Sociedad de Ingeniería de Audio

Artículo de Congreso

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

Reconocimiento de patrones rítmicos en música de percusión a partir de señales de audio

Bernardo Marenco[1] y Martín Rocamora[2]

[1] Instituto de Matemática y Estadística Rafael Laguardia, Universidad de la República
Montevideo, 11300, Uruguay

[2] Departamento de Procesamiento de Señales, Instituto de Ingeniería Eléctrica, Universidad de la República
Montevideo, 11300, Uruguay

bmarenco@fing.edu.uy, rocamora@fing.edu.uy

RESUMEN

Se presenta un sistema en desarrollo para el reconocimiento automático de patrones rítmicos en música de percusión a partir de señales de audio. La clasificación se lleva a cabo utilizando cadenas ocultas de Markov. El desempeño es muy bueno al entrenar y validar con archivos de audio sintéticos, pero no generaliza adecuadamente al validar con grabaciones reales.

0. PRESENTACIÓN DEL PROBLEMA

En este trabajo se aborda el reconocimiento automático de patrones rítmicos en música de percusión, a partir de señales de audio. El enfoque presentado se centra en el candombe, un género musical de origen africano propio del Uruguay.

El ritmo del candombe surge de la interacción de los patrones rítmicos de tres tambores diferentes, denominados chico, piano y repique. El repique es el improvisador, responsable de generar interés, sorpresa y variedad musical. Existe un patrón característico de su toque, del que se realizan diversas variaciones. Además, suelen ejecutarse figuras que se alejan bastante de ese patrón, por lo que su toque tiene un alto grado de complejidad.

En [1], Luis Jure postula una serie de principios generativos para el toque de repique a partir de un axioma y reglas transformacionales. El axioma describe al patrón característico del repique, mientras que las reglas transformacionales indican algunas de las posibles variaciones que se utilizan para modificar dicho axioma y así obtener nuevos patrones rítmicos.

Este trabajo propone una metodología para la clasificación automática de algunos de los patrones rítmicos propuestos.[1]

[1] El trabajo fue parcialmente financiado por la Agencia Nacional

0.1. Principios generativos

Jure plantea como axioma un patrón rítmico que abarca un compás, tal como se muestra en notación musical en la Figura 1. La figura de negra da el pulso del ritmo, en un compás de 4/4. Luego, identifica tres niveles constitutivos del axioma, que denomina Inicio (I), Núcleo (N) y Final (F), como se muestra en la Figura 1.

Dado que estos niveles tienen duraciones diferentes, para este trabajo se dividió el núcleo en dos patrones de igual duración. Así, se tienen 4 patrones a reconocer que duran un pulso: Inicio (I), Núcleo 1 (N1), Núcleo 2 (N2) y Final (F), como se indica en la Figura 1.

Las reglas transformacionales que se plantean para el axioma consisten esencialmente en la repetición de estos patrones, la sustitución de alguno por otro y el adorno de los patrones modificando o agregando golpes. Por una descripción detallada, ver [1].

1. Metodología

El reconocimiento de los patrones rítmicos se lleva a cabo utilizando cadenas ocultas de Markov (HMMs por sus siglas en inglés). Se entrena una HMM por cada patrón a reconocer. Por más información sobre el uso de HMMs como herramienta de clasificación, ver [2].

1.1. Entrenamiento

La Figura 4 muestra un diagrama de bloques del proceso de entrenamiento de una HMM. Primero, el audio de entrenamiento es segmentado en pulsos. El entrenamiento se realiza con audios sintéticos, por lo que los tiempos de comienzo de cada pulso están completamente determinados desde la generación de los datos.

Luego, se calculan los primeros 10 coeficientes cepstrales de frecuencias mel (MFCCs por sus siglas en inglés) de la señal. Se descarta el primer coeficiente, dado que es una medida global de la energía de la señal. Se agrega además como característica el flujo espectral de la señal (normalizado usando media móvil). Todas las características son calculadas con un ancho de ventana de 40 ms y salto de 10 ms.

Este vector de características será el vector de observaciones de la HMM; se asume que tiene distribución gaussiana en $\mathbb{R}^{10}$.

Para los estados ocultos de la HMM se usa una cadena de derecha a izquierda, que es una concatenación de cadenas como la de la Figura 3. Cada estado oculto representa la presencia de un golpe en el frame de audio o un silencio. Para determinar en qué estado oculto se está en cada frame de audio, es necesario conocer la ubicación de los golpes en la señal; se asume que esta información se conoce de antemano (lo que de hecho sucede para audios sintéticos).

En la Figura 5 se muestra un ejemplo de lo que resulta de esta etapa: para cada trama se tiene una descripción del espectro (los MFCCs 2 a 10) más el flujo espectral para esa trama, y la asignación a alguno de los estados ocultos. Con esta información se ajustan los parámetros de la HMM para maximizar la probabilidad de esa secuencia de observaciones, usando el algoritmo Baum-Welch [3].

1.2. Clasificación

Para clasificar un audio, las primeras dos etapas son iguales a las del proceso de entrenamiento: se segmenta el audio en pulsos y se extraen las características. Luego, para cada pulso, se calcula la probabilidad de haber observado la secuencia de vectores de características según cada HMM y se clasifica el pulso como perteneciente a la clase cuya HMM da la mayor probabilidad.

2. Resultados experimentales

Se plantea un esquema de pruebas de complejidad incremental. Primero se evalúa el funcionamiento del sistema para audio sintético en el que solo aparecen los patrones I, N1, N2 y F, sin adornos. En este caso el sistema reconoce sin problemas todos los patrones de forma correcta. Cabe señalar que los patrones N1 y N2 solo se diferencian en que los golpes de mano y palo están intercambiados. Por esta razón, que el sistema detecte correctamente cada patrón indica que está teniendo en cuenta el contenido tímbrico de la señal.

Luego, se evalúa utilizando un audio sintético que es una transcripción de una interpretación real[2]. En ese caso no solo aparecen los patrones originales, sino que también hay algunos adornados. Se observa que el sistema es capaz de absorber estos adornos: por ejemplo, el primer patrón de la Figura 2 es un inicio al que se le agrega un golpe de palo en el tercer tiempo; el sistema lo reconoce como un I.

Si el sistema entrenado con señales sintéticas es utilizado para procesar una grabación real el desempeño decae drásticamente. Esto se debe principalmente a que las características acústicas de la señal real no se corresponden con las de las grabaciones sintéticas.

3. Discusión y trabajo futuro

Los resultados obtenidos con el sistema planteado son auspiciosos cuando se trata de reconocimiento en audios sintéticos. Si se usan audios reales, el sistema no funciona como se espera. Esto es razonable, ya que el entrenamiento se realiza únicamente con audios sintéticos. Así, surge como continuación natural la incorporación de patrones reales en la etapa de entrenamiento; el mayor obstáculo a esto es la baja disponibilidad de transcripciones. Otra posibilidad es adaptar las características de la señal, para que los dos escenarios sean comparables, por ejemplo usando regresión lineal por máxima verosimilitud, de forma similar a lo que se hace en procesamiento de voz [4].

de Investigación e Innovación.

[2] La transcripción puede verse en la figura 19 de [1].

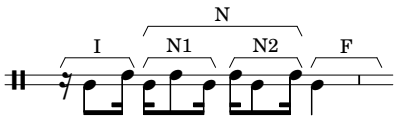

Figura 1: División del axioma en sus tres niveles constitutivos: inicio (I), núcleo (N) y final (F). El núcleo se divide a su vez en N1 y N2.

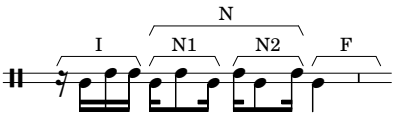

Figura 2: Ejemplo de adorno en I.

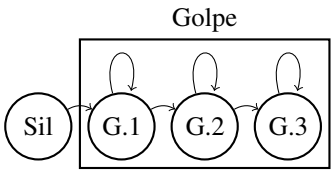

Figura 3: Diagrama de transiciones entre estados ocultos.

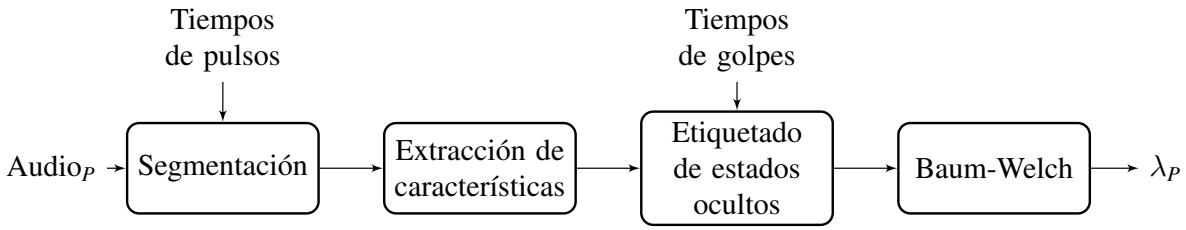

Figura 4: Diagrama de bloques del proceso de entrenamiento. Audio$_P$ representa el audio de entrenamiento asociado al patrón P y λ_P es el conjunto de parámetros de la HMM al finalizar el entrenamiento.

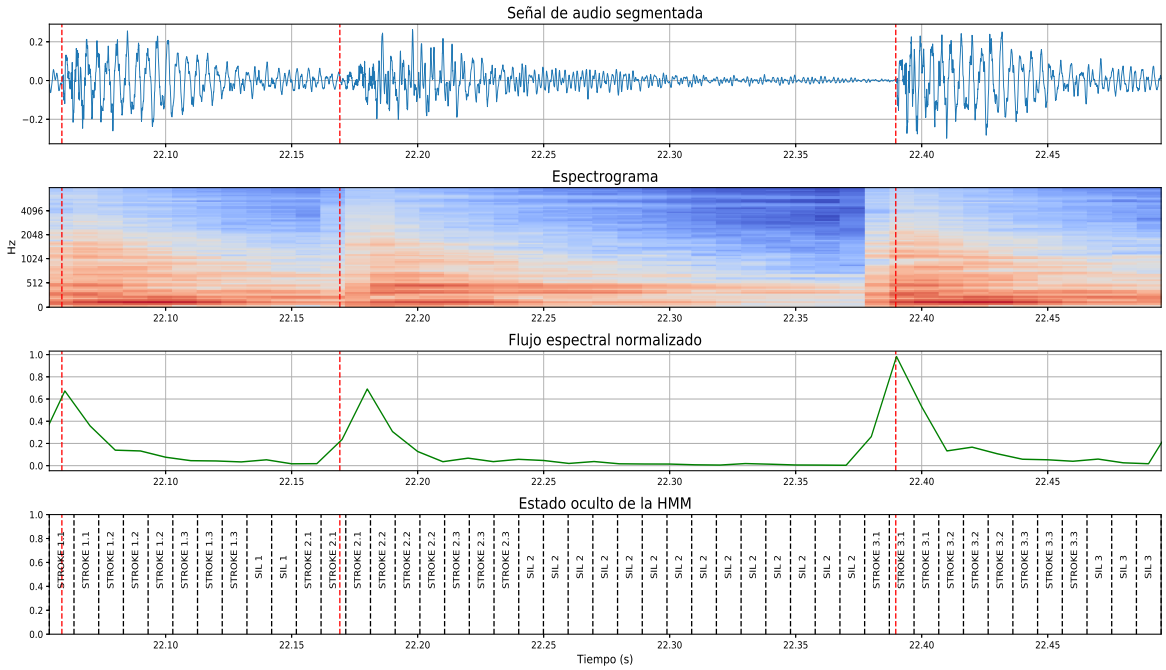

Figura 5: Señal de entrenamiento segmentada, espectro en escala mel, flujo espectral y división en tramas, con estados ocultos etiquetados. Las líneas punteadas rojas marcan la ubicación de los golpes.

REFERENCIAS

[1] Luis Jure, "Principios generativos del toque de repique del candombe," in *La música entre África y América*, C. Aharonián, Ed., Montevideo, Uruguay, 2013, Centro Nacional de Documentación Musical Lauro Ayestarán, pp. 263–291.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[3] Leonard E. Baum, John Alonzo Eagon, et al., "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc*, vol. 73, no. 3, pp. 360–363, 1967.

[4] Christopher J. Leggetter and Philip C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech & language*, vol. 9, no. 2, pp. 171–185, 1995.