada o estirada para reducir las diferencias entre las dos secuencias a un mínimo. DTW fue originalmente diseñado para alinear dos secuencias, las cuales eran ambas conocidas a priori. Para nuestro propósito solo una de las dos secuencias se conoce a priori (el audio objetivo), el alineamiento tiene que ser llevado a cabo a medida que la otra secuencia llega. Recientemente en [6-7], se presentó una variante de DTW, el doblado en-línea es una adaptación propuesta precisamente para realizar seguimiento de interpretaciones musicales en línea. En [6-7] se utilizan los incrementos súbitos de energía por bandas para caracterizar la señal. El algoritmo de doblado en-línea intenta predecir el doblado óptimo avanzando por ventanas de tamaño fijo (un parámetro del algoritmo), este método tiende a desviarse del alineamiento óptimo ya que el error es acumulativo y puede incluso perder completamente la

pista al seguimiento. En [2] se usan los Modelos Ocultos de Markov (HMM por sus siglas en Inglés) son usados como técnica de alineamiento. Un HMM es un proceso doblemente estocástico con un proceso estocástico subyacente que no es observable (está oculto) y que solo puede ser observado a través de otro proceso estocástico que produce la secuencia de observaciones acústicas [8]. Encontrar los parámetros de un HMM dada una secuencia de observaciones acústicas es un problema conocido como "entrenamiento" que se resuelve mediante el algoritmo de Baum-Welch [9]. Usar un HMM (por ejemplo para hacer seguimiento) es un problema conocido como "evaluación" que se resuelve con el "procedimiento hacia adelante" o bien con el "procedimiento hacia atrás" [10]. El problema con los HMM es que el diseñador debe decidir la topología adecuada (Qué estados se deben conectar con cuales estados) y aunque el algoritmo "Viterbi" es de utilidad para esto, aún tiene que tomar decisiones críticas como el número de estados.

La propuesta

Los enfoques del estado del arte utilizan solo información local para resolver el problema de seguimiento en-línea de interpretaciones musicales, esto implica un error acumulativo, esto se debe a que tanto DTW como HMM asumen que el alineamiento realizado es correcto para todos los instantes previos y parten de esa premisa para realizar el alineamiento en cada momento. La propuesta expuesta en este trabajo consiste en no realizar alineamiento sino trasladar el problema al de búsquedas en espacios métricos, es decir, utilizando un índice de proximidad, tomar el segmento de audio mas recientemente grabado de la interpretación musical y buscarlo en el audio objetivo. Esta idea se propone combinarla con el uso de una huella de audio basada en la entropía de la señal que ha demostrado ser sumamente robusta.

Firma de Audio Basada en entropía espectral Multi-Banda

La determinación de una huella de audio muy robusta fue propuesta en [11]. La misma firma de audio fue exitosamente utilizada para monitoreo automático de estaciones de radio en [12] con resultados excelentes. Para extraer la firma de audio, la señal se divide en marcos de tiempo de 185 ms con traslape de 75%, de esta forma, un vector de características será determinado cada 46 ms; a cada marco se le aplica la ventana de Hann y luego se aplica la transformada rápida de Fourier; se determina la entropía de Shannon para las primeras 24 bandas críticas de Bark (frecuencias entre 20Hz y 7700 Hz); por cada banda se verifica si la entropía ha aumentado o no comparándola con la correspondiente al marco anterior. La ecuación (1) indica como se determina el bit de la firma que corresponde a la banda b y marco n, denotado F(n,b).

$$F(n,b) = \begin{cases} 1 \; si\,[H_b(n) - H_b(n-1) > 0 \\ 0 \qquad en\ caso\ contrario \end{cases} \quad (1)$$

$H_b(n)$ es la entropía correspondiente a la banda b marco n y $H_b(n-1)$ es la

entropía correspondiente a la misma banda, marco anterior.

Índice de proximidad: La presente propuesta implica realizar búsquedas de la firma de segmentos de audio de aproximadamente un segundo en todas las posiciones posibles de la firma del audio objetivo. Para realizar la búsqueda de manera rápida es necesario hacer uso de un índice de proximidad, se eligió el árbol de Burkhard-Keller o BK-tree por su sencillez [13]. Se hizo uso de la librería SISAP [14] para la implementación.

Experimentos

Fue posible conseguir 62 parejas de piezas musicales, cada pareja está formada por dos interpretaciones distintas de la misma pieza musical, por ejemplo, de la quinta sinfonía de Beethoven contamos con la interpretación hecha por la Orquesta filarmónica de Berlín conducida por Karajan y la interpretación hecha por la Orquesta Filarmónica de Viena conducida por Kleiber.

Por cada pareja de interpretaciones tomamos la primera de ellas como el audio objetivo y la segunda como la interpretación a la que se dará seguimiento. Se extrajo la firma de audio de la interpretación que fungiría como audio objetivo y cada pedazo de firma correspondiente a un segundo de duración se insertó en el BK-tree de la pieza musical. Los pedazos consecutivos (de un segundo) tenían entre ellos un traslape de 23/24, entonces,, para una canción de 4 minutos de duración no se insertan al índice solo 240 pedazos de firma sino 5700 aproximadamente, de ahí la importancia de usar índices de búsqueda. La operación se repitió por cada audio objetivo de manera que se construyeron 62 índices. Una vez construidos los índices se usaron para el seguimiento de las segundas interpretaciones musicales de cada pareja. En lugar de solo buscar al segmento más parecido se buscaba a los K más parecidos, luego se elegía entre ellos al más cercano (en tiempo) a la última posición reportada por el sistema. En la **Figura 1** se muestran los resultados obtenidos después de todos los ensayos, se observa que al aumentar K, disminuye el régimen de errores de seguimiento. Se probaron tres diferentes distancias entre segmentos cortos: La distancia de Hamming, de Levenshtein y DTW, las últimas dos permiten que al comparar los dos segmentos cortos de un segundo estos se alineen, esta modificación solo mejora el régimen de errores para valores de K muy bajos. En la **Figura 2** se muestra el seguimiento de una interpretación musical para diferentes valores de K. El eje horizontal corresponde con la evolución de la interpretación en segundos y el eje vertical indica la posición en que cada segmento de audio de un segundo fue localizado en el audio objetivo también en segundos.

Conclusiones

Los experimentos nos muestran que es posible realizar el seguimiento de interpretaciones musicales en tiempo real mediante búsquedas de segmentos cortos de solo un segundo de duración en el audio objetivo utilizando un índice

de proximidad. También se observó la capacidad del sistema de recuperarse de los errores de localización, es decir que los errores no se acumulan como en las técnicas tradicionales, esto se debe al hecho de que el alineamiento no se hace con información local solamente sino global (la búsqueda se hace en todo el audio objetivo y no solo en una región cercana a la posición actual). Observamos como al incrementar K (El número de vecinos) estos errores disminuyen drásticamente e incluso desaparecen. Finalmente hay que destacar la ventaja de este método que permite iniciar el seguimiento en cualquier momento aunque la interpretación haya iniciado mucho antes.

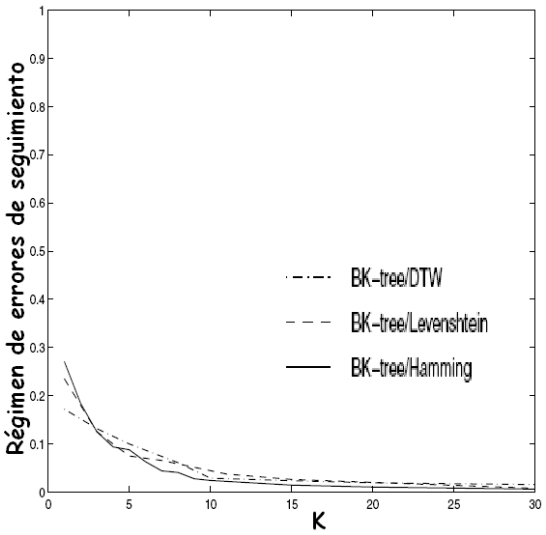

Fig. 1 Resultados de los experimentos. El régimen de errores disminuye al aumentar K independientemente de la medida de distancia

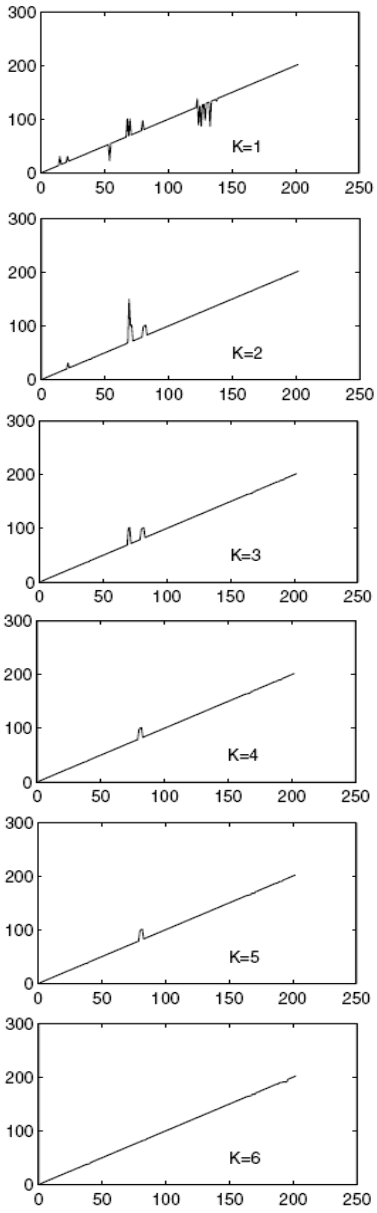

Fig. 2 Seguimiento de una interpretación musical para diferentes valores de K. Ocurren fallas para valores bajos de K

Referencias

[1] W.A. Sethares and R.D. Morris and J.C. Sethares, “Beat tracking of musical performances using low level audio features”, IEEE Transactions on Speech and Audio Processing, Vol 2, Mar. 2005.

[2] P. Cano and A. Loscos and J. Bonada, "Score-performance matching using hmms", in ICMC1999, Audiovisual Institute, Pompeu Fabra University, (Spain) ,1999.

[3] N. Orio and F. D´echelle, "Score following using spectral analysis and hidden Markov models", in Proceedings of the ICMC, pp. 151–154, 2001.

[4] N. Orio and S. Lemouton and D. Schwarz, "Score following: state of the art and new developments", in Proceedings of the, conference on New Interfaces for Musical Expression, National University of Singapore, pp. 41, 2003.

[5] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics and Speech Signal Processing (ASSP), pp 43–49, 1978.

[6] S. Dixon, "Live tracking of musical performances using on-line time warping", in 8th International Conference on Digital Audio Effects (DAFx 2005), Austrian Research Institute for Artificial Intelligence, (Vienna), Sept. 2005.

[7] S. Dixon and G. Widmer, "Match: A music alignment tool chest", in 6th International Conference on Music Information Retrieval (ISMIR), Austrian Research Institute for Artificial Intelligence, (Vienna), 2005.

[8] L. Rabiner and B. Juang "An introduction to hidden markov models". IEEE ASSP Magazine, Vol 3(1), pp 4–16, 2003.

[9] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", Technical Report TR-97-021, Department of Electrical Engineering and Computer Science U.C. Berkeley, Apr. 1998.

[10] R. L. Rabiner "A tutorial on hidden markov models and selected aplications in speech recognition", Proceedings of the IEEE, vol 77(2), pp 257–286,1989.

[11] A. Camarena-Ibarrola and E. Chavez, "On musical performances identification, entropy and string matching", in MICAI 2006. LNCS (LNAI), pp. 952–962. 2006.

[12] A. Camarena-Ibarrola and E. Chavez and E. S. Tellez, "Robust radio broadcast monitoring using a multi-band spectral entropy signature", in 14th Iberoamerican Congress on Pattern Recognition, pp. 587–594. 2009.

[13] W. A. Burkhard and R. M. Keller, "Some approaches to best-match file searching". ACM, vol 16(4), pp 230–236,1973.

[14] K. Figueroa and E. Chávez and G. Navarro, "The sisap metric indexing library", http://www.sisap.org/library/metricspaces.tar.gz

Autores:

J. Antonio Camarena Ibarrola. Ingeniero Electricista egresado de la Universidad Michocana en 1986. Obtuvo el grado de Maestro en Ciencias Computacionales en 1996 en el Instituto Tecnológico de Toluca y de Doctor en 2008 en la Universidad Michoacana.

email: camarena@umich.mx

Dirección de los autores: edif. 'Omega2' PB, C.U. Morelia, Mich. Méx.