Detección de pitch: un entrenador de canto

Juan J. Bertinetti y M. Virginia Gómez
Trabajo práctico final de "Procesamiento digital de señales", II-FICH-UNL.

***Resumen*—Muchas personas quieren entrenarse en el canto y necesitan de una forma objetiva para saber si están entonando bien o no. Aquí se presenta un Entrenador de Canto que analiza la entonación del usuario según una nota o melodía objetivo o de prueba, basándose en calcular el *pitch* o frecuencia fundamental del canto a lo largo del tiempo, y comparándolo con la frecuencia de las notas musicales que se deben entonar.**

Se detalla el método usado para hacer esta detección del pitch, así como los problemas encontrados.

***Palabras clave*—canto, pitch, frecuencia fundamental, armónicos, voz humana.**

I. Introducción

Se presenta aquí un Entrenador de Canto, que calcula la entonación del canto grabado del usuario a lo largo del tiempo, y lo compara con una nota o melodía de prueba, presentando resultados visuales y cualitativos sobre su performance (qué tan buena fue su entonación).

Se usa para el mismo el sistema internacional de notación musical, en el cual un La de 440 Hz se encuentra en la cuarta octava (La4 o A4). El DO0 (Do en la octava 0) está en la región de la frecuencia más baja audible cerca de los 16 Hz.

En este sistema, la frecuencia de una nota puede ser calculada mediante la fórmula:

$$f(n,o) = 440\, e^{\left((o-4)+\frac{(n-10)}{12}\right)\ln(2)} \qquad (1)$$

donde n es un número de 1 a 12 (escala cromática, 12 semitonos en una octava) y o es un número de 0 a 10.

Para el entrenamiento, se generan notas o melodías objetivo o de prueba, mediante tonos puros con las frecuencias correspondientes a las notas deseadas según (1). Luego el usuario trata de entonarlas a la vez que graba su canto. El Entrenador de Canto analiza luego la grabación según el objetivo y muestra una representación visual de la entonación a lo largo del tiempo.

Como expone [1], hay una gran variedad de métodos para detectar el pitch: métodos en el dominio del tiempo (tasa de eventos temporales, autocorrelación, espacio de fase), en el dominio de la frecuencia (métodos basados en Fourier, basados en filtros, basados en el cepstrum) y estadísticos en el dominio de la frecuencia (máxima verosimilitud y redes neuronales). Aquí se decidió implementar uno basado en una combinación de 2 transformadas de Fourier.

Para la implementación, se trabajó con señales muestreadas a 22050 Hz.

II. Voz humana y armónicos

Se sabe que la voz humana tiene un rico contenido de armónicos. Los armónicos son componentes de frecuencias que son múltiplos enteros de la frecuencia fundamental de la señal.

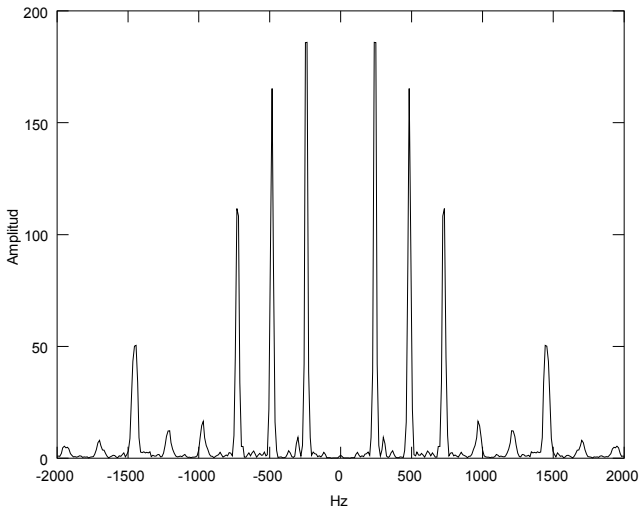

Fig. 1: Espectro de una señal de voz

Si se hace la transformada de Fourier a una señal de voz, se puede observar una serie de picos dominantes equiespaciados que corresponden a los armónicos (Ver Fig. 1), donde el pitch de la señal corresponde a la frecuencia del primero de estos picos (en las frecuencias positivas), es decir, la frecuencia fundamental es la frecuencia del primer armónico.

En general, la amplitud de estos picos tiende a cero, siendo la amplitud de los armónicos más altos mucho menor que la amplitud de la fundamental. Por lo tanto para hallar el pitch bastaría con buscar el máximo en el espectro. Sin embargo en ocasiones la amplitud de los armónicos suele ser mayor, y hasta puede suceder que la fundamental no se encuentre, que sea "virtual" [2].

III. Detección del pitch

Para hacer la detección del pitch a lo largo del tiempo de la melodía cantada, nos basamos en el método propuesto por [2]. El mismo consiste en hacer una combinación de 2 transformadas de Fourier, por lo tanto lo llamaremos "la transformada de la transformada". El mismo resuelve el problema de que la fundamental no se encuentre.

A. Explicación del método

El método de "la transformada de la transformada" consiste en aplicar la transformada de Fourier al valor absoluto de la transformada de Fourier de la señal limitada a las frecuencias positivas. De esta forma, tomando el valor absoluto de esta segunda transformada (llamaremos a esto TF(TF)), obtenemos una serie de picos decrecientes en amplitud también equiespaciados, donde el primero y más prominente (sin tener en cuenta el que aparece en el índice 0) corresponde al pitch (ver Fig. 2). La frecuencia F_0 correspondiente a este pico ubicado en el índice i, es $F_0 = (f_m/2)/i$, donde f_m es la frecuencia de muestreo.

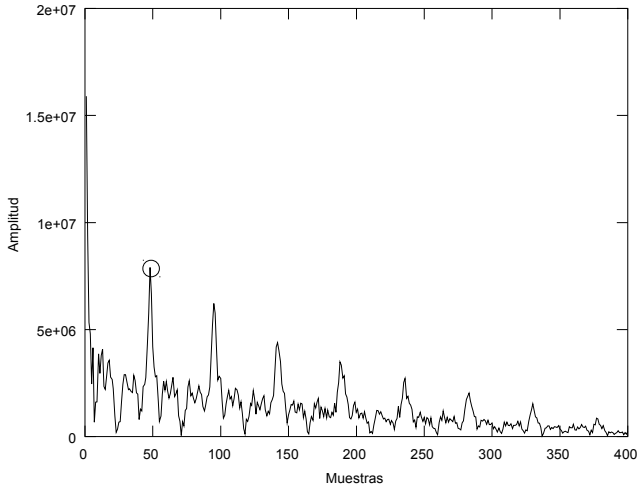

Fig. 2: "Transformada de la transformada" (TF(TF))

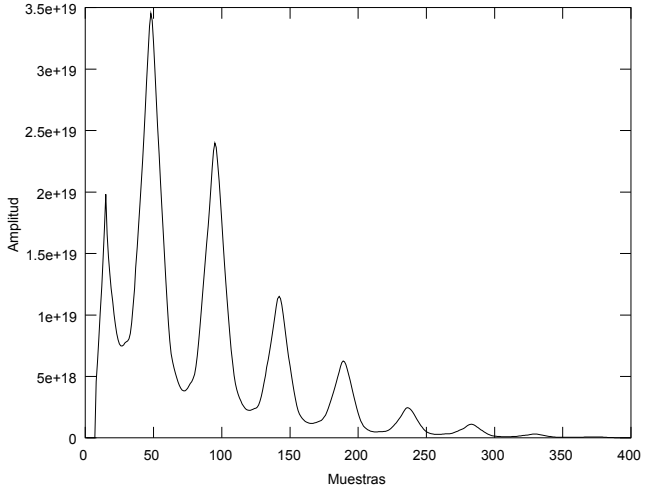

Fig. 3: ft_2: TF(TF) luego del suavizado y elevado al cubo

Para mejorar la extracción de los picos correspondientes a la segunda transformada, [2] propone usar el método FT^n con $n = 1$, o transformada de Fourier de orden 1 [3]. Básicamente consiste en aplicar Fourier no a la señal, sino a la derivada de ésta, aproximada mediante la fórmula $y[n] = f_m \cdot s[n] - f_m \cdot s[n-1]$, donde s[n] es la señal original.

B. Encontrar el pico correspondiente

En algunos casos, el método de "la transformada de la transformada" también tiene el problema de que algunos picos tienen más amplitud que el que necesitamos, por lo que encontrar el máximo local más grande tampoco nos sirve ya que tendremos saltos entre las octavas (detectamos el pitch pero en una octava incorrecta).

Para solucionar el problema, primero de todo aplicamos un proceso de suavizado a TF(TF) con el fin de eliminar los vecinos muy cercanos a los máximos locales que pueden ser confundidos como otros máximos locales. Para esto usamos un filtro que realiza un suavizado triangular de 15 elementos, es decir con la ponderación más fuerte en el centro dejando a los costados valores menores con el fin de fortalecer los valores más altos.

Para hacer más pronunciada la amplitud de los picos con respecto a los valores chicos, elevamos el resultado de este filtrado al cubo (Fig. 3). Llamaremos a esto ft_2.

Luego sólo nos queda encontrar el segundo pico que es el que necesitamos. Para esto establecemos un umbral (8% del valor máximo en ft_2 mostró dar buenos resultados) y procedemos a buscar los máximos locales en ft_2. Si el encontrado está por encima del umbral establecido, lo contamos, en caso contrario buscamos uno siguiente (esto es porque pueden aparecer mini-picos de poca amplitud entre 2 picos reales). Cuando encontramos el segundo máximo local que pasó el umbral, guardamos su índice.

C. Algoritmo

El algoritmo entonces para detectar el pitch en la señal cantada a lo largo del tiempo es el siguiente:

1. Tomamos una cierta cantidad de muestras de la señal original con una ventana (por ejemplo Hanning). En la implementación usamos 2048 muestras que con una frecuencia de muestreo de 22050 Hz corresponde aproximadamente a 92 ms. Ya que lo que se analiza es un canto (un tono sostenido por un largo instante de tiempo) no se nos hace necesario analizar segmentos más chicos (como 20 ó 30 ms), por lo que usamos esta cantidad de muestras que nos permite obtener mejores resultados.
2. Analizamos si la porción de señal corresponde a un sonido sordo o sonoro. Un sonido sordo no posee frecuencia fundamental por lo que debemos descartarlo para la detección del pitch. Para esto medimos si su energía es relativamente baja. Un sonido sonoro posee mucha mayor energía que uno sordo.
3. Realizamos la derivada de la porción de la señal con el fin de aplicar la FT^1.
4. Aplicamos la transformada de Fourier, y nos quedamos con el rango correspondiente a las frecuencias positivas.
5. Aplicamos la transformada de Fourier al valor absoluto de la obtenida en el punto anterior. Para reducir los cálculos subsiguientes, nos quedamos con el valor absoluto sólo de las primeras 140 muestras. Podemos hacer esto ya que la voz humana está normalmente entre el rango de 80 Hz y 1100 Hz (lo que equivale a un MI2 hasta un DO6), estando en un extremo un bajo (cantante lírico masculino) entre un MI2 hasta un DO4, y en el otro extremo un soprano entre un DO4 y un DO6. Por lo tanto, un índice de 140 nos da como mínimo una F_0 de (22050/2)/140 = 78.75 Hz.
6. A las muestras con las que nos quedamos en el punto anterior le aplicamos el suavizado y lo elevamos al cubo.
7. Buscamos el segundo pico teniendo en cuenta el umbral.
8. Tomamos otra porción de la señal y repetimos.

IV. ANÁLISIS DE LA PERFORMANCE DEL CANTO

Teniendo un método robusto para detectar el pitch, el usuario graba su canto basándonse en las notas objetivo y lo analiza.

Para detectar el pitch objetivo, al ser tonos puros sí podemos simplemente buscar el pico en el espectro.

Teniendo la evolución en el tiempo del pitch del canto, comparamos los instantes de tiempo correspondientes a notas separadas entre el canto y el objetivo (si el objetivo es DO3 RE3 MI3, primero comparamos lo correspondiente a DO3). Para determinar de una manera el pitch que el usuario cantó durante ese momento que duró la nota, usamos una medida adecuada como la moda.

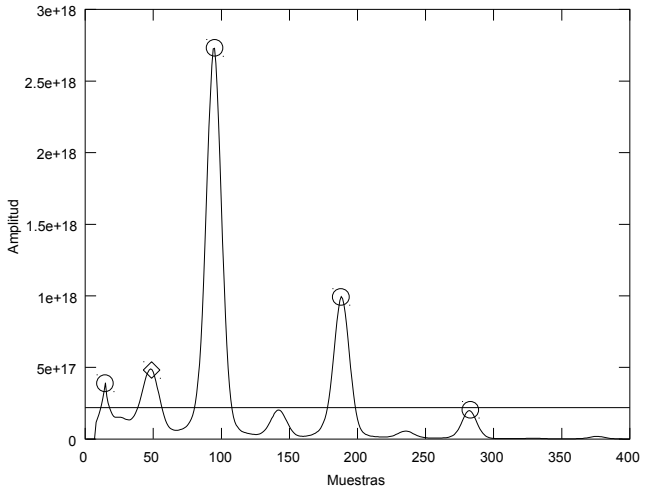

Fig. 4: Círculos: picos correctos. Rombo: pico incorrecto detectado como 2º. Recta: umbral

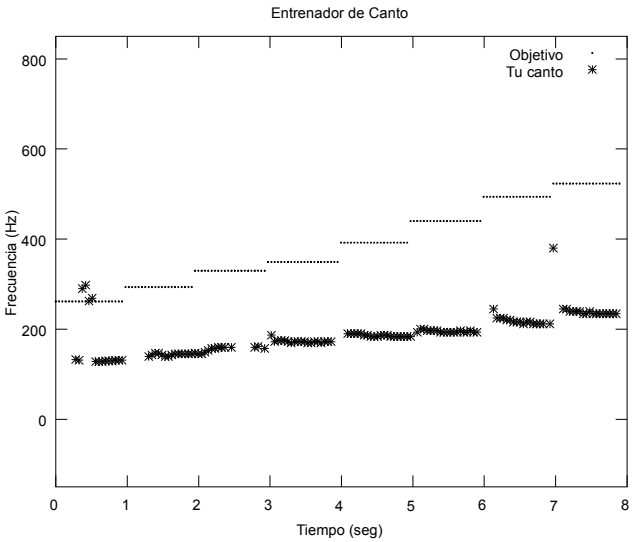

Fig. 5: Resultado al tratar de entonar la octava 4

V. Resultados y discusión

Este método que usamos para encontrar el pico que nos interesa en TF(TF) (suavizar, elevar al cubo y buscar máximos locales teniendo en cuenta un umbral) dio generalmente buenos resultados y resultó ser robusto frente a saltos de octavas.

Sin embargo, todavía hay ocasiones donde puede funcionar mal, donde hay un pequeño máximo local entre 2 picos reales que es relativamente chico pero lo suficientemente grande como para pasar el umbral, por lo que es detectado incorrectamente (Fig. 4). Podríamos solucionarlo especificando un umbral más alto, pero en este caso suele fallar más ya que hay ocasiones donde el primer pico tiene una amplitud muy baja por lo que no pasa el umbral, detectando en este caso al 3º pico como 2º.

Jugando con el valor del umbral, combinado a elevar a potencias más elevadas para realzar los picos importantes, se podría solucionar este problema, pero esto depende mucho de cada señal, por lo que los valores aquí presentados son los recomendados ya que dieron los mejores resultados.

VI. Conclusiones

El método presentado para detectar el pitch del canto resultó ser generalmente robusto. Utilizando el método de la transformada de la transformada propuesto por [2], junto con las estrategias usadas para detectar correctamente el pico de importancia, el pitch es detectado de manera correcta en la gran mayoría de los casos. Cuando no es así, en el resultado visual aparecen puntos aislados que pueden ser ignorados (Fig. 5).

Una persona que quiera entrenarse en el canto y necesite una guía para saber su entonación objetivamente, puede usar el Entrenador de Canto aquí presentado.

VII. Trabajo futuro

Como trabajo futuro se podría tratar de resolver adecuadamente el problema mencionado anteriormente, de manera de ser 100% robusto (o lo más cercano posible).

También está la implementación del Entrenador de Canto para que funcione en tiempo real, de manera de ir teniendo un resultado instantáneo a medida que el usuario canta.

Referencias

[1] D. Gerhard. “Pitch Extraction and Fundamental Frequency: History and Current Techniques”, *Technical Report TR-CS 2003-06*, Noviembre, 2003.

[2] S. Marchand. “An Efficient Pitch-Tracking Algorithm Using A Combination of Fourier Transforms”, *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, Limerick, Ireland, Diciembre 6-8,2001.

[3] M. Desainte-Catherine y S. Marchand. "High Precision Fourier Analysis of Sounds using Signal Derivatives", *Journal of the Audio Engineering Society*, vol. 48, no. 7/8, pp. 654–667, Julio/Agosto 2000.