［日本語版］

（図解）音声合成の仕組み

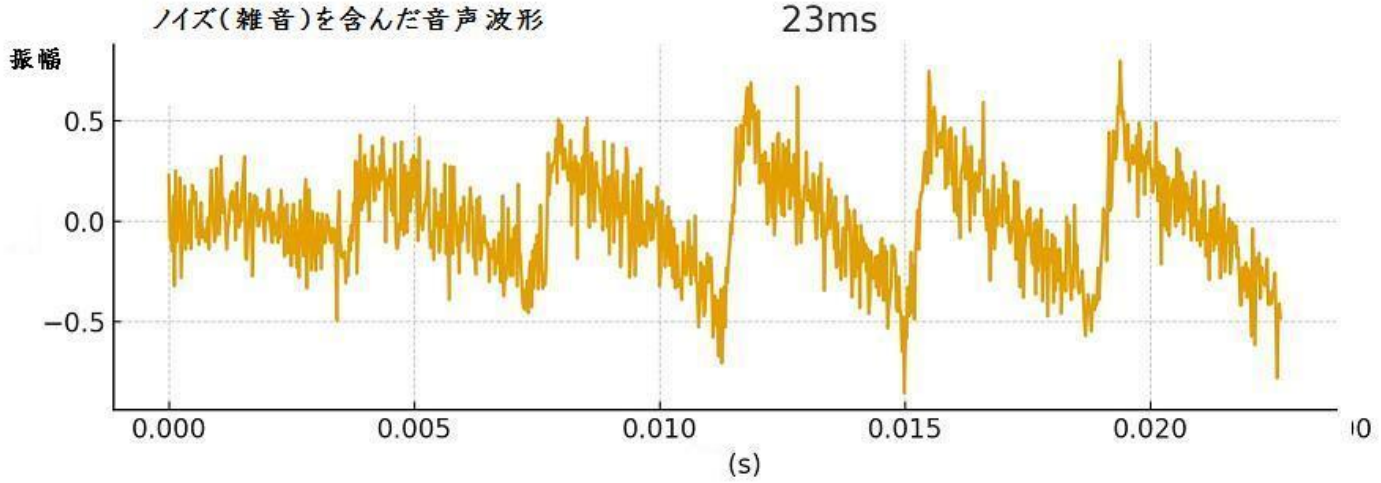
テキスト（ワープロ）ファイルで入力した内容を男性の政治家が話したように偽装する生成 AI の利用が活発になってきています。その仕組みの基礎理論を知る為には、フーリエ変換と逆フーリエ変換の理解が必要です。音声の合成には必ずフーリエ変換と逆フーリエ変換を使います。

フーリエ変換には、量子力学でも使われる複雑な波動関数が使われます。また、その理解も必要になりますが、既にそれらの複雑な波動関数を組み込んだプログラム（アプリケーション）が提供されています。利用するユーザは、変換したい政治家の適当な複数の音声を生成 AI プログラムに学習させるだけです。

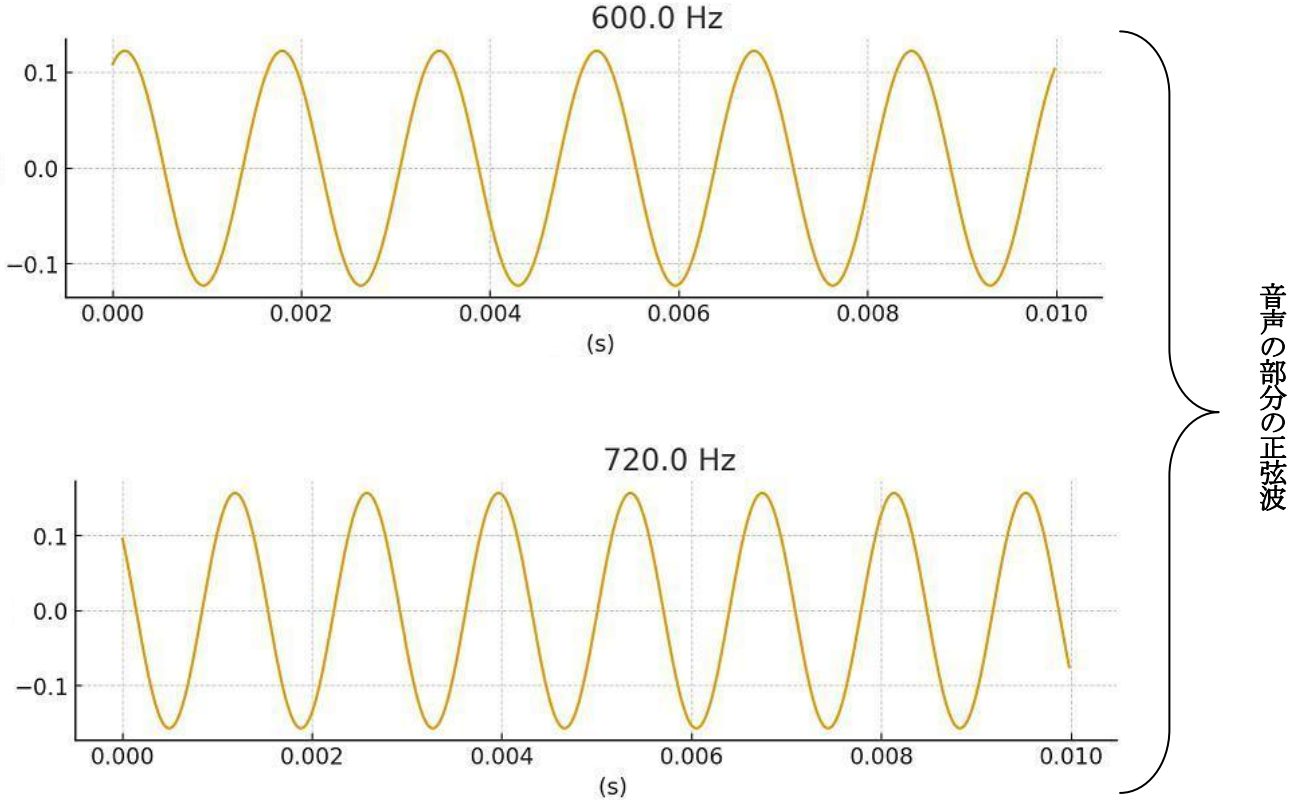
ここでは、波動関数は一切使わず、フーリエ変換と逆フーリエ変換をノイズを含んだ音声からノイズを除去する例を用いて図解します。音声は合成は、ノイズ除去の応用でしかありません。

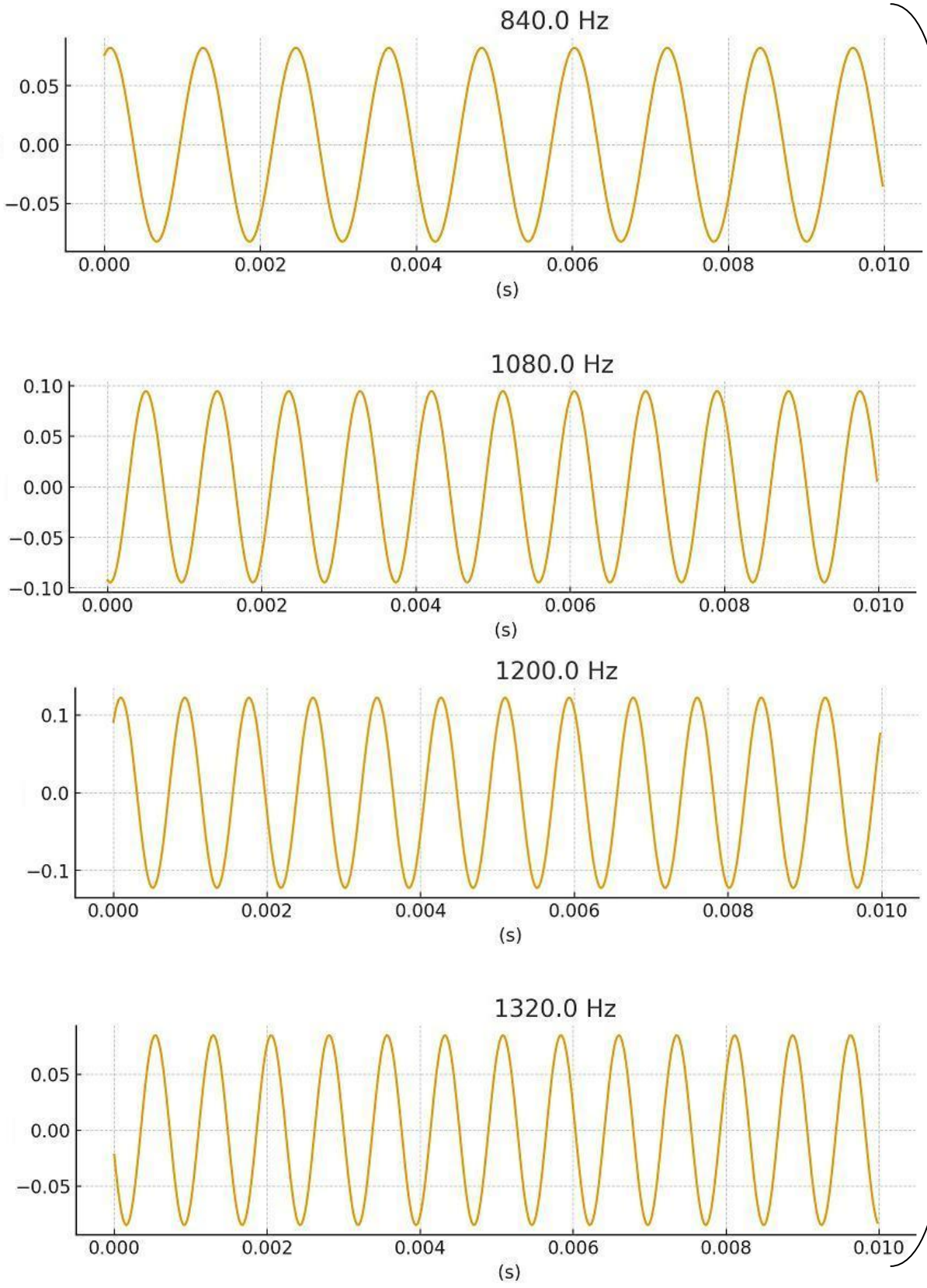
［フーリエ変換とは］

① どのような音声でも必ず複数の正弦波（サインカーブ）で表わすことができる。
例に使うのは、以下のようなノイズ（雑音）を含んだ音声です。

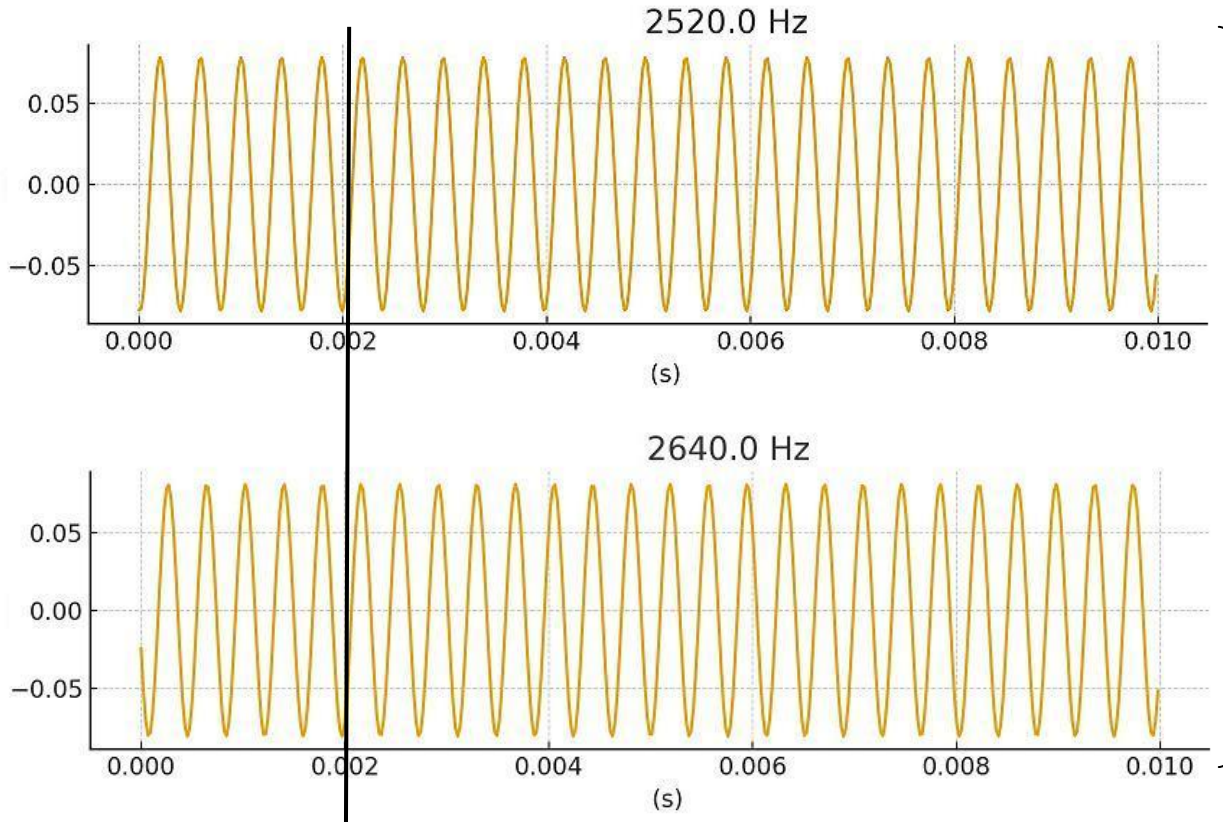


② 音声とノイズを複数の正弦波（サインカーブ）で表わした例（全てではない）。



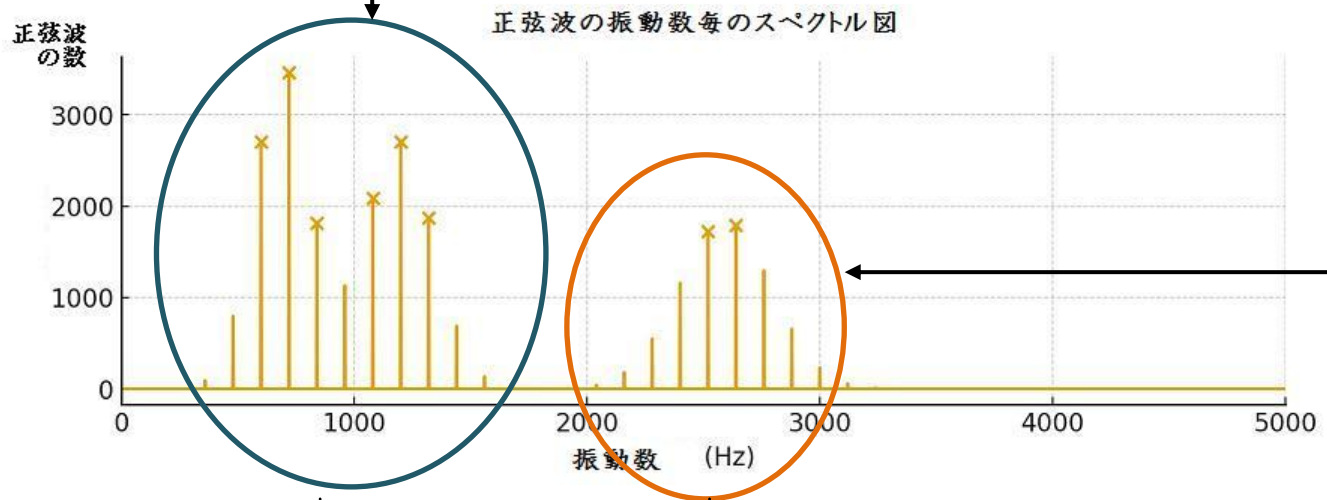

840.0 Hz
0.05
0.00
−0.05
0.000
0.002
0.004
0.006
0.008
0.010
(s)
1080.0 Hz
0.10
0.05
0.00
−0.05
−0.10
0.000
0.002
0.004
0.006
0.008
0.010
(s)
1200.0 Hz
0.1
0.0
−0.1
0.000
0.002
0.004
0.006
0.008
0.010
(s)
1320.0 Hz
0.05
0.00
−0.05
0.000
0.002
0.004
0.006
0.008
0.010
(s)

音声の部分の正弦波



ノイズの部分の正弦波

③異なる正弦波（サインカーブ）を振動数毎にスペクトル図で表わすことができる。
これが、フーリエ変換の重要なポイント。

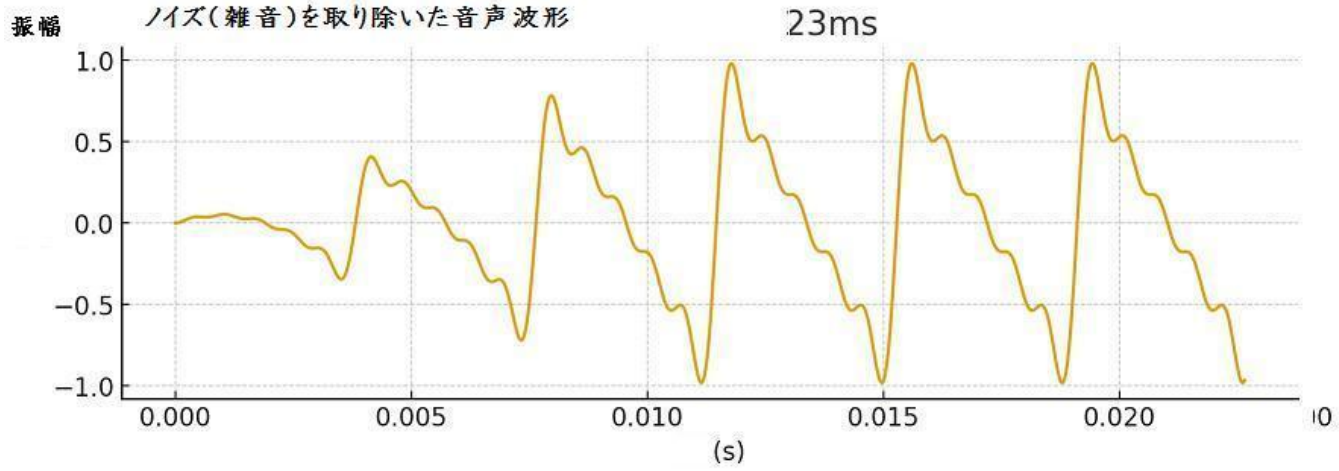


スペクトルの違いでノイズの部分を取り除くことができる。

音声の部分はそのまま残すか、修正（合成）可能

この部分だけを元に戻す（逆フーリエ変換）。

［逆フーリエ変換した結果の波形］



以上、図を見て直感的に理解できる、フーリエ変換と逆フーリエ変換の解説です。

[English Board]

(Illustrated) How Voice Synthesis Works

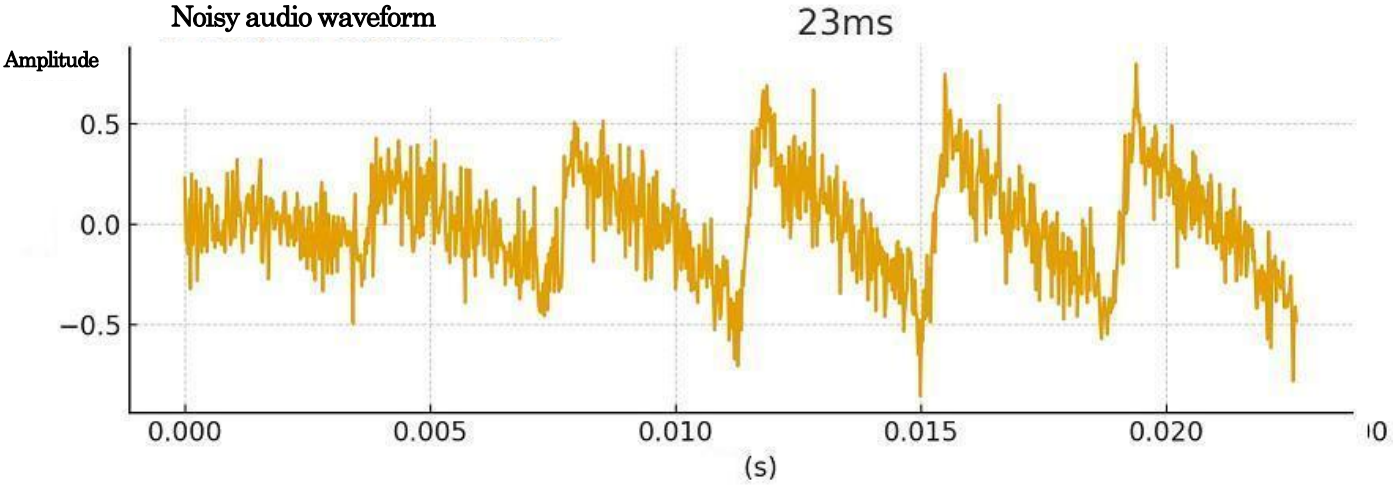
The use of generative AI to disguise text (word processor) files as if spoken by male politicians is becoming increasingly active. To understand the fundamental theory behind this mechanism, one must grasp Fourier transforms and inverse Fourier transforms. These transforms are always used in speech synthesis.

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.

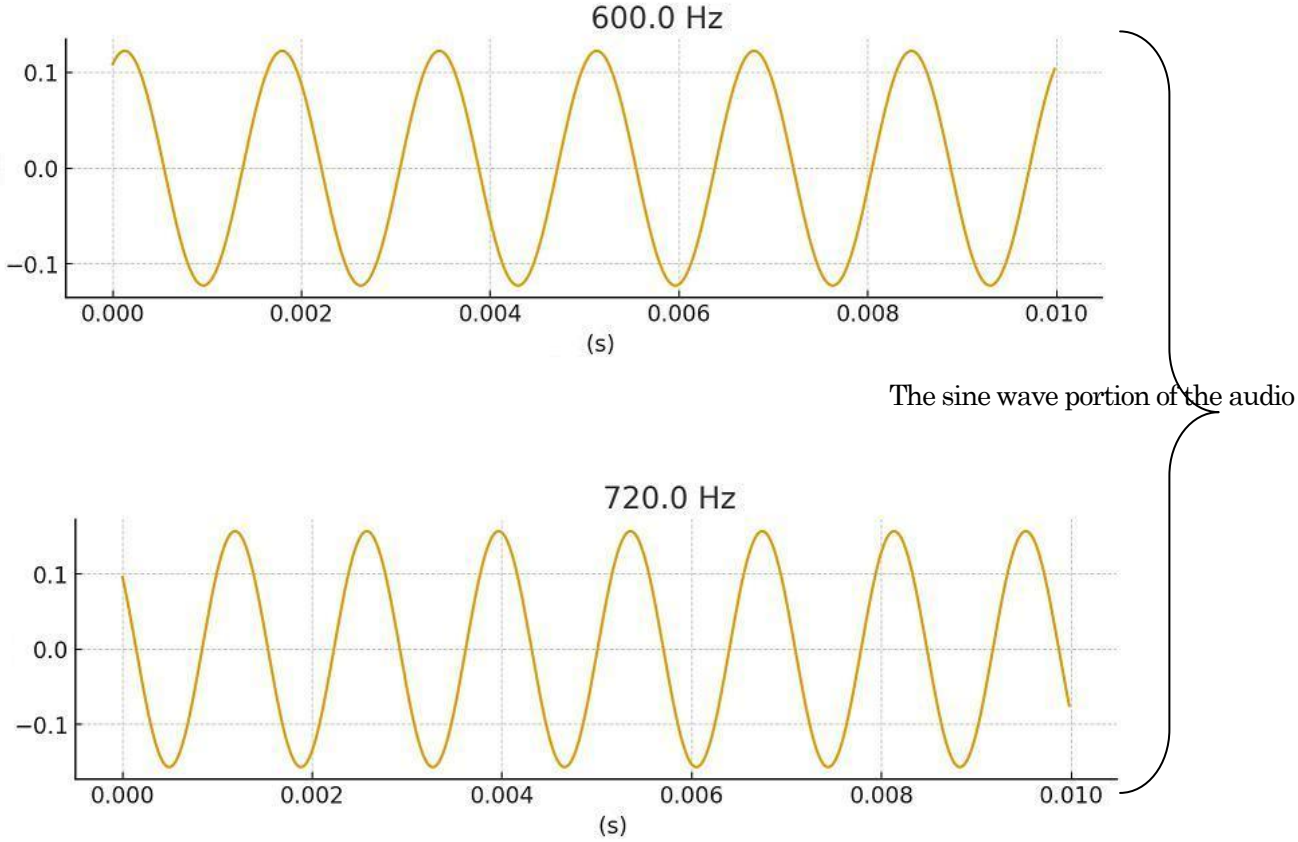
Here, we illustrate the process of removing noise from audio containing noise using Fourier transforms and inverse Fourier transforms, without using wave functions at all. Audio synthesis is merely an application of noise removal.

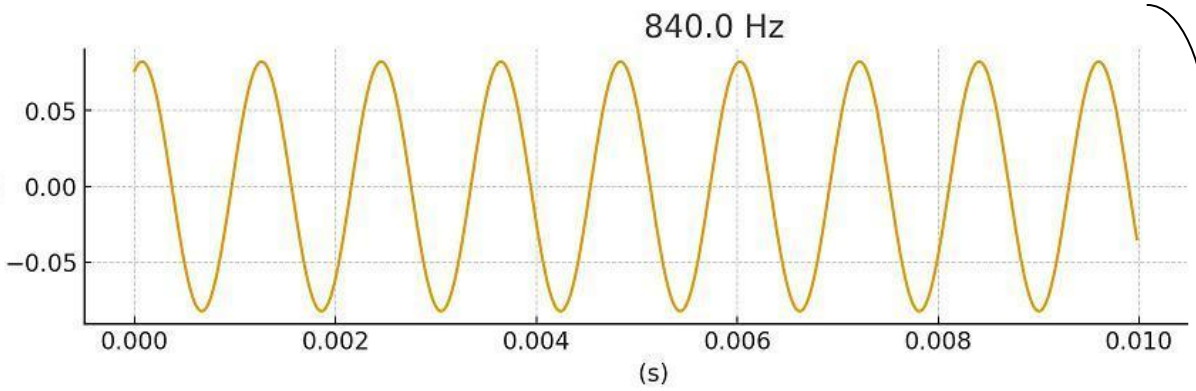
[What is the Fourier Transform?]

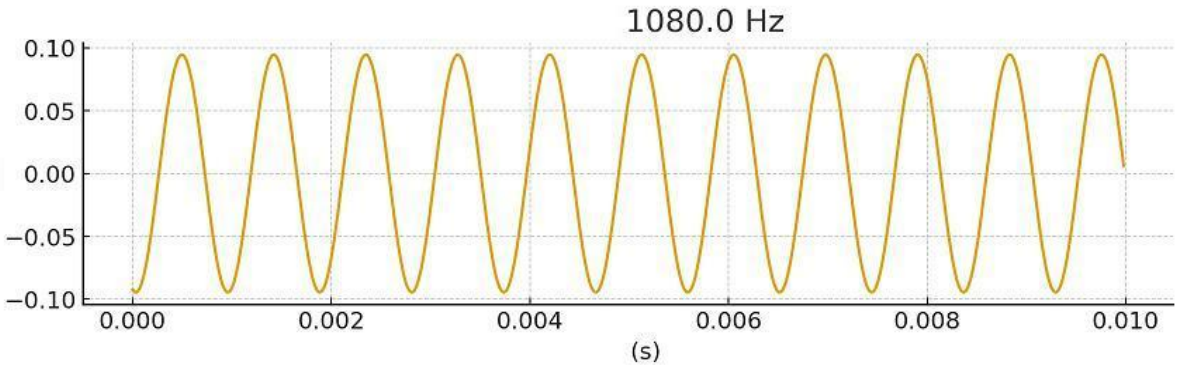
① Any audio signal can always be represented as a combination of multiple sine waves.
The example we'll use is an audio signal containing noise like the following.



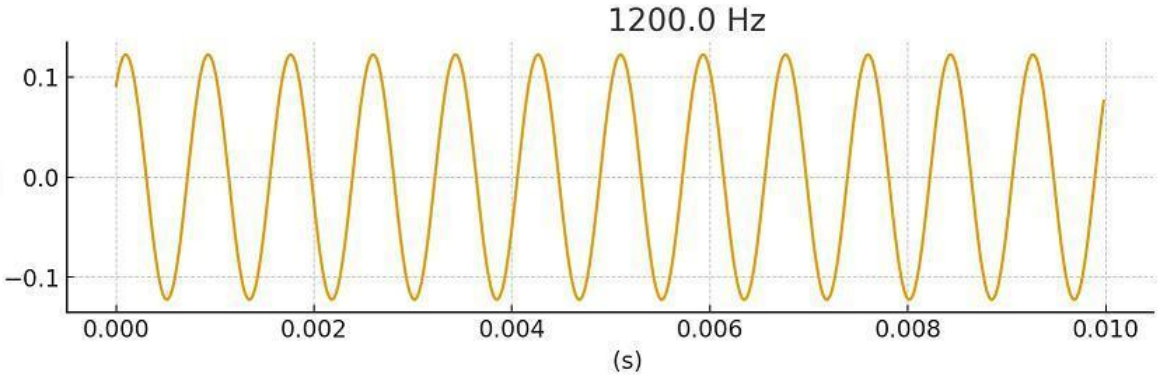
② An example of representing sound and noise using multiple sine waves (sign curves) (though not all).

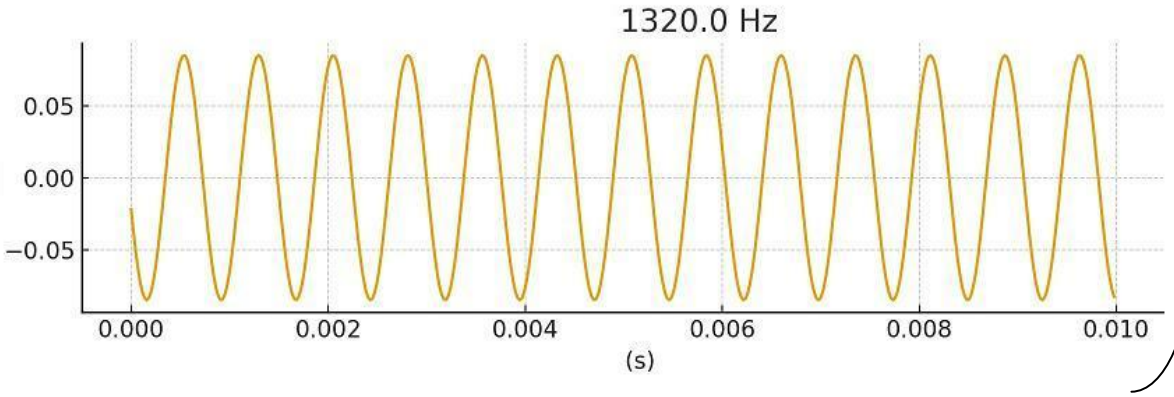


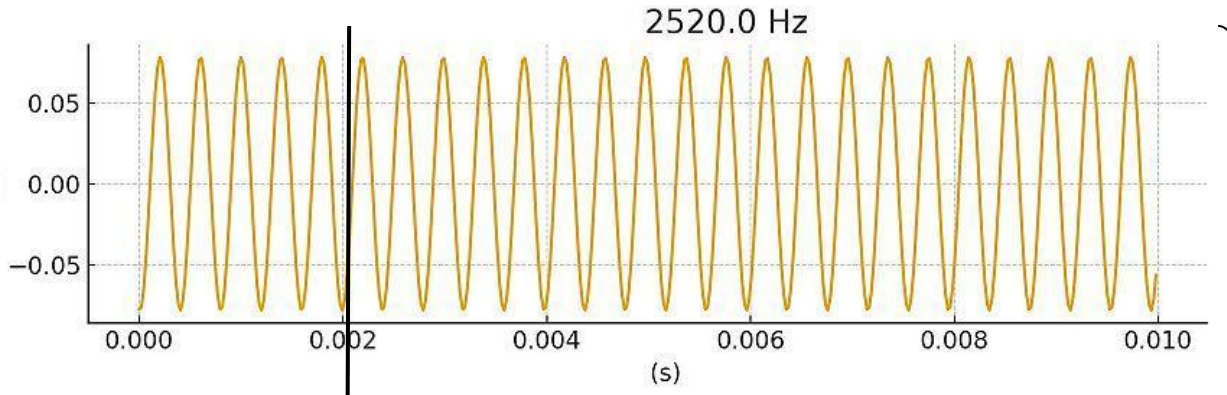




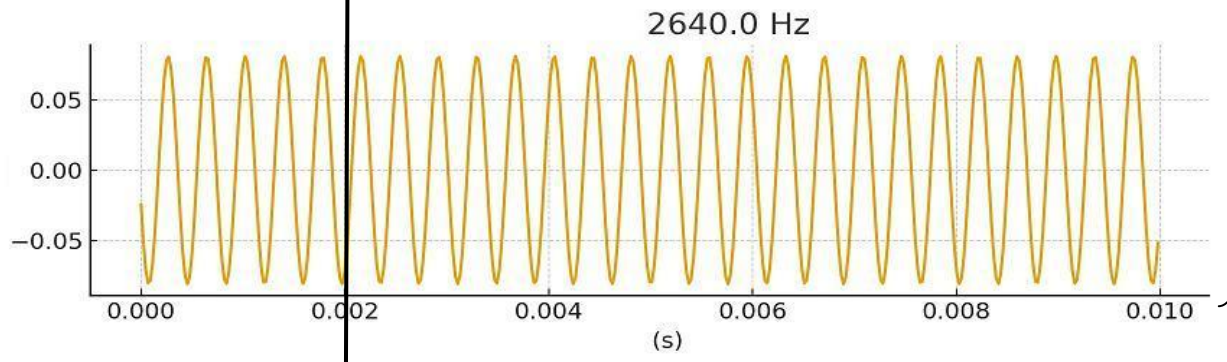
The sine wave portion of the audio



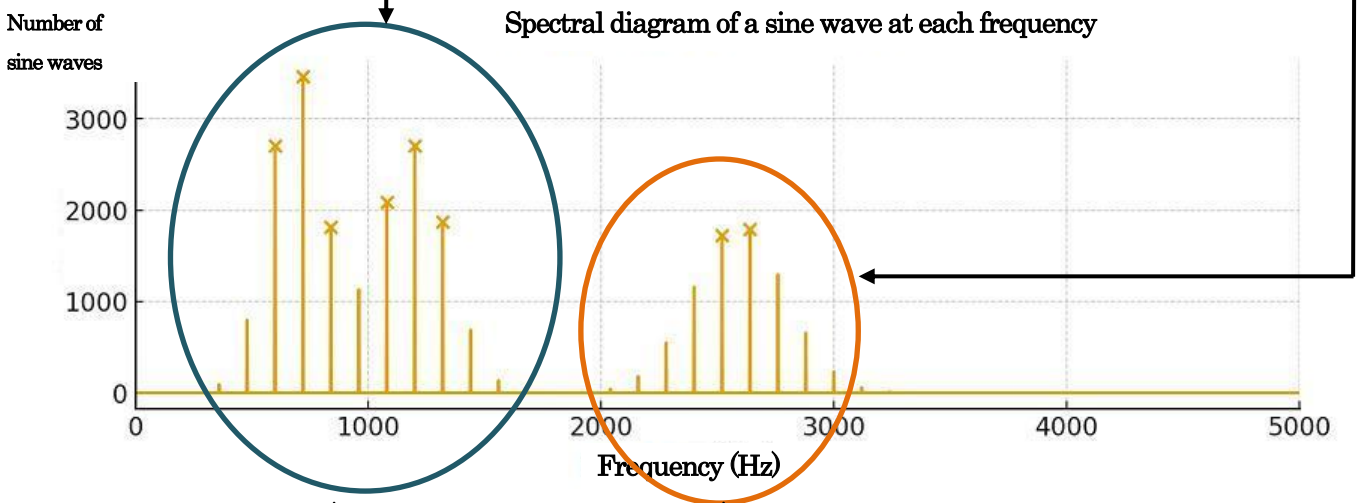




The sine wave in the noise portion



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

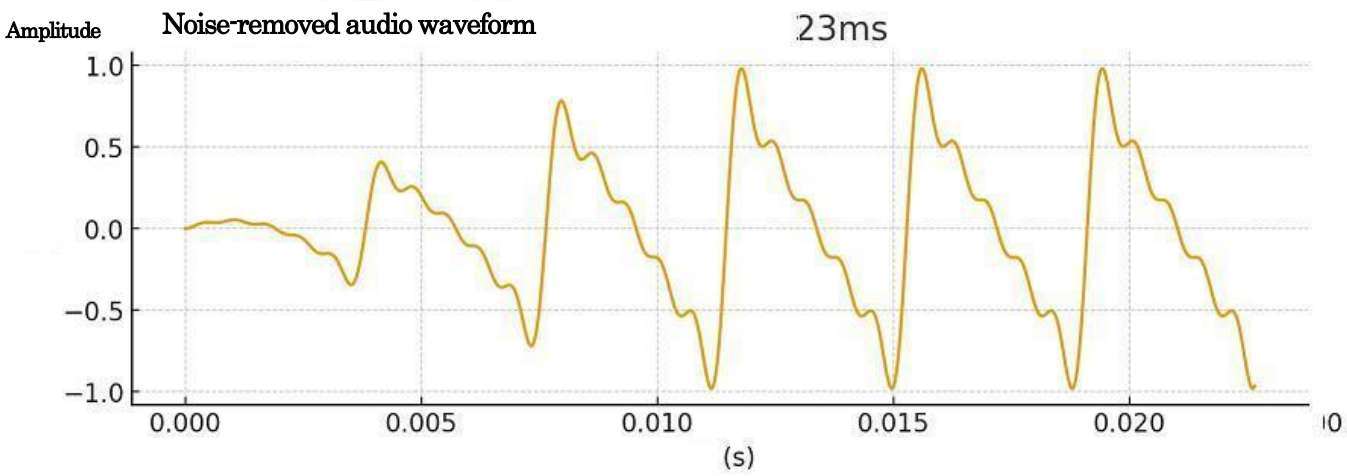


Noise sections can be removed based on spectral differences.

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.