Episode 35 (Audio decomposition and synthesis (Fourier transform and inverse transform))

(Illustrated) How Voice Synthesis Works

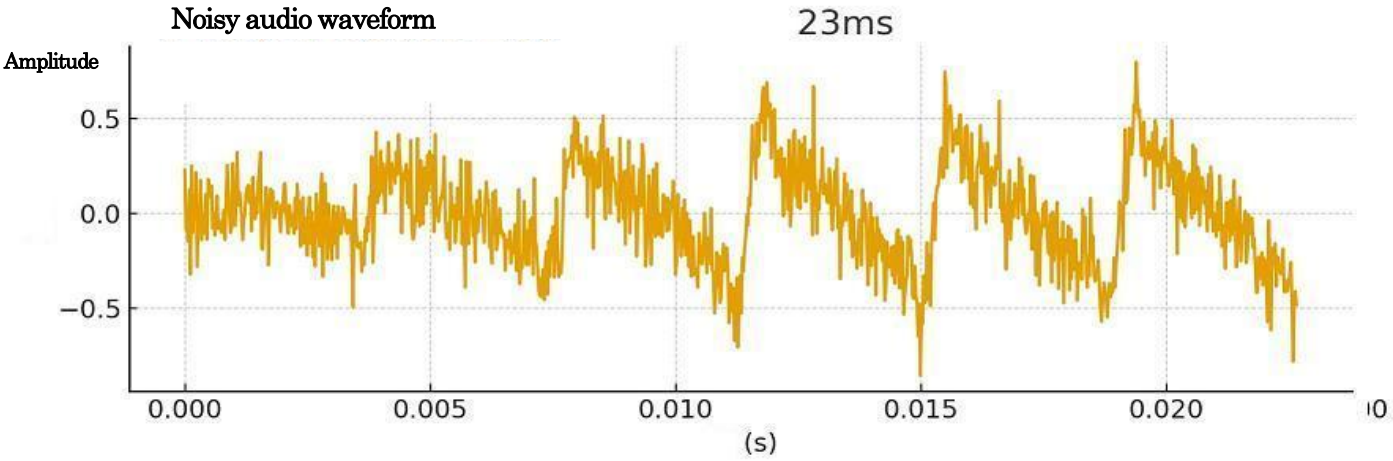
The use of generative AI to disguise text (word processor) files as if spoken by male politicians is becoming increasingly active. To understand the fundamental theory behind this mechanism, one must grasp Fourier transforms and inverse Fourier transforms. These transforms are always used in speech synthesis.

The Fourier transform utilizes complex wave functions also employed in quantum mechanics. While understanding these is necessary, programs (applications) incorporating these complex wave functions are already available. Users need only train the AI program with several suitable audio samples of the politician they wish to transform.

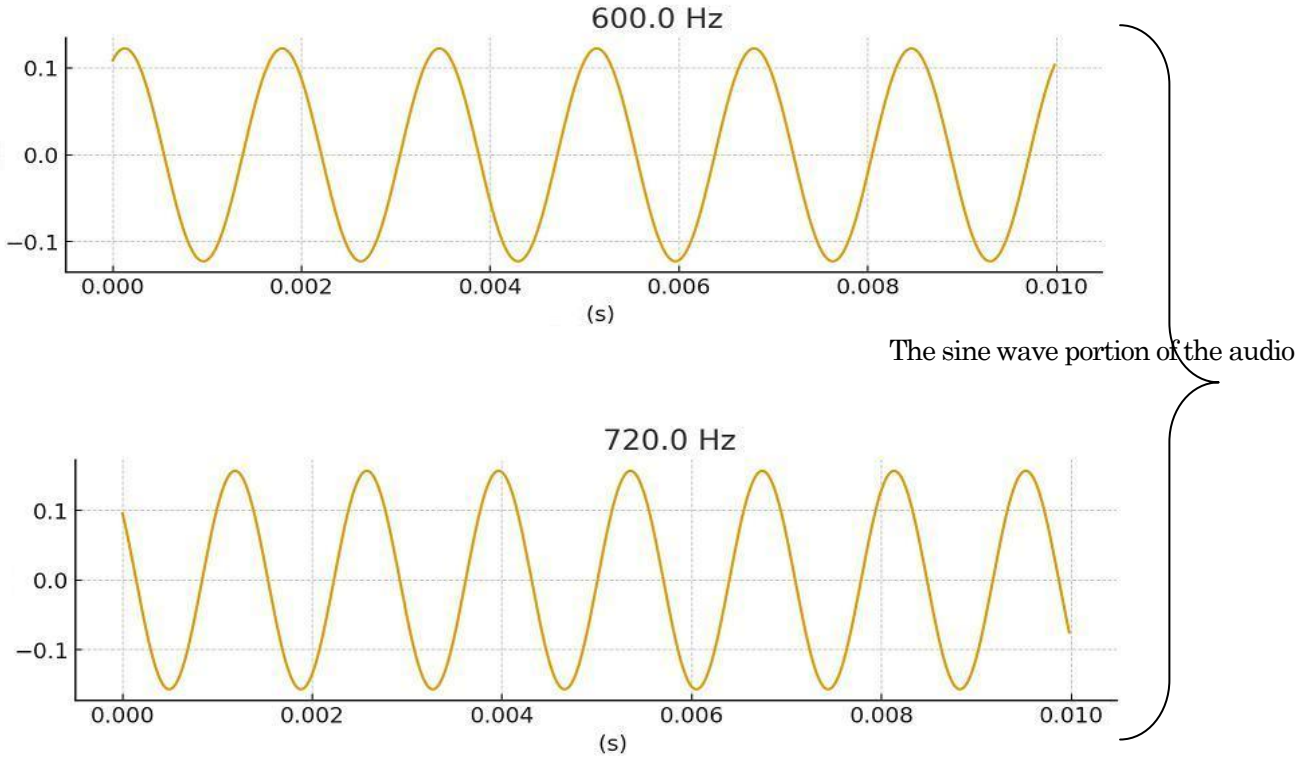
Here, we illustrate the process of removing noise from audio containing noise using Fourier transforms and inverse Fourier transforms, without using wave functions at all. Audio synthesis is merely an application of noise removal.

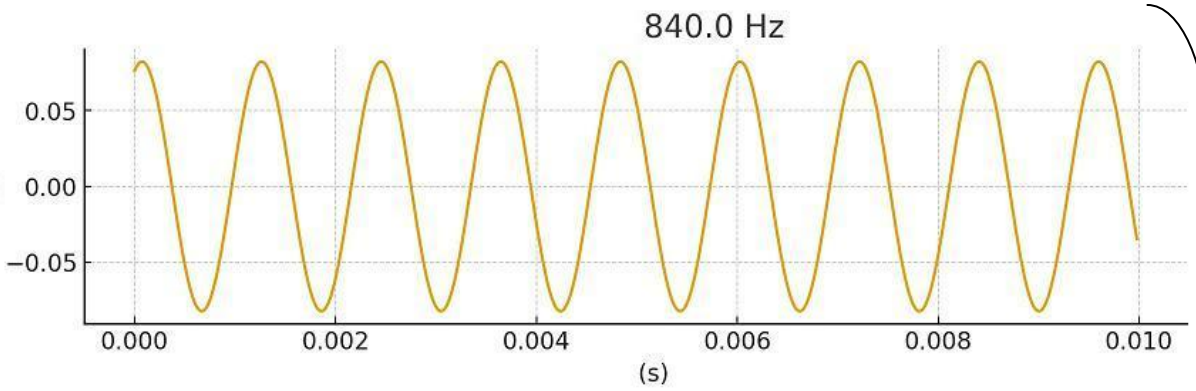
[What is the Fourier Transform?]

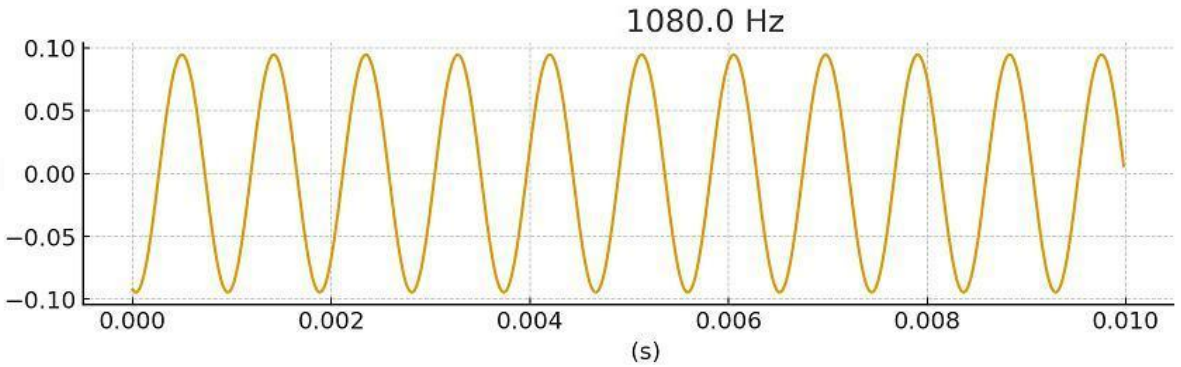
① Any audio signal can always be represented as a combination of multiple sine waves.
The example we'll use is an audio signal containing noise like the following.



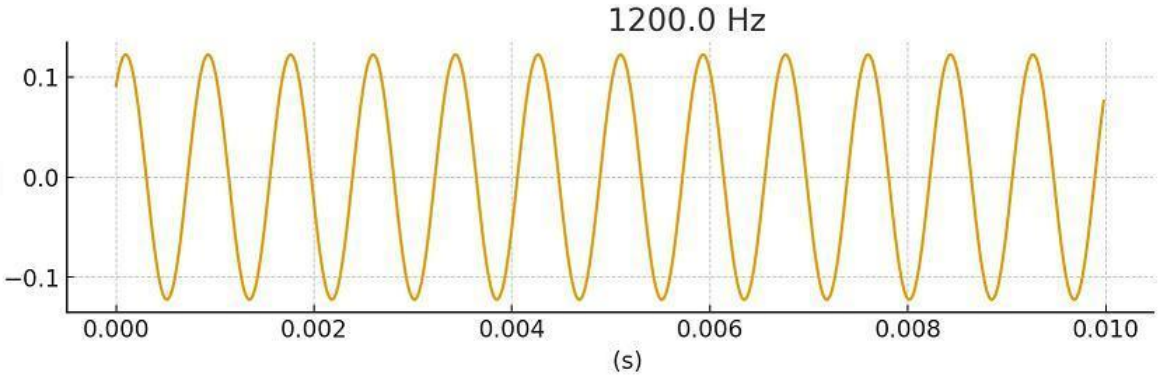
② An example of representing sound and noise using multiple sine waves (sign curves) (though not all).

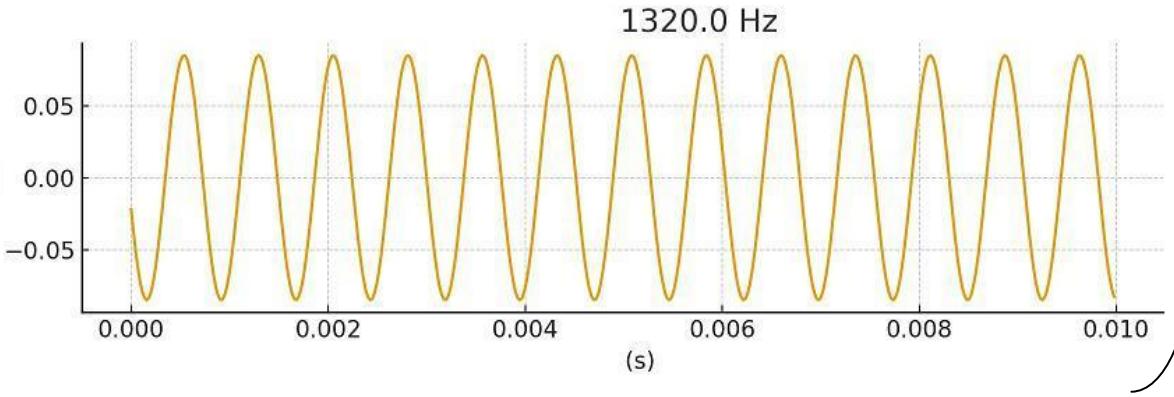


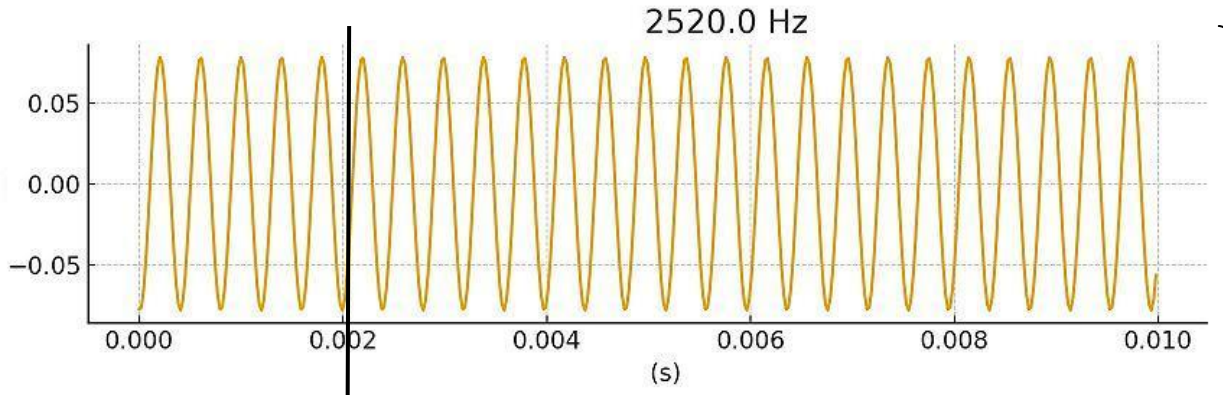




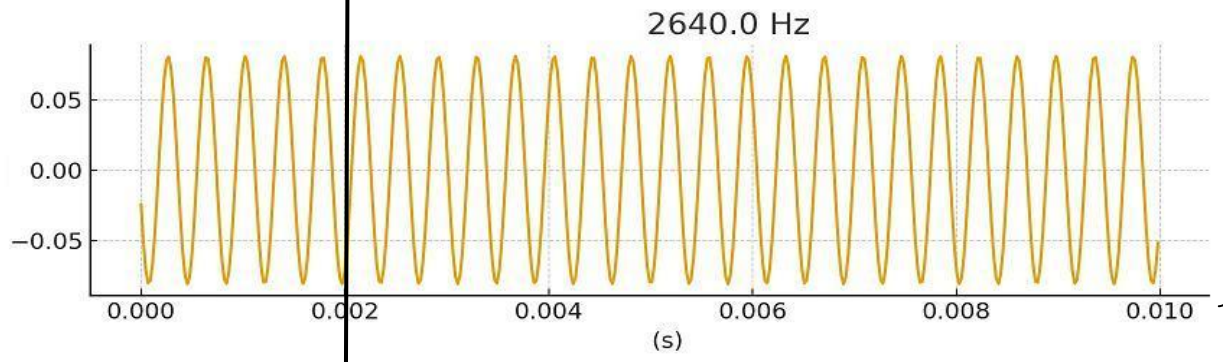
The sine wave portion of the audio



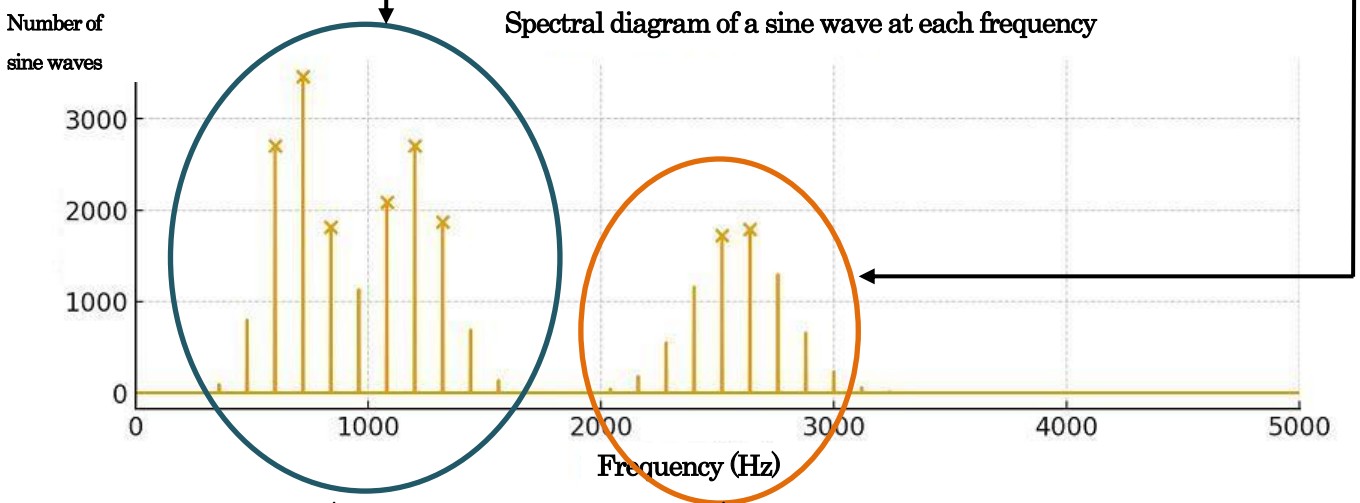




The sine wave in the noise portion



③Different sine waves (sine curves) can be represented in a spectrum diagram for each frequency.
This is the key point of the Fourier transform.

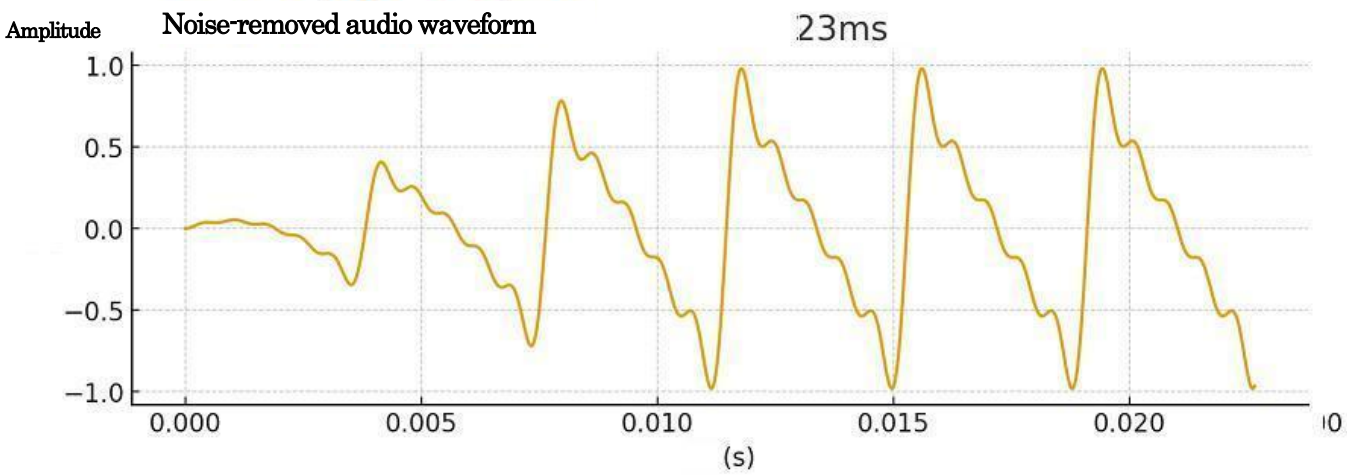


Noise sections can be removed based on spectral differences.

The audio portion can be left as-is or modified (synthesized).

Restore only this part (inverse Fourier transform).

[Waveform of the inverse Fourier transform result]



The above is an explanation of the Fourier transform and inverse Fourier transform that can be intuitively understood by looking at the figure.