C4

INGENIERÍA DE CARACTERÍSTICAS PARA CLASIFICACIÓN DE SEÑALES SONORAS

Gómez Bellido, Jesús. Luque Sendra, Amalia. Carrasco Muñoz, Alejandro. Grupo de investigación TEP022 y TIC150. Departamentos de Ingeniería del Diseño y Tecnología Electrónica. Universidad de Sevilla.

RESUMEN

Muchos autores han demostrado que el canto de anuros puede ser un indicador del cambio climático. Por esta razón, la clasificación del sonido del canto de anuros se ha convertido en un tema importante para biólogos y otros científicos del clima.

A lo largo de este artículo se describirá la investigación que se llevará a cabo para encontrar el mejor camino de realizar una clasificación automática de los cantos de anuros.

Palabras claves: *características, dimensionalidad, sonido, clasificación, secuencias.*

ABSTRACT

Several authors have shown that the sounds of anurans can be used as an indicator of climate change. For this purpose anuran sound automatic classification has become an important issue for biologists and other climate scientists.

In this paper, we will describe the research to find the best way to perform an automatic classification of anurans sounds.

Keywords: *feature, dimensionality, sound, classification, sequences.*

INTRODUCCIÓN

El cambio climático se está convirtiendo en una de las preocupaciones más exigentes para toda la humanidad. Por esta razón, se están definiendo y monitoreando muchos indicadores para observar la evolución del calentamiento global. Algunos de estos indicadores tienen que ver con el impacto del calentamiento de la biosfera, midiendo el cambio en la población de algunas especies animales.

De hecho, la producción de sonido en animales ectotérmicos está fuertemente influenciada por la temperatura ambiente (Marquez and Bosch 1995) que puede afectar varias características de su sistema de comunicación acústica. Como resultado, la temperatura puede afectar significativamente los patrones de llamada de canciones modificando el inicio, duración e intensidad de los episodios de llamada.

Por lo tanto, el análisis y la clasificación de los sonidos producidos por ciertas especies animales ha destacado como indicador de los cambios de temperatura. Particularmente interesante son los resultados obtenidos por el análisis de sonidos de anuros (Llusia et al. 20139.

En trabajos previos (Luque et al. 2016), se ha propuesto un método automático para la clasificación de sonidos. De acuerdo a estos trabajos, se divide el sonido en pequeños segmentos uniformes, llamados *frames*, y a cada uno de ellos se le caracteriza con 18 parámetros usándose la norma MPEG-7 para la definición. Sin embargo, el enfoque de hacer uso del conjunto de características MPEG-7 no es el único que se puede aplicar para la clasificación de sonidos.

Probablemente el enfoque más común para la extracción de características del sonido sea el basado en los *Mel Frequency Cepstral Coefficients* (MFCC), los cuales derivan del cepstrum

del sonido. En (Young et al. 2002) se describe un algoritmo detallado para la obtención de los MFCC.

METODOLOGÍA

La ingeniería de características es un campo de la minería de datos (inteligencia artificial, aprendizaje automático, ...) que trata de procesar la información en bruto que se posee sobre un determinado problema y transformarla en un conjunto de parámetros (características) que puedan ser utilizados más fácilmente en tareas posteriores (clasificación, predicción, etc).

En el contexto de este trabajo, la clasificación de sonidos, la ingeniería de características tiene tres etapas fundamentales:

1. Extracción de características. A partir de una señal sonora se abordará la obtención de un conjunto de características mediante dos enfoques:
    - MFCC
    - MPEG-7
2. Construcción de características. Partiendo de las características anteriores, se elaborarán nuevas características que puedan reflejar el carácter secuencial de los sonidos.
3. Selección de características. Se seleccionan las características más relevantes para la clasificación de los sonidos del conjunto obtenido en los procesos anteriores.

RESULTADOS Y DISCUSIÓN

Extracción de características

Partiendo de la información del sonido sin procesar, magnitud del sonido en sucesivos instantes de tiempo *(s_i).* El volumen de información puede llegar a ser, en muchos casos, muy elevado y con información redundante. En esta etapa se aplicarán diversas técnicas de representación del sonido que reducen el número de valores.

- Por ejemplo, una muestra de sonido de una duración de 10 segundos, muestreada a 44,1 kHz, supone exactamente 444.100 valores.
- Para la mayoría de las técnicas, se divide el sonido en *frames* de una duración determinada. Siguiendo con el ejemplo anterior, para una duración del *frame* de 10 ms, se generan un total de 441 valores en cada *frame* y un total de 1.000 *frames*.

La mayoría de técnicas extrae un conjunto de características *(e)* de cada *frame* del sonido.

- Siendo *N* el número de valores que contiene un *frame* y S_j el conjunto de valores de sonido de j-ésimo *frame*. Se tiene que,

$$Sj = s_k \forall (j-1) \cdot N + 1 \leq k \leq j \cdot N$$

- El conjunto e_j de parámetros extraídos del j-ésimo *frame* se calcula como función de los valores del sonido en ese *frame* $e_j = f(S_j)$**.**
- Por lo tanto, continuando con el ejemplo anterior, si utilizamos 18 parámetros para representar un *frame*, el número total de valores necesarios para describir el segmento de 10 segundos es de 18.000 valores, un 4% del total original.

Extracción de características MPEG-7

La norma MPEG-7, es un estándar (Carpentier 2005) de la organización internacional para la estandarización, desarrollado por el grupo MPEG, grupo que se encarga de desarrollar normas para la representación codificada del audio y el vídeo digital, que define un conjunto de

parámetros para la representación de un sonido. Estos parámetros se obtienen de diversos métodos de análisis, como son:

- Procesamiento en el dominio del tiempo
- Procesamiento en el dominio de la frecuencia
- Codificación predictiva lineal (LPC)

MPEG-7 no está dirigido a ninguna aplicación en particular, simplemente es un estándar para la descripción del material multimedia: habla, audio, vídeo, imágenes y modelos 3D.

Extracción de características MFCC

La obtención de los Coeficientes Mel Cepstral *(MFCC, Mel-Frequency Cepstrum Coefficients)* ha sido considerada como una de las técnicas de parametrización de la voz más importante y utilizada (Davis and Mermelstein 1980). El objetivo de esta transformación es obtener una representación compacta, robusta y apropiada para posteriormente obtener un modelo estadístico con un alto grado de precisión. En la Figura 1 se recoge el resultado de la extracción de parámetros MFCC para un sonido.

Uno de los inconvenientes que tienes los parámetros MFCCs es que su método de obtención no está normalizado, pudiendo haber ligeras variantes entre una y otra implementación. Aunque no es de ámbito general, la norma (Standard 2003) se define como una forma de obtención de estos parámetros para telefonía. Durante nuestros estudios este será, con alguna adaptación, el método empleado.

La adaptación realizada se debe al hecho de que la norma utiliza *frames* de 25 ms, a una frecuencia de muestreo de 16 kHz. Esto supone que cada *frame* está formado por 400 valores. En nuestro caso, la frecuencia de muestreo será de 44,1 kHz, no contemplada en la norma. Por ello se propone adaptar la duración del *frame* a un valor de 10 ms, la misma que en MPEG-7, lo que supone que estará formado por 441 valores.

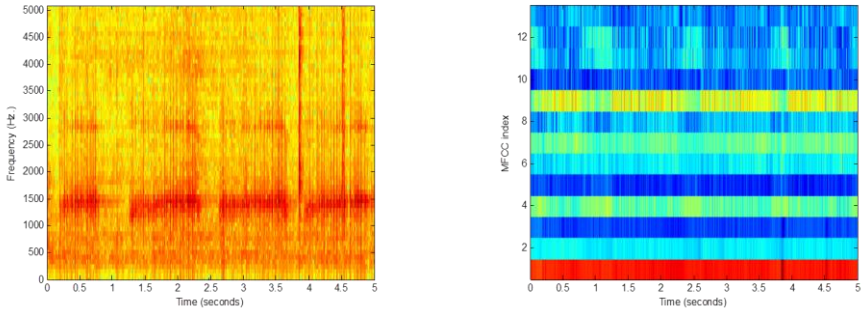

Figura 1: Espectograma (izquierda) y parámetros MFCC (derecho).

Construcción de características

En esta etapa, se parte del conjunto de características e, obtenido en la extracción de características.

Durante esta etapa, se opera en ellos con distintas transformaciones, construyendo una o varias características nuevas, para el j-ésimo frame, se tiene $c_j = f(e)$.

- Es típico el uso de parámetros de tendencias, por ejemplo, los **ΔMFCC** o los **ΔΔMFCC** *(*Sharma et al. 2014).
- También pueden generarse características que contengan información regional del sonido, como puede ser la variación de un parámetro en un microsegundo en torno a un *frame.*

Estas características construidas se añaden a las extraídas, para construir el conjunto de características π que representa un sonido. Por tanto, se tiene para el j-ésimo *frame*, $\pi = e_j \cup c_j$.

Selección de características

En numerosas ocasiones el conjunto de características π obtenidas en las etapas anteriores tiene un cierto grado de redundancia, por lo que puede ser reducido sin grave deterioro de la potencia representativa de los mismos.

Por ello, se puede seleccionar un subconjunto *p* de características de tal forma que, para el j-ésimo *frame,* $p_j \subset \pi_j$.

CONCLUSIONES

En este artículo, se han descrito las etapas fundamentales de las que consta la ingeniería de característica y que serán objeto de estudio durante esta tesis, además de los dos principales métodos de extracción de características que se desarrollarán, MFFC y MPEG-7.

Para concluir, se ha de mencionar que esta tesis partirá de estudios previos realizados por los grupos de investigación, por lo tanto, se pretende dar mayor formalidad a estos estudios en el apartado de la ingeniería de características.

AGRADECIMIENTOS

A los grupos de investigación TEP022 y TIC150 de la Universidad de Sevilla por los trabajos previos a esta tesis.

A la Escuela Politécnica Superior de Sevilla por la organización de la Jornadas Doctorales que ha permitido la publicación del presente trabajo.

BIBLIOGRAFÍA

Carpentier, G. (2005). Information technology—multimedia content description interface—part 4: Audio, amendment 2: High-level descriptors. In Motion Picture Expert Group (ISO/IEC JTC 1 SC29.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4):357–366.

ETSI. 202 050 v1. 1.3: Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI standard, 2002.

Llusia, D., Márquez, R., Beltrán, J. F., Benitez, M., and Do Amaral, J. P. (2013). Calling behaviour under climate change: geographical and seasonal variation of calling temperatures in ectotherms. Global change biology, 19(9):2655–2674.

Luque, J., Larios, D. F., Personal, E., Barbancho, J., and León, C. (2016). Evaluation of mpeg-7-based audio descriptors for animal voice recognition over wireless acoustic sensor networks. Sensors, 16(5):717.

Márquez, R. and Bosch, J. (1995). Advertisement calls of the midwife toads alytes (amphibia, anura, discoglossidae) in continental Spain. Journal of Zoological Systematics and Evolutionary Research, 33(3-4):185–192.

Sharma, S., Shukla, A., and Mishra, P. (2014). Speech and language recognition using mfcc and delta-mfcc. International Journal of Engineering Trends and Technology (IJETT), 12(9):449–452.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. Cambridge university engineering department, 3:175.