Clasificador de canciones según géneros musicales

Gianfranco Fagioli, Victor Franco Matzkin y Gaspar Ezequiel Oberti
Trabajo práctico final de "Procesamiento Digital de Señales", II-FICH-UNL.

***Resumen*— Los géneros musicales permiten agrupar la música en conjuntos de canciones con características similares relacionadas a la señal de audio, que son percibidas por el oído humano y nos permiten poder distinguir en base a nuestra experiencia entre un género y otro. En este trabajo, teniendo en cuenta artículos realizados por Tzanetakis [1] y Scheirer [2], se trata de imitar la habilidad de distinguir entre géneros, pero utilizando características propias de la señal, que se pueden obtener a partir de mediciones en el dominio temporal y frecuencial. Haciendo uso de un clasificador estadístico, luego se procede a estimar el género de un conjunto de canciones, teniendo una tasa de efectividad del 72 %.**

***Palabras clave*— géneros, extracción de características, ritmo, clasificación musical**

I. Introducción

La clasificación de canciones puede ser una tarea difícil. Puede hacerse en base a la información del disco en el que fue publicado o bien agrupando canciones que posean características musicales similares. Aquí se busca una forma de llegar a un conjunto de valores para cada canción que permitan distinguir, en todo el conjunto de canciones, a qué género pertenece. El concepto de género puede variar, ya que el *género* como tal es una forma de categorizar a las canciones a partir de una determinada percepción del ser humano. Esto quiere decir que es posible inventar nuevos subgéneros en base a géneros conocidos o generalizar en base a otros ya existentes, pero teniendo siempre en cuenta que deben poseer diferencias auditivas perceptibles. Debido a esta subjetividad en el concepto de género, no se puede hacer una predicción con absoluta efectividad a través de las técnicas aquí utilizadas. En este trabajo, se propone, a partir fragmentos de señales digitales de audio que son ventaneadas, realizar mediciones generales y hallar posteriormente un valor representativo para el fragmento en base a los valores en los que varía una característica determinada (en este caso usamos la media y la varianza de cada característica). El formato de los archivos de audio es el mismo para todas las canciones (22050 Hz, .wav), por lo que se procesarán de la misma forma sin importar de qué archivo se trata. Además se propone una forma de extraer características rítmicas de las canciones a partir de relaciones calculadas según la morfología de un histograma que indica la variación de una estimación de los BPM (golpes por minuto) de cada ventana del recorte. Posteriormente, las características se normalizan, quedando listas para ser procesadas por un clasificador.

II. Extracción de Características

La base de este trabajo está en el proceso de extracción de características, que consiste en caracterizar un segmento de audio mediante una representación numérica compacta. Se desarrollarán dos tipos de características: de timbre y de ritmo.

A. Características del timbre

El timbre es una de las cuatro cualidades que caracteriza al sonido, y es el que nos permite diferenciar dos sonidos de igual frecuencia fundamental e intensidad (por ejemplo, la misma nota tocada con dos instrumentos diferentes). Para representar la textura del timbre, se realiza un análisis de Fourier para un conjunto de ventanas de 512 muestras (a 22Khz son 43 ms), y según la necesidad se utilizan los datos del dominio temporal o frecuencial. Por lo tanto, una vez obtenidas las ventanas en el dominio del tiempo y sus respectivos espectros de frecuencias, se obtiene:

1. Centroide Espectral: Se define como el *centro de masa* del espectro de magnitudes, y es una medida del *brillo* del sonido (si se hace una analogía con el brillo de una imagen).

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}$$

   donde $M_t[n]$ es la magnitud de la TDF en la ventana t y en la muestra n, y N es la cantidad de muestras en esa ventana.
2. Roloff Espectral: El *roloff* espectral es un punto que se define como el enésimo percentil de la distribución espectral de potencia (N generalmente es 85 o 95 %). El punto de roloff es la frecuencia por la cual se concentra el $N\,\%$ de la magnitud de la distribución.

$$C_t = \sum_{n=1}^{R_t} M_t[n] = 0{,}85 * \sum_{n=1}^{N} M_t[n]$$

   En este caso N corresponde al $85\,\%$. de la distribución.
3. Flujo espectral: El flujo espectral es una medida de cuánto cambia la potencia del espectro de una ventana a otra. Éste se calcula comparando la potencia del espectro de una ventana de la señal contra la potencia del espectro de la ventana anterior, mediante la norma 2. Esta medida es independiente de la potencia media, ya que se trabaja con espectros normalizados.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2$$

4. Cruces por cero en el dominio del tiempo: Da una medida del ruido en general de la señal.

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |signo(x[n]) - signo(x[n-1])|$$

5. Coeficientes cepstrales en la escala de Mel (MFCC): Partiendo de espectro frecuencial de la señal normalizada, se lo pasa por un banco de filtros lineal en escala de mel (de manera que al ir a la escala en Hz se tenga más contenido relacionado a las bajas frecuencias, tal como lo hace el oído humano). Al resultado obtenido,se lo pasa a escala logarítmica, se le aplica la transformada de coseno discreta y una vez en el dominio de las cuefrencias se realiza un liftrado de tantos coeficientes como se necesite. A diferencia del análisis de voz, aquí se toman los primeros 5 coeficientes para el vector de características, ya que según Tzanetakis[1] de esta forma se provee el mejor rendimiento para la clasificación.

Hasta este punto, se tiene para cada ventana un conjunto de 9 características, por lo que resta tomar la media y la varianza de cada característica usando todas las ventanas, duplicando entonces la cantidad de características.

B. *Características del ritmo*

El ritmo, definido por Clarke como *fenómenos temporales de pequeña y mediana escala* puede ser una herramienta muy útil cuando se quiere clasificar canciones. Se define por el *Tempo* (una medida de la rapidez con la que fluye el ritmo). Esto se mide en golpes por minuto (BPM), de manera que 1 Hz representan 60 BPM. Es por esto que en este trabajo el análisis del ritmo se realiza en fragmentos más largos que en el caso anterior, en el que se utilizaban ventanas de unos pocos milisegundos.

Scheirer propone en su trabajo[2] una forma para extraer el tempo en BPM para un fragmento, a través del uso de una serie de filtros para separar la señal en varias bandas de frecuencia, calcular la envolvente, sumar las señales, hacer la autocorrelación y arrojando un solo valor final, que será el pico máximo de esta autocorrelación. Tzanetakis por otro lado propone utilizar la transformada Wavelet Discreta Diádica (DDWT) para obtener esta división de las bandas de frecuencia, por lo que se siguió este enfoque, calculando mediante una función la división en 6 bandas (1 vector de coeficientes de aproximación y 6 vectores de coeficientes de detalle). A partir de los coeficientes de detalle, se halla la envolvente. Al hacer la DDWT se puede observar que los coeficientes de detalle para cada banda poseen cantidades de muestras diferentes (la última banda tendrá N/2 muestras, la penúltima N/4 y así sucesivamente), por lo que se realiza un sobremuestreo antes de seguir con la detección de la envolvente. La envolvente se extrae con el objetivo de obtener la morfología de la banda que se está analizando y consta de una serie de operaciones que se le realizan a esta banda:

1. Rectificación de onda completa: Se pasa la amplitud negativa de la señal a las amplitudes positivas.

$$y[n] = |x[n]|$$

2. Filtrado pasa bajos: En este paso se suaviza la señal rectificada para eliminar componentes de ruido.

$$y[n] = (1 - \alpha)x[n] + \alpha y[n - 1]$$

donde α en este caso vale 0.99.

3. Submuestreo: Debido a que los resultados de la autocorrelación son los mismos si se submuestrea (hasta un cierto límite), se realiza un submuestreo para reducir la cantidad de cálculos.

$$y[n] = x[kn]$$

4. Remover la media: Se realiza para que la señal esté centrada en cero a la hora de aplicar la autocorrelación.

$$y[n] = x[n] - E[x[n]]$$

Donde $E[x[n]]$ es el valor esperado de $x[n]$.

Una vez obtenida la envolvente para cada banda, éstas se sobremuestrean (ya que el algoritmo de la DWT proporciona bandas con diferentes tamaños) para tener el mismo tamaño, se suman y posteriormente se realiza la autocorrelación:

$$y[k] = \sum_n x[n]x[n - k]$$

El resultado de la autocorrelación es una señal que tendrá picos en las posiciones correspondientes a donde la señal es mas parecida a sí misma. Se obtienen nuevas características a partir de esta señal:

- Amplitudes relativas del primer y segundo pico (amplitud sobre la suma de amplitudes).
- Relación de amplitud entre el segundo pico dividido por la amplitud del primero.
- Período del primer y segundo pico (en qué momento ocurren).
- Suma de la señal autocorrelacionada.

Estas 6 características se agregan a las características anteriores, quedando 24 en total.

III. Clasificación

Para la clasificación se usa un clasificador basado en el análisis discriminante. Este análisis consiste en describir las diferencias entre grupos de datos u objetos que poseen una cierta cantidad de variables, asumiendo que las diferentes clases poseen características relacionadas basándose en distribuciones Gaussianas diferentes. El conjunto de datos utilizado GTZAN se compone de recortes de canciones provenientes de grabaciones de CD, radio y micrófono, con el objetivo de poder simular la mayor cantidad de condiciones posibles. El *dataset* posee 500 canciones, las cuales pueden ser de 5 géneros diferentes (clásico, rock, hip-hop, reggae, metal), teniendo 100 canciones para cada género. De éstas, para realizar las pruebas se tomaron 90 canciones para entrenar el clasificador y 10 para probarlo posteriormente.

IV. Resultados

Con el conjunto de datos dado, se tuvo un porcentaje de aciertos del 72 %. La matriz de confusión es la siguiente:

TABLA I
MATRIZ DE CONFUSIÓN DE LOS GÉNEROS.

	cla	dis	hip	reg	roc
cla	9	0	0	0	1
dis	0	6	4	0	0
hip	0	1	8	1	0
reg	1	1	3	4	1
roc	1	0	0	0	9

Los aciertos de la clasificación se pueden ver en la diagonal y los errores en el resto de la matriz.

V. Conclusiones

A este punto, se pudo comprobar lo mencionado anteriormente relacionado a la subjetividad del concepto de género, esto quiere decir que, para este tipo de análisis, los límites de donde termina un género y comienza el otro no están perfectamente definidos como para hacer una clasificación exacta o casi humana. Se considera que el porcentaje de aciertos obtenido es relativamente bueno teniendo en cuenta que si se determina un genero al azar entres los posibles 5 géneros en este caso, se obtiene una media esperada de 20 % de aciertos.

Referencias

[1] G. Tzanetakis, *"Musical Genre Classification of Audio Signals"*, in *IEEE Trans Speech Audio Processing*, vol. 10, no. 5, pp. 293-302. Jul. 2002.

[2] E. Scheirer, *"Tempo and beat analysis of acoustic musical signals"*, in *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588-601. Jan. 1998.