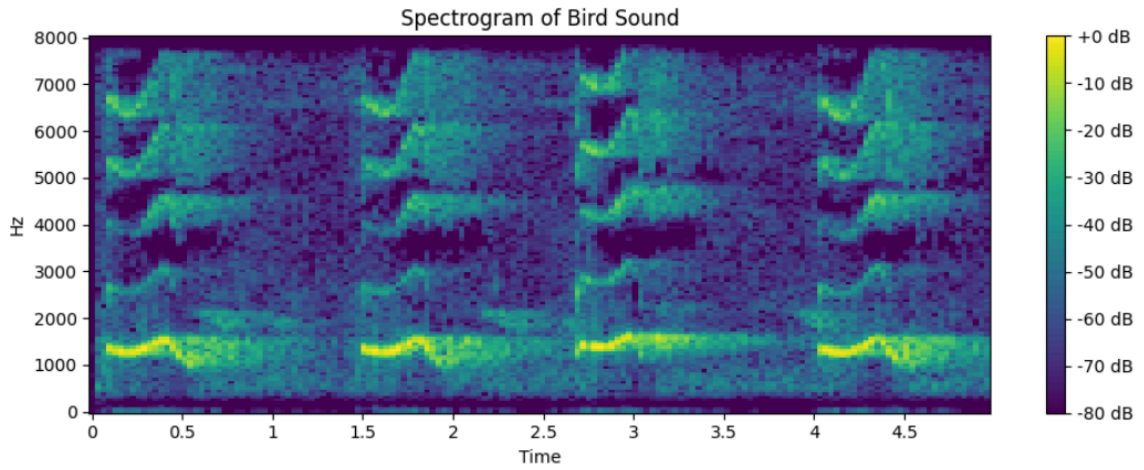
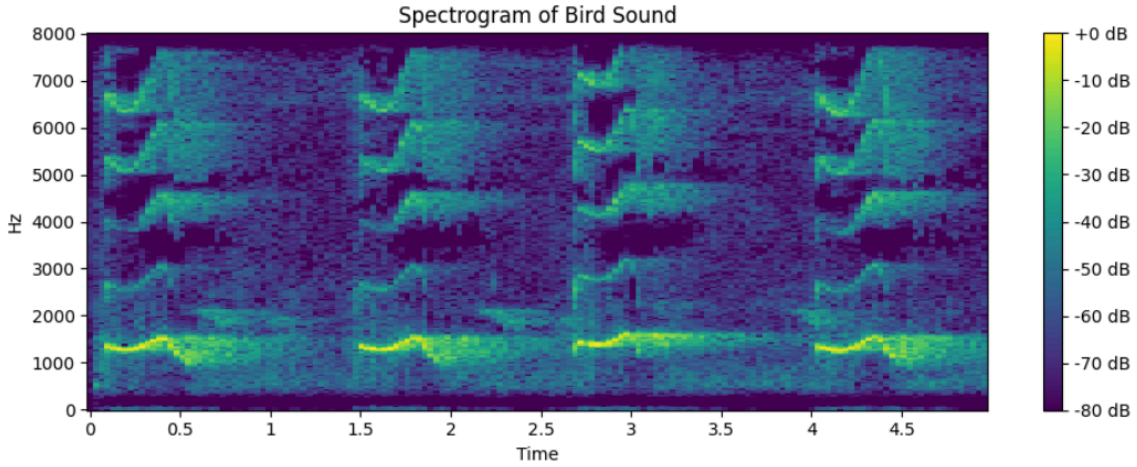
3.7 Exercises

1. Generate a spectrogram of a bird sound with different FFT sizes (e.g., 256, 512, 1024). Compare the time–frequency resolution.

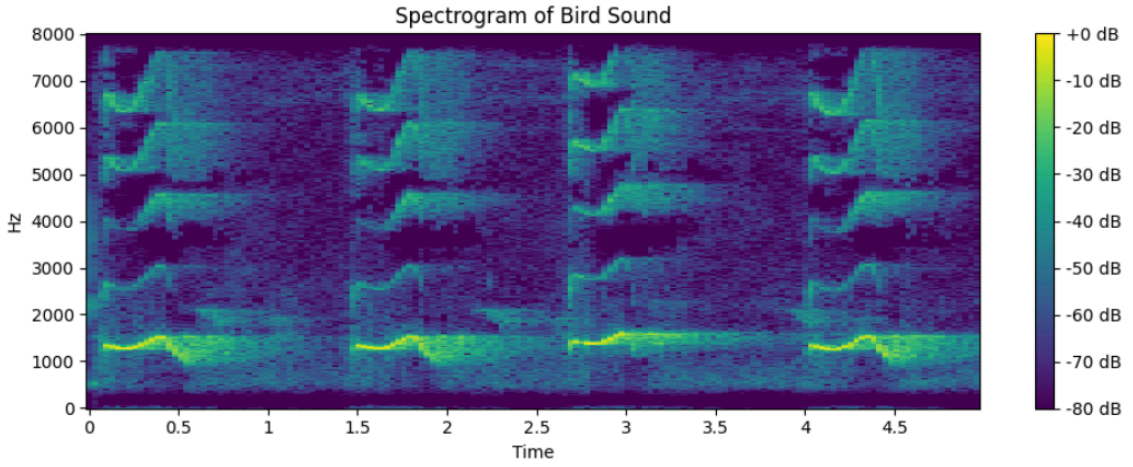
n_fft=256



n_fft=512



n_fft=1024



In a spectrogram, the sound is divided into small boxes of time and frequency. The horizontal axis (X) shows time, the vertical axis (Y) shows frequency, and the colour inside each box shows how strong the sound is at that moment and pitch. These boxes are created by running the Fast Fourier Transform (FFT) on short chunks of the signal. The parameter n_fft controls how many samples go into each chunk.

When n_fft is small (e.g., 256), each FFT window is short. This makes the spectrogram update quickly in time, so the boxes are narrow along X. The result is good time resolution: you can see exactly when a bird chirp starts and ends. But because the window is short, the FFT cannot separate nearby frequencies well. The boxes become taller along Y, giving poor frequency resolution, and the spectrum looks blocky or smeared vertically.

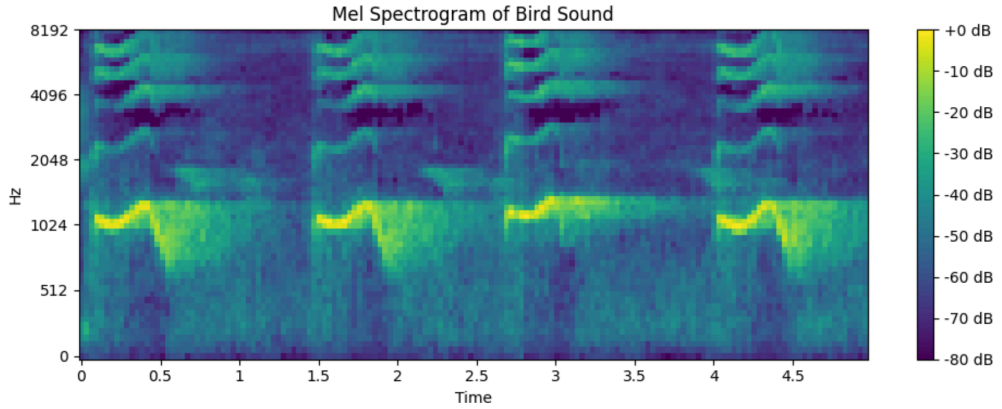
When n_fft is large (e.g., 1024), each FFT window is long. This improves frequency resolution: the boxes are shorter along Y, so you can distinguish finer pitch differences such as harmonic details in a bird call. However, the tradeoff is worse time resolution. The boxes become wider along X, so a very short chirp is stretched horizontally, making it look blurred in time.

Another effect is seen in the colour intensity. With a larger n_fft, the same sound energy is spread across more frequency bins. This reduces the energy per bin, so the dB values are lower, and the colours appear darker. Conversely, with a smaller n_fft, the energy is concentrated in fewer bins, giving brighter colours.

Comparison of `n_fft` sizes for bird sounds

n_fft Size	Frequency Resolution	Time Resolution	Spectrogram Effect (Bird Sound)
Small (256)	Low (broad bins, can't separate close frequencies)	High (updates often, good for short chirps)	Bird chirps look sharp in time but smeared across frequency.
Medium (512)	Balanced	Balanced	A compromise: both frequency and time are moderately clear.
Large (1024)	High (fine frequency detail, can separate close pitches)	Low (events are blurred in time)	Bird songs show clear harmonic structure, but short chirps look stretched.

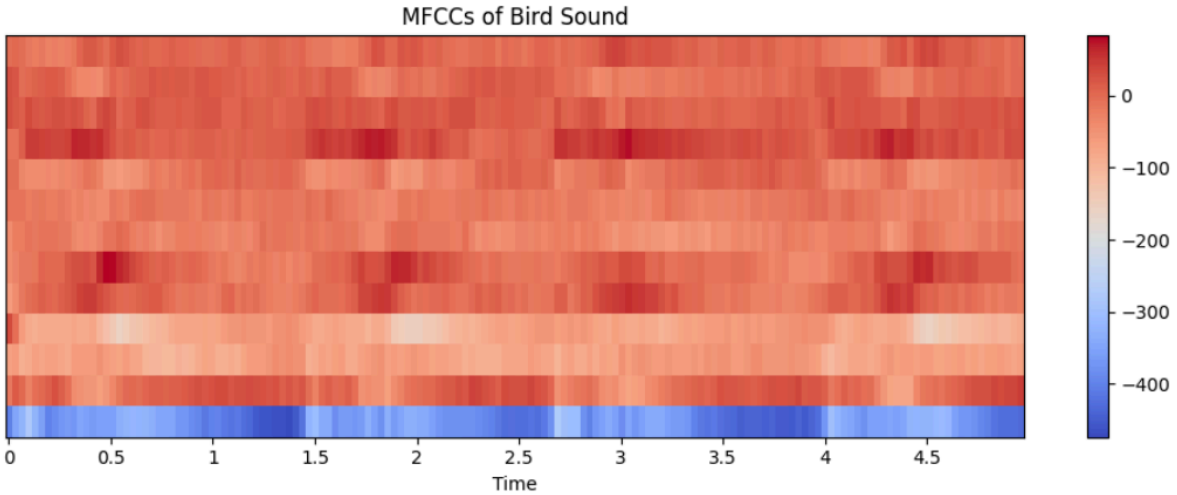
2. Compare the resolution of Mel filters in the low-frequency range versus the high-frequency range. Why does the Mel scale emphasise low frequencies more?



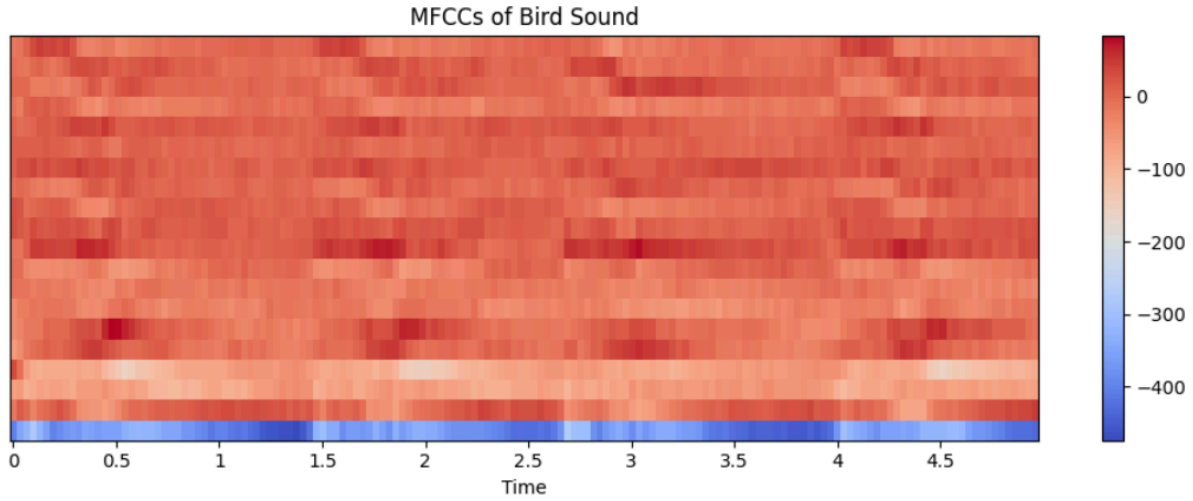
In a Mel filter bank, the resolution of filters is much higher at low frequencies and lower at high frequencies. This means that filters are dense in the low-frequency region, each covering a narrow frequency band, while they become sparse at higher frequencies, each covering a wider band. The reason is rooted in human hearing: we are much more sensitive to pitch differences at low frequencies (for example, the difference between 200 Hz and 300 Hz is clearly noticeable), but much less sensitive at high frequencies (the difference between 7000 Hz and 7100 Hz is almost indistinguishable). The Mel scale reflects this psychoacoustic property, ensuring that machine learning models emphasise the parts of the spectrum that are most perceptually meaningful. This makes Mel-based representations particularly effective for tasks like bird sound detection and speech recognition, where critical information is often carried in the lower frequency ranges.

3. Extract MFCCs from the same bird sound using 13 coefficients and then 20 coefficients. Compare the feature sets.

n_mfcc=13



n_mfcc=20

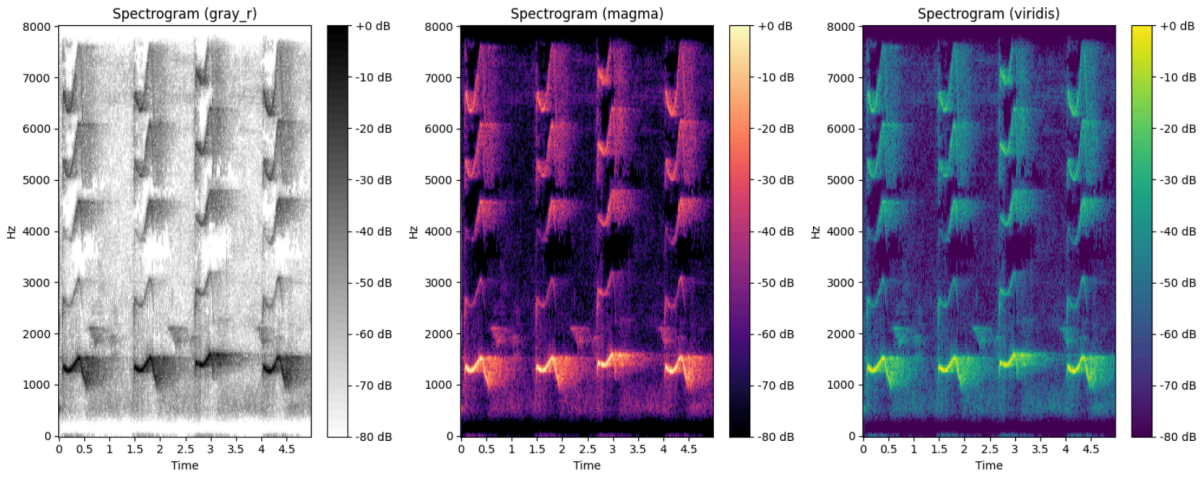


The number of Mel-Frequency Cepstral Coefficients (MFCCs) chosen controls how much detail is captured from the sound. Using 13 coefficients is standard in speech recognition, as it captures the broad spectral envelope (timbre) in a compact form. This makes the features stable, efficient, and less prone to overfitting.

Increasing to 20 coefficients adds more detail, allowing subtle frequency patterns and richer harmonics to be represented. This can be useful in bird sound analysis, where calls are more complex than human speech. However, it also introduces more sensitivity to noise and less compact features. In visualisations, this is seen as shorter boxes along the y-axis and lighter colours, showing energy spread across more coefficients.

13 coefficients are sufficient for speech, but 20 (or more) may benefit bioacoustics by capturing richer detail, at the cost of added complexity.

4. Try different colourmaps (gray_r, magma, viridis) for spectrogram visualisation. Which one makes the patterns easiest to see?



When visualising spectrograms, the choice of colourmap does not change the underlying data but strongly affects how easy it is to interpret the patterns. Using **gray_r** gives scientific

clarity, with darker areas showing stronger energy, but it can sometimes make subtle differences harder to notice. The **magma** colourmap provides a smooth gradient from dark to bright, which makes intensity changes stand out more clearly. Meanwhile, **viridis** is perceptually uniform and colourblind-friendly, making fine structures such as harmonics and short chirps easier to distinguish. Overall, viridis and magma are often preferred for highlighting detailed sound patterns, while gray_r remains useful for simple and consistent scientific presentations.

3.8 Reflection

Why does increasing FFT size improve frequency resolution but worsen time resolution?

A larger FFT size analyses more samples per window, which allows finer separation of close frequencies (better frequency resolution). However, because the window is longer, the spectrogram updates less often, so short events like bird chirps appear stretched or blurred in time. This is the tradeoff between frequency and time resolution.

Why are Mel spectrograms better aligned with human hearing than linear spectrograms?

Mel spectrograms use the Mel scale, which spaces filters densely at low frequencies and sparsely at high frequencies. This matches the way humans (and many animals) perceive sound—sensitive to small pitch changes at low frequencies but less sensitive at high ones. As a result, Mel spectrograms highlight the most perceptually relevant information.

Why are MFCCs popular in speech recognition, but sometimes less effective in bird sound analysis?

MFCCs are widely used in speech recognition because they capture the broad spectral envelope (timbre) in a compact way with relatively few coefficients, making them efficient and less prone to overfitting. However, bird calls often contain richer harmonics and rapid frequency changes. Standard MFCCs with 13 coefficients may miss these fine details, so they can be less effective unless more coefficients are used.

How might the choice of features (spectrogram vs. Mel vs. MFCC) affect neural network performance?

The feature type shapes how a neural network learns. A raw spectrogram preserves all details but may include redundancy and noise. Mel spectrograms reduce irrelevant detail and emphasise perceptually important features, which often improves learning. MFCCs provide compact representations well-suited for speech, but for complex bird sounds, they might remove useful harmonic information. Choosing the right feature representation can therefore directly influence accuracy and generalisation.