



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen



Introduction of Sound, Speech, and Singing Voice

Xueyao Zhang

2024/04

About me

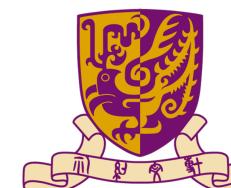


Xueyao Zhang (张雪遙)

- ◆ **Second-year PhD student**, Supervised by Prof Zhizheng Wu
School of Data Science, CUHK-Shenzhen
Homepage: <https://www.zhangxueyao.com/>
- ◆ **Amphion v0.1's co-founder**
Project: <https://github.com/open-mmlab/Amphion> (3.9k stars)
- ◆ **Research interest:** “AI + Music”, especially on:
 - Singing Voice Processing
 - Music Generation

The screenshot shows a Hugging Face Spaces interface for the "Amphion Singing Voice Conversion: DiffWaveNetSVC" demo. The page title is "Amphion Singing Voice Conversion: DiffWaveNetSVC". A hint at the top says: "This demo provides an Amphion DiffWaveNetSVC pretrained model for you to play. The training data has been detailed [here](#)". Below it, a section titled "Source Audio" includes a "Source Audio" button and a note: "Hint: We recommend using dry vocals (e.g., studio recordings or source-separated voices from music) as the input for this demo. At the bottom of this page, we provide some examples for your reference." A large callout box highlights the URL: "SVC Online Demo: https://huggingface.co/spaces/amphion/singing_voice_conversion". The interface features sections for "Target Singer" (with Jian Li 李健 selected), "Key Shift Values" (0), "Diffusion Inference Steps" (1000), "Pitch Shift Control" (Auto Shift selected), and buttons for "Clear" and "Submit".

Let your favorite singer
sing your favorite song!



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

How to create a sound using Python?

- CSC3160: Fundamentals of Speech and Language Processing
 - Lecture 2: [Colab notebook](#)

正弦波是一种周期性波动的数学模型，其公式通常写作：

$$y(t) = A \sin(2\pi ft + \phi)$$

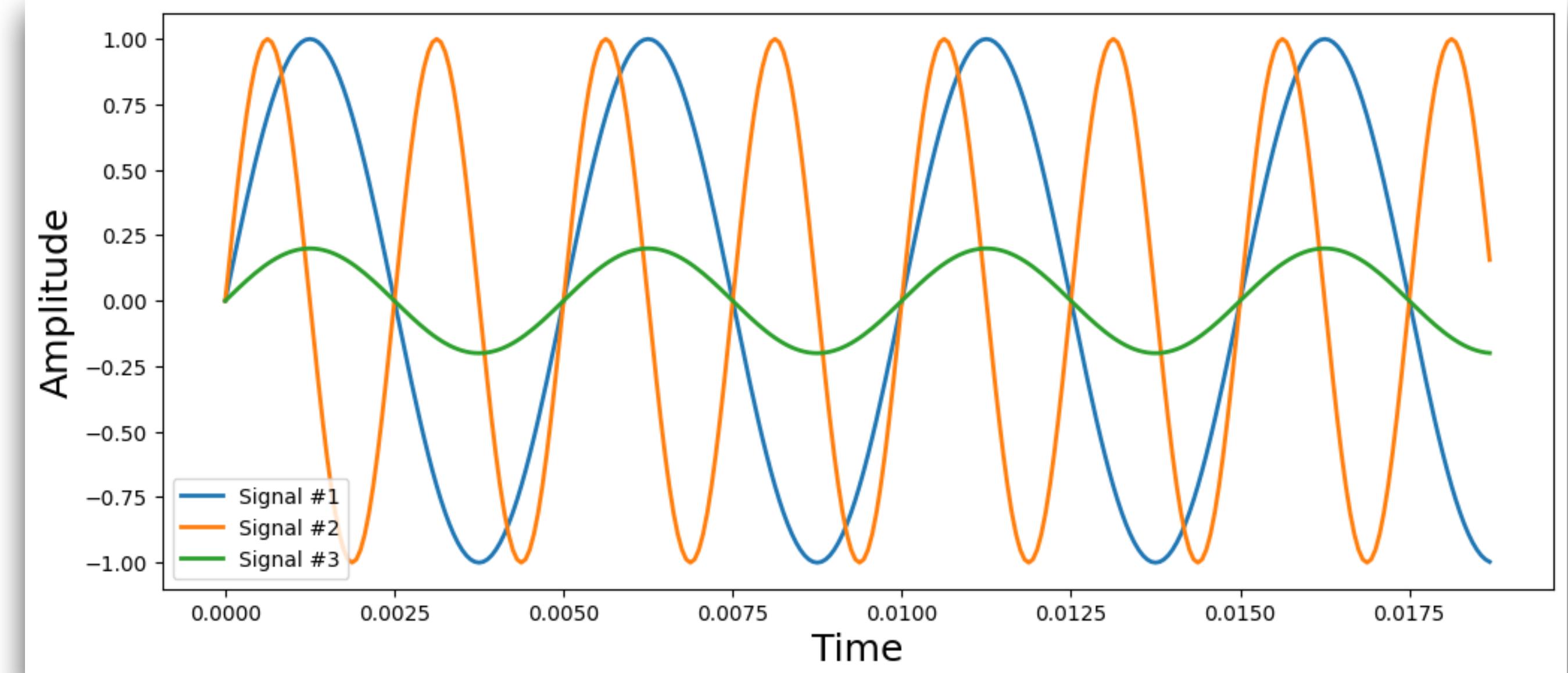
其中：

- $y(t)$ 表示时间 t 时刻的波形值。
- A 是波形的振幅，即波形的最大值和最小值之差的一半。
- f 是频率，表示单位时间内波形重复的次数，单位是赫兹 (Hz)。
- t 是时间，通常以秒为单位。
- ϕ 是相位，表示波形在时间轴上的偏移量，单位是度或弧度。

正弦波在物理学中非常重要，它可以用来描述许多自然现象，如声波、电磁波、机械振动等。

```
# Time points
time = np.arange(beginTime, endTime, samplingInterval);

# Create three sine waves
signal1 = np.sin(2*np.pi*signal1Frequency*time)
signal2 = np.sin(2*np.pi*signal1Frequency*2*time)
signal3 = 0.2*np.sin(2*np.pi*signal1Frequency*time)
```



Three elements of sound: Pitch, Loudness, and Timbre

Perceptual Property

Pitch

Loudness

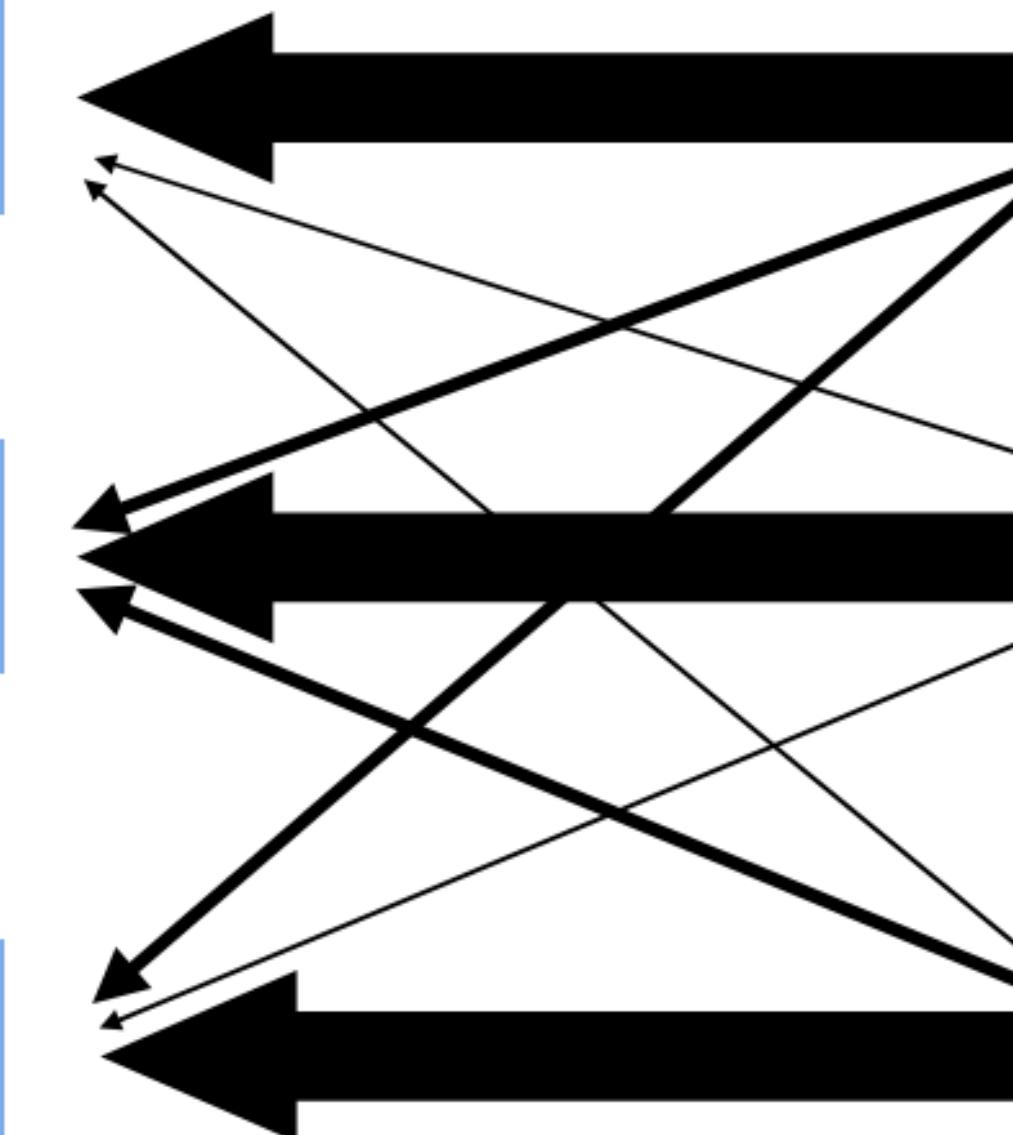
Timbre

Physical Property

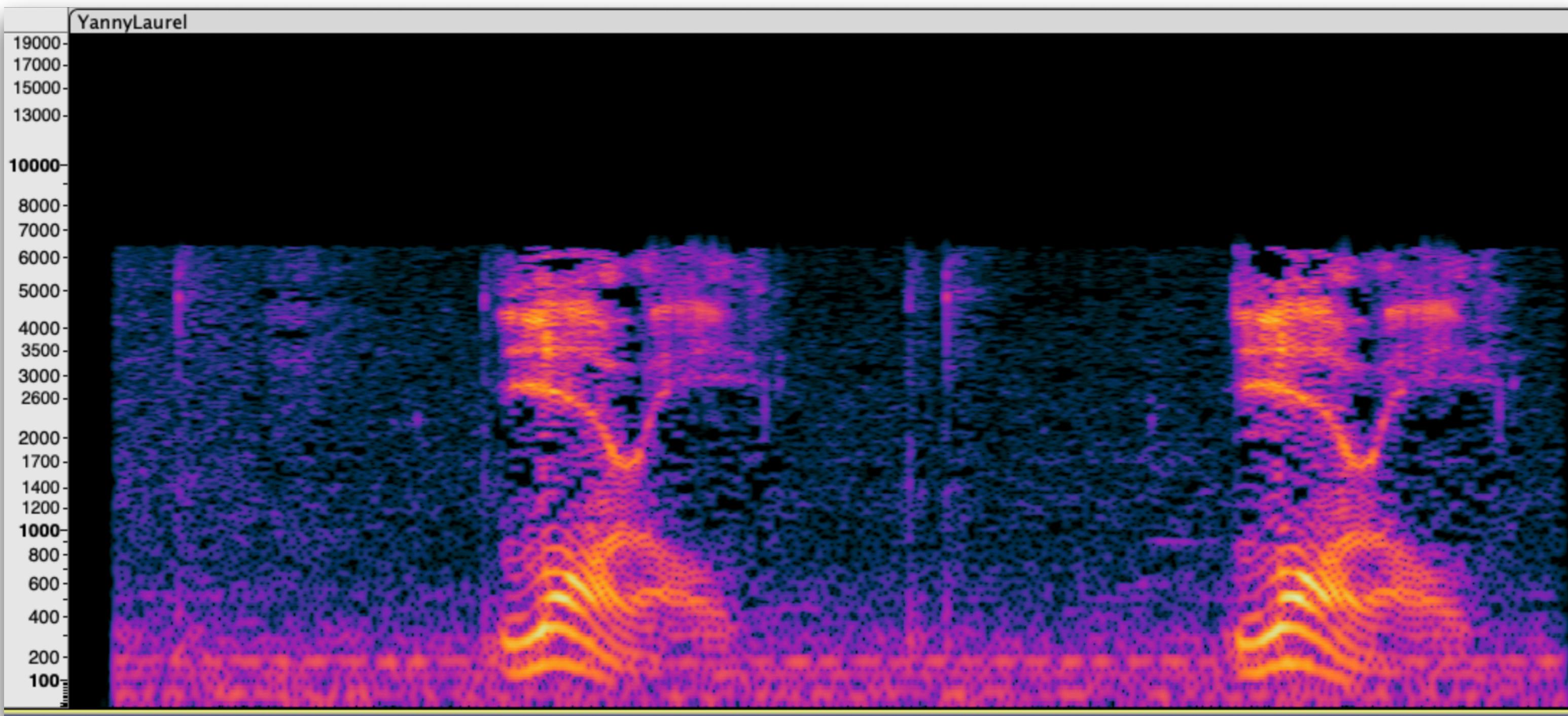
Frequency

Intensity

Time variation
&
Spectral content



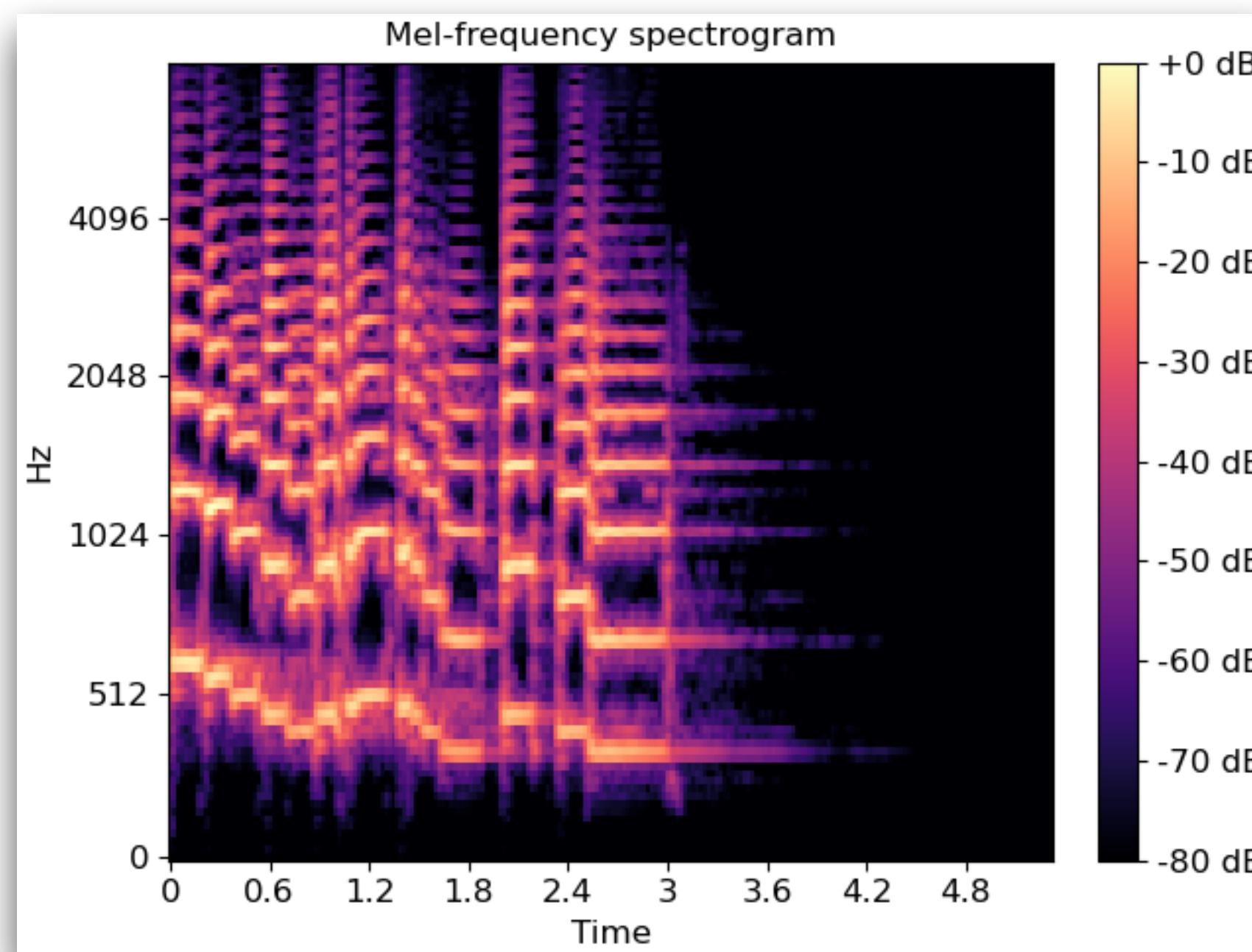
Perceptual Property: “Yanny” or “Laurel”?



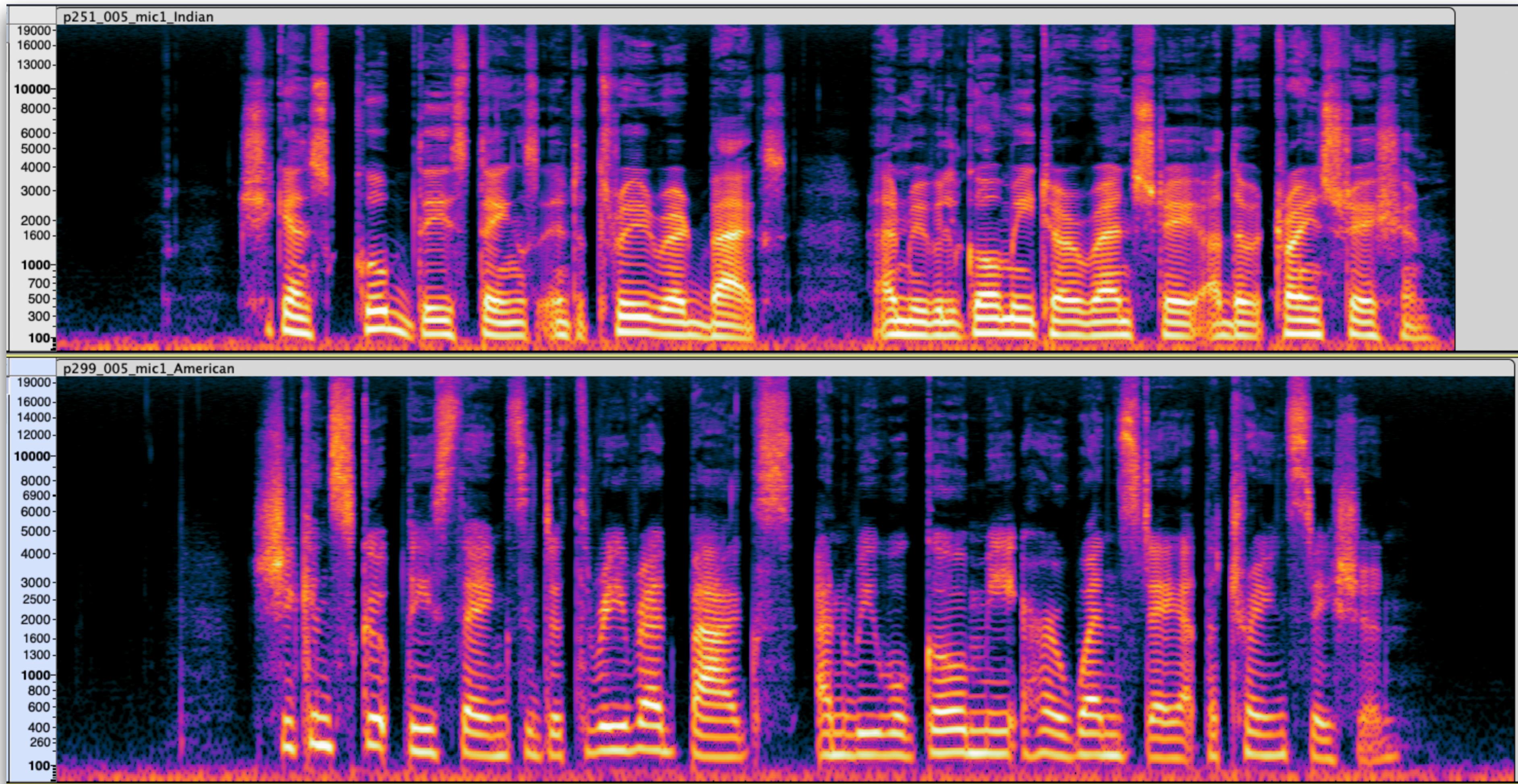
“Yanny” or “Laurel”?

Spectrogram: Visualization of Sound

- Librosa
 - [librosa.feature.melspectrogram](#)

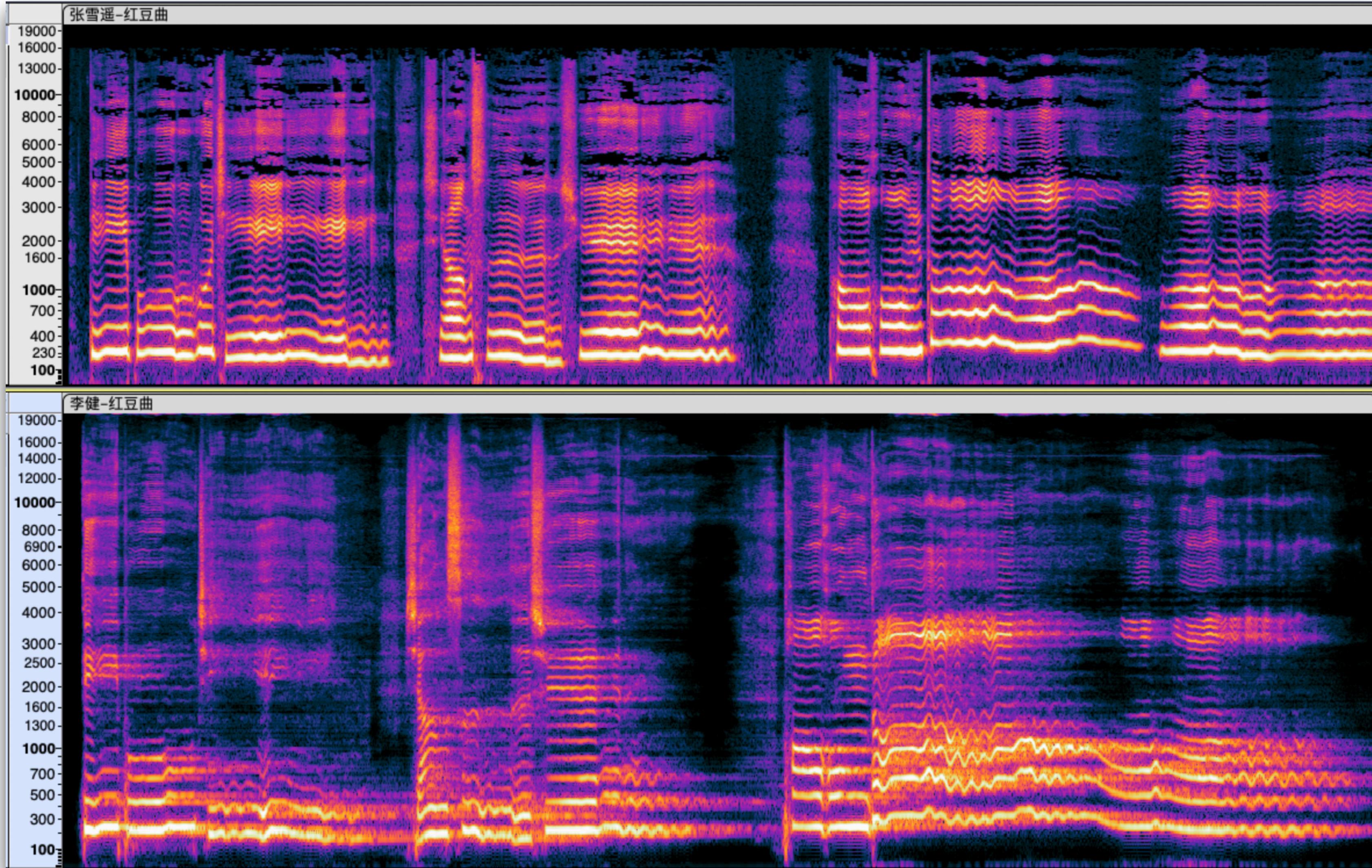


Speech — Sound that owns semantic information



A same utterance spoken by different people in different accents

Singing Voice — A More Beautiful Speech



Magic of Singing Voice Conversion

	Source	Conversion Results [1]	Ground Truth
韩红 to 李健			
齐秦 to 李健			
张学友 to 李健			-
林志炫 to 李健			-
陶喆 to 李健			-

Source	Reference	Results
李健《异乡人》	Xueyao	
Peking Opera Performer		
Peking Opera《苏三起解》	Xueyao	

[1] Xueyao Zhang, et al. Leveraging Content-based Features from Multiple Acoustic Models for Singing Voice Conversion. Machine Learning for Audio Workshop, NeuIPS 2023.

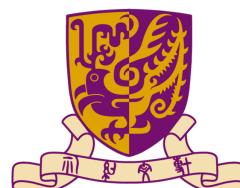


香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen



THANKS



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen