Identificación de nota musical y familia instrumental

Saul Ivan Rivas Vega

Aprendizaje Automatizado

Contenido

- Introducción
- Extracción de características
- Entrenamiento y evaluación del modelo
- Evaluación extendida
- Conclusión

Introducción

- Nota Musical y Tono

  Una nota musical es el nombre de un tono: Re Do Fa, G C A, etc.

  El tono es como percibimos el sonido con el fin de ordenarlos en una escala

  Asociado a la frecuencia fundamental (f0) y se toman como lo mismo fuera de la psicoacústica

- Timbre

  Nos permite diferenciar fuentes sonoras con el mismo volumen y tono

Kim, J. W., et al. (2018). CREPE: A Convolutional Representation for Pitch Estimation.
Schnupp, J. et al. (2011). *Auditory neuroscience: Making sense of sound*. MIT Press.
Siedenburg, K., et al. (2019). *Timbre: Acoustics, Perception, and Cognition*. Springer

Introducción

- Nota Musical, Tono y Timbre

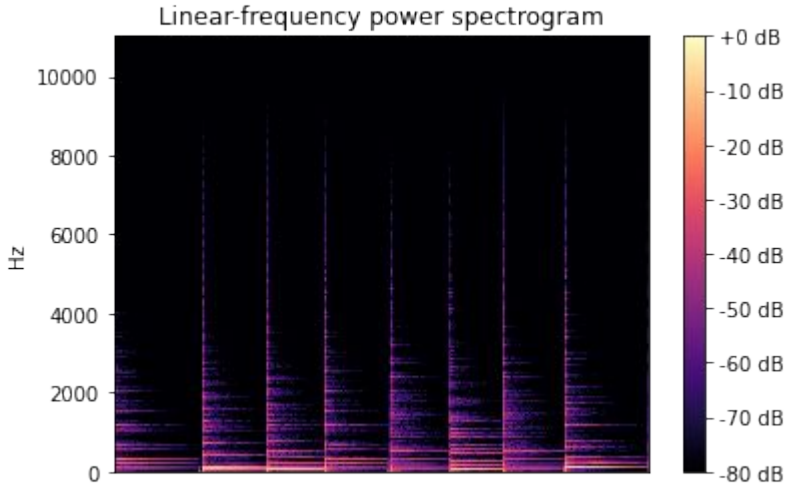
```
{Mix(SinOsc.ar(300*[0.5],0,0.1))}.play
```


```
(
SynthDef(\e,{
    var sig, env;

    sig=Mix(SinOsc.ar(300*[0.5,1,1.19,1.56,2,2.51,2.66,3.01,4.1],0,0.1));

    env= EnvGen.kr(Env.perc(0.1,2),4,doneAction:2);

    Out.ar(0,[sig*env,sig*env]);
}).send(s)
)
```

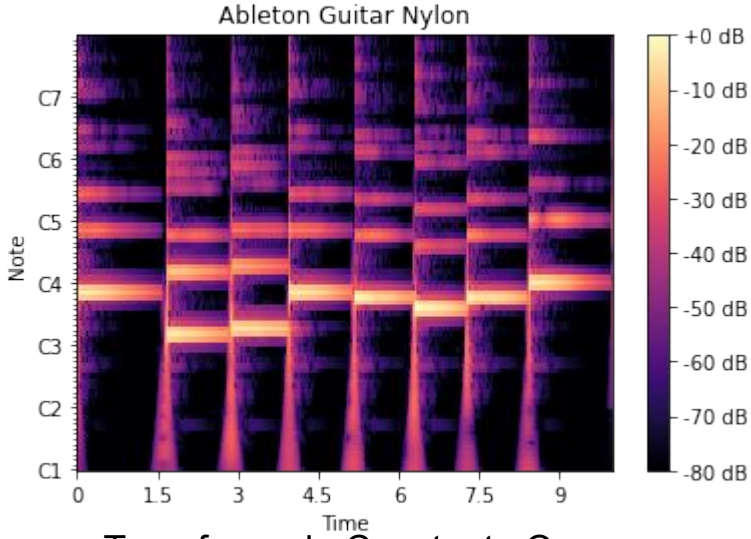

Kim, J. W., et al.. (2018). CREPE: A Convolutional Representation for Pitch Estimation.
Schnupp, J., et al. (2011). *Auditory neuroscience: Making sense of sound*. MIT Press.
Siedenburg, K., et al. (2019). *Timbre: Acoustics, Perception, and Cognition*. Springer

Introducción

- Transformada de Constante Q



Transformada de Fourier



Transformada Constante Q

C1 - 32.70 Hz, C2 - 65.41 Hz, C3 - 130.81 Hz, C4 - 261.63 Hz, …, C7 - 2093.00 Hz

Brown, J. C. (1991). Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, *89*(1)

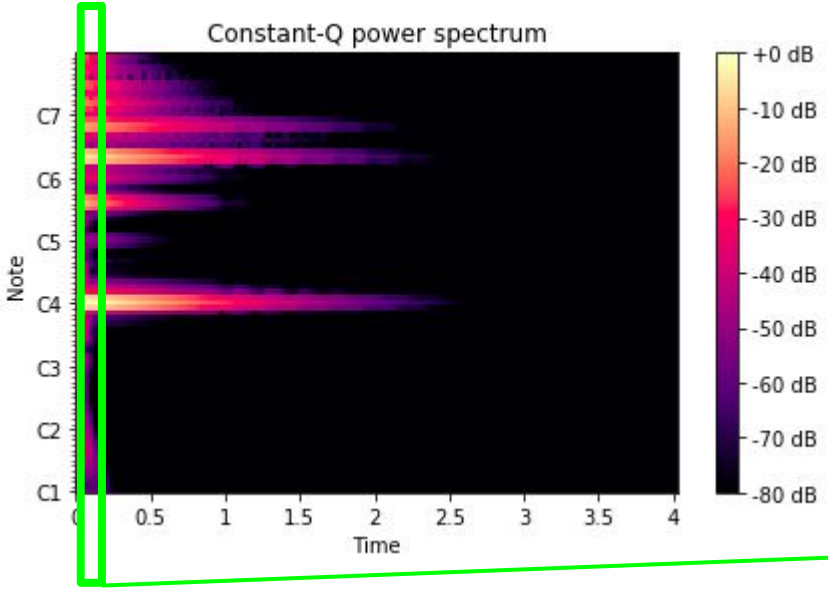
Extracción de características

- Dataset
  - **NSynth** contiene 305,979 clips de audio de notas musicales, obtenidos de 1,006 instrumentos grabando clips de monofónicos de 4 segundos con anotaciones de nota musical en el rango del formato MIDI (21-108) con 5 velocidades (25, 50, 75, 100, 127).
- Un subconjunto de 1554 audios fue utilizado. Hay 14 instrumentos por cada una de las 3 familias (Guitarras, Cuerdas y Vientos) y 37 notas por cada uno desde Do1 (C1) a Do4 (C4).

Engel, J., et al. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders.
McFee, B., et al. (2015). *librosa: Audio and Music Signal Analysis in Python*.

Extracción de características

- Biblioteca Librosa
  - 16 frames por audio, 24864 muestras
  - 224 muestras por categoria (nota-familia)



	segment_name	note_24	note_25
0	guitar_acoustic_001-042-127_seg_0	0.1548396758272465	0.21162235488528283
1	guitar_acoustic_001-042-127_seg_1	0.14865352420076297	0.20248682732713844
2	guitar_acoustic_001-042-127_seg_2	0.13246161845622467	0.17783167815307035
3	guitar_acoustic_001-042-127_seg_3	0.10867733895496602	0.14073855298161575
4	guitar_acoustic_004-033-127_seg_0	0.2665652028886773	0.43013401153111186
...	...	...	...
3251	guitar_acoustic_026-060-127_seg_3	0.3128918375472502	0.31833682526523954
3252	guitar_acoustic_002-042-127_seg_0	0.45726032931360555	0.4228486327314009
3253	guitar_acoustic_002-042-127_seg_1	0.4387334557786869	0.4036912168580886
3254	guitar_acoustic_002-042-127_seg_2	0.38990515722765556	0.35416741550726255
3255	guitar_acoustic_002-042-127_seg_3	0.3178505825897622	0.2831098900920131

Engel, J., et al. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders.
McFee, B., et al. (2015). *librosa: Audio and Music Signal Analysis in Python*.

Entrenamiento y evaluación del modelo

Clasificador Bayesiano Ingenuo

$$Y \leftarrow \underset{y_k}{\arg\max} P(Y = y_k) \prod_{i}^{d} P(X_i | Y = y_k)$$

Estimación por máxima verosímilitud

$$\mathcal{N}(X_i; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}}$$

$$\prod_{i=1}^{k} p_i^{[x=i]},$$

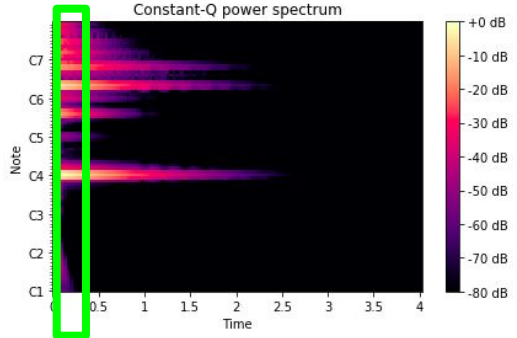
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{q}_k = \frac{c_k}{n}$$

$$\hat{\sigma}_F^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

10 repeticiones de una validación cruzada de 7 particiones. En cada partición las muestras de 12 instrumentos por categoría fueron usados de entrenamiento y 2 para prueba.

note_106	note_107	NOTE_CLASS
7.803614e-07	5.115749e-07	family_string_note_26
2.255030e-04	1.986155e-04	family_string_note_26
4.423557e-04	9.610463e-04	family_string_note_26
9.777358e-04	2.081144e-03	family_string_note_26
1.244309e-03	2.766589e-03	family_string_note_26

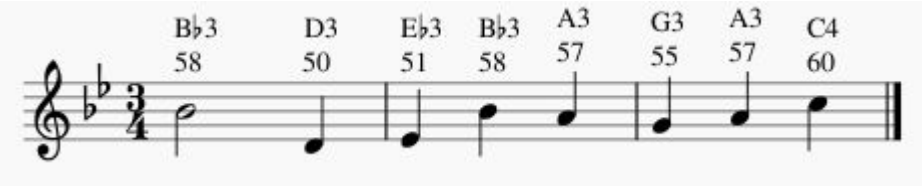


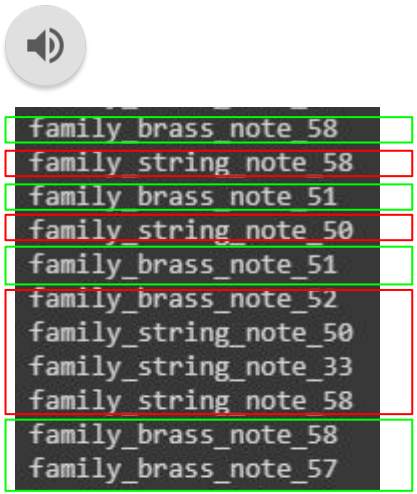
X = (0.213124,...,1.62345) Y = ?

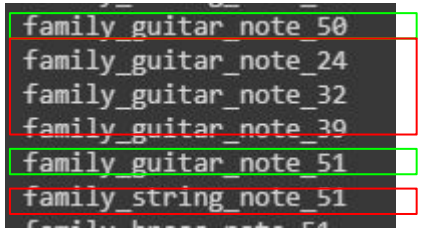
Exactitud: 93.638% en Entrenamiento y 88.616% en Prueba.

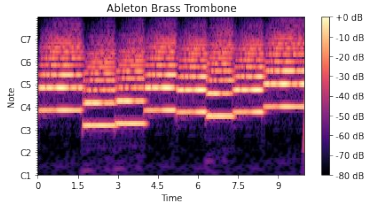
Maron, M. E. (1961). "Automatic Indexing: An Experimental Inquiry". Journal of the ACM.
Narasimha Murty, M.; Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.

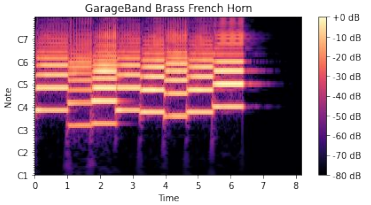
Evaluación extendida

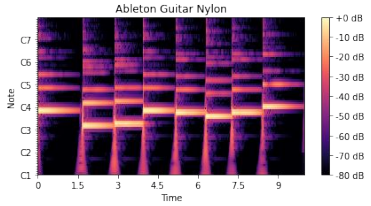


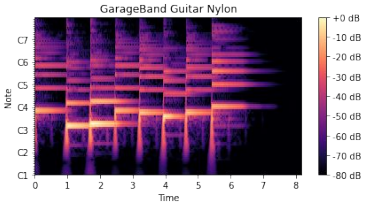


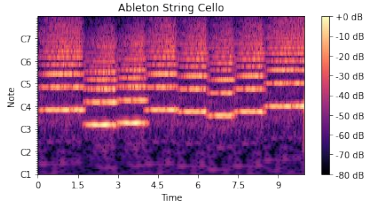


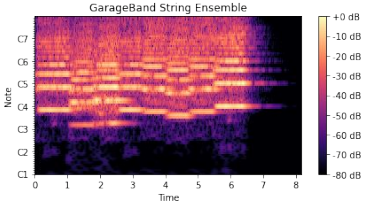












Bastante acertado al identificar las notas presentes.
Hay variaciones cuando la nota es sostenida.
En la guitarra hay variaciones entre notas y en los audios de vientos y cuerdas hay variaciones entre familia.

Conclusión

En general el sistema tuvo buenos resultados en el conjunto de datos sin embargo en un caso de uso más general (aún monofónico, sin ruido y sin reverberación), tiene un comportamiento inestable en el tiempo. Esto se puede explicar por el uso de solo características frecuenciales ignorando la envolvente temporal y otras propiedades del timbre. Por ejemplo pueden ser similares en el momento de las notas sostenidas pero el ataque y decaimiento puede distinguirlos fácilmente.

Aunque se usa información por ventana de muestreo el tiempo que toma de minutos por cada uno no lo hace adecuado para análisis en tiempo real. Su caso de uso podría ser más sobre la búsqueda y organización de audios por notas y familias presentes en ellos.

https://sirv.top/f