クレジット：
Mathematics and Informatics Center メディアプログラミング入門 2020 山肩洋子

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

メディアプログラミング入門

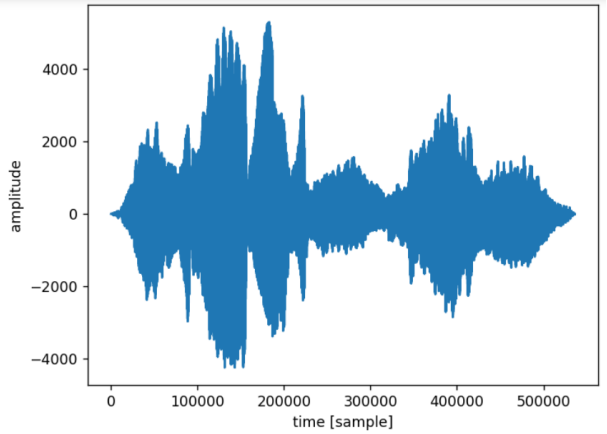
第2回：音の情報処理

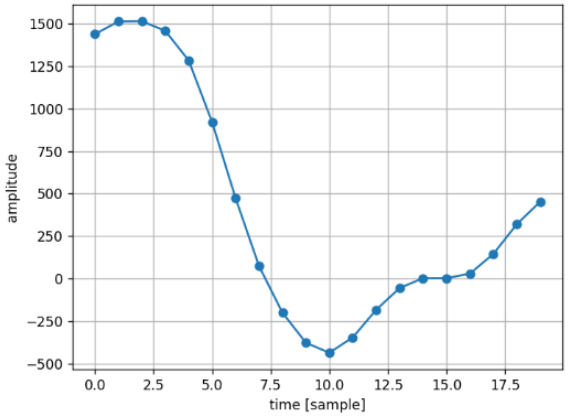
火５@本郷 2020年6月9日
情報理工学系研究科 数理・情報教育研究センター
准教授 山肩 洋子

第3回 音を分析して何の楽器の音かを当てよう

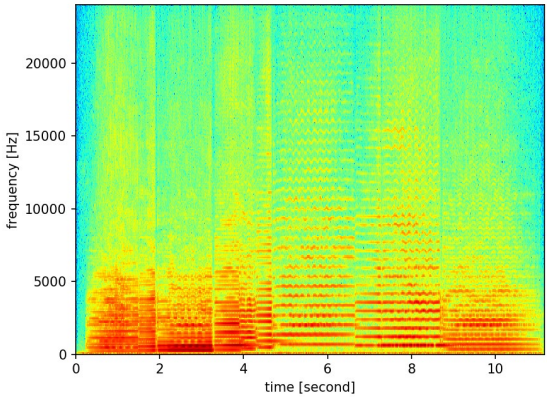
講義内容：聴覚の仕組みと音声や楽器の音響特性、マイクロフォン・スピーカの仕組み、コンピュータにおける音声データの表現や基礎的な解析手法を学ぶ

演習内容： 音声情報の入力（マイクロフォンの仕組み、サンプリング、量子化）、周波数分解（フーリエ変換、周波数フィルタリング、逆フーリエ変換）、スペクトログラム（短時間フーリエ変換）、聴覚特性、代表的な音特徴（MFCC)





マイクで受けた粗密波（連続値）を
コンピュータに取り込むためには？



音の周波数変化の可視化
スペクトログラム

本日の学習内容

- 音の入出力デバイス
  - 音情報って何？マイクロフォンとスピーカの仕組み
  - デジタル音声情報の可視化と再生
  - 連続値の離散化：サンプリング、量子化
- 周波数解析
  - フーリエ変換：
    その音はどのような周波数の音が混ざったものかを分析
  - 周波数フィルタリング・逆フーリエ変換：
    特定の周波数帯の音を除去してみよう
  - 短時間フーリエ変換（スペクトログラム）：
    周波数成分が時間とともに変化する音情報の分析
- 音響特徴抽出
  - スペクトル包絡とスペクトル微細構造（ケプストラム特徴）
  - MFCC

音声ファイルを読み込んで可視化: SoundProcessing1

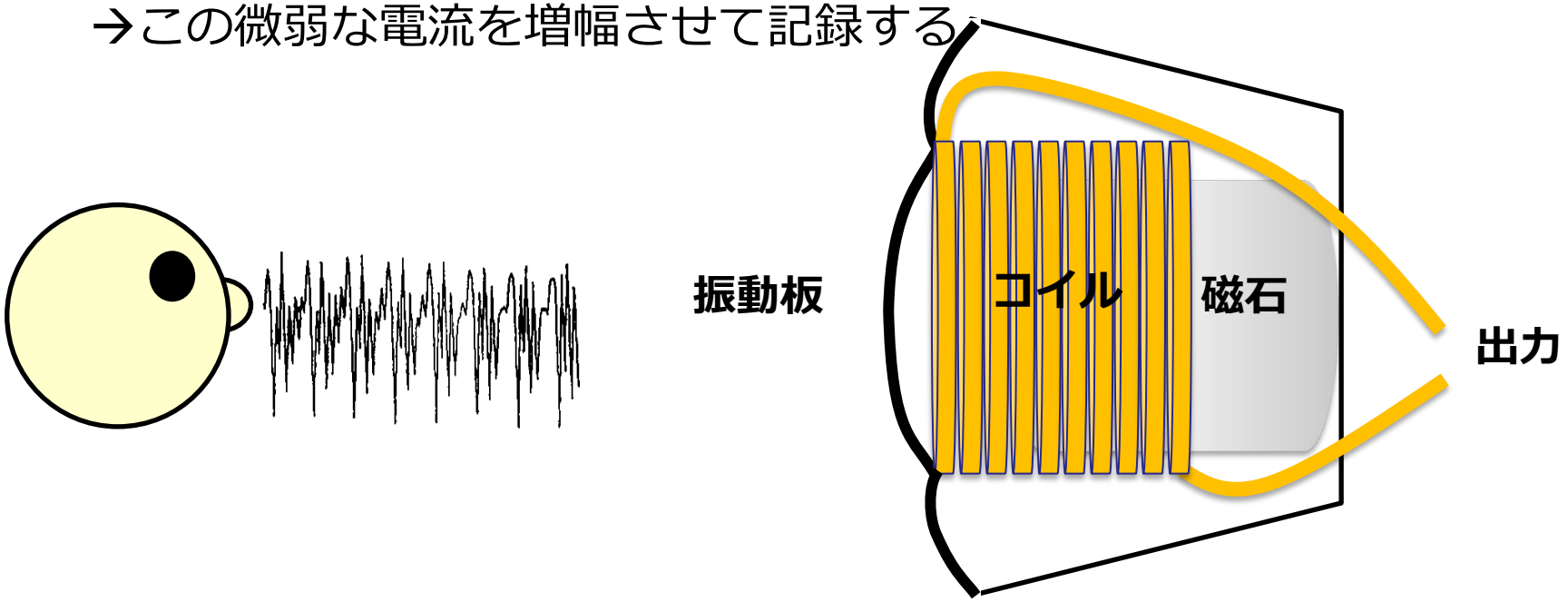
‘sounds/Violin-music.wav’の拡張子はwav
→ 「RIFF waveform Audio Format」という
音声ファイルフォーマット

- ファイルには音声情報だけでなく、そのファイルのパラメータが記録されている
  - チャンネル数：モノラルなら１、ステレオなら２
  - **サンプリングレート（1秒間に何回記録するか）**：48kHzや44.1kHzなど
  - **ビット深度（1つのサンプルを何ビットで記録するか）**：8bit, 16bit, 24bitなど
- その他の音声ファイルフォーマット
  - Wav以外にもmp3, AAC (mpeg映像等に使われる）、WMA (Windows Meta Audio, マイクロソフトの音声ファイル形式）などがあり、AppleやSonyなどメーカ独自フォーマットもある
  - Wavは非圧縮フォーマット（情報欠損はないがデータサイズは大）、mp3、AAC, WMAなどは非可逆圧縮フォーマット（サイズは小さいが情報欠損が起きる）

演習： SoundProcessing1.ipynb

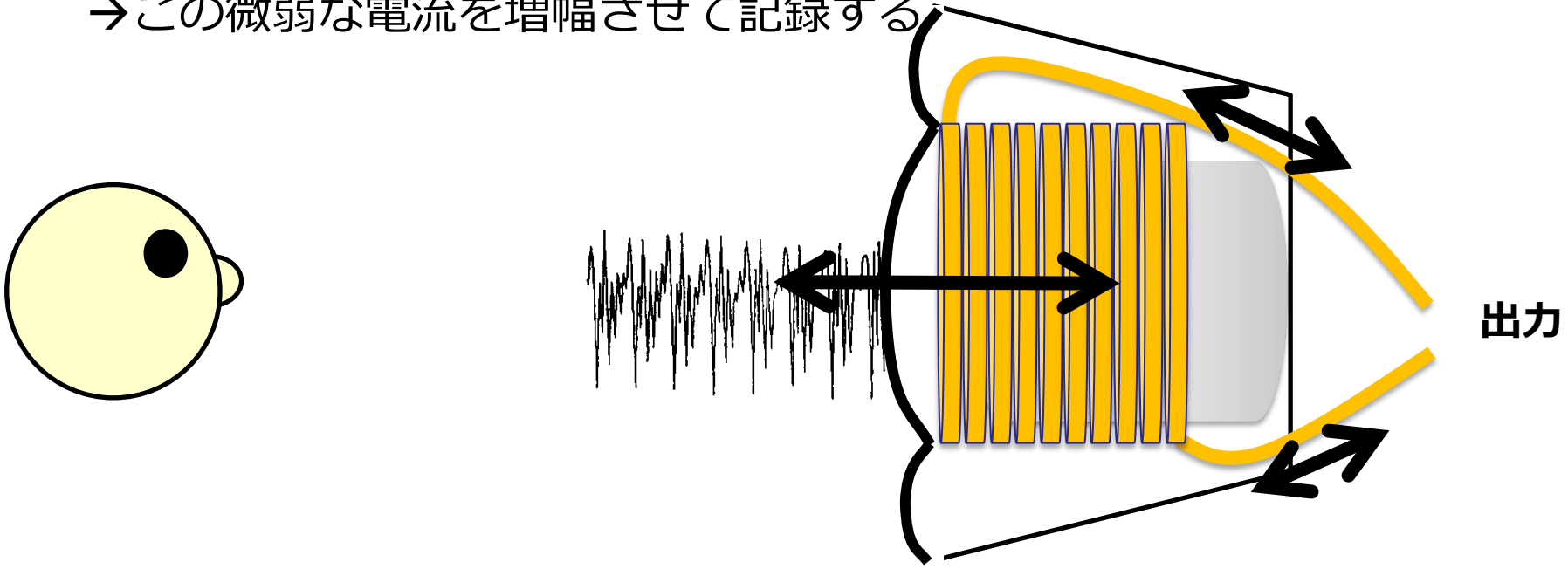
音声の収録
―ダイナミックマイク（ムービングコイル型）の原理―

- 振動板に音波（粗密波）が当たると振動板が揺れる
- 振動板が揺れるとコイルが揺れる
- 磁界に対してコイルが動くと電流が流れる
  →この微弱な電流を増幅させて記録する



音声の収録
–ダイナミックマイク（ムービングコイル型）の原理–

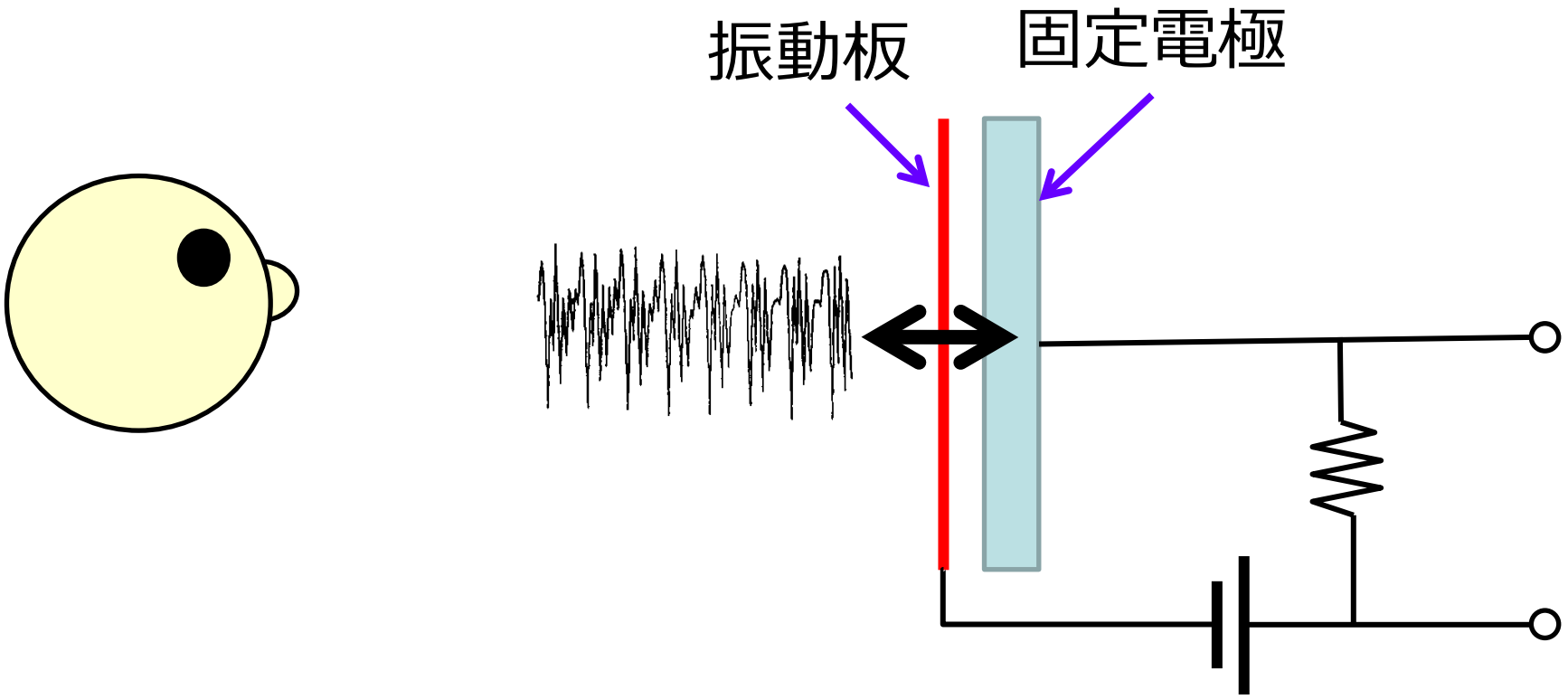
- 振動板に音波（粗密波）が当たると振動板が揺れる
- 振動板が揺れるとコイルが揺れる
- 磁界に対してコイルが動くと電流が流れる
  →この微弱な電流を増幅させて記録する



メリット：仕組みが単純で壊れにくい、電源が不要
デメリット：振動版にコイルが連結しているため、週録音に歪が生じやすい

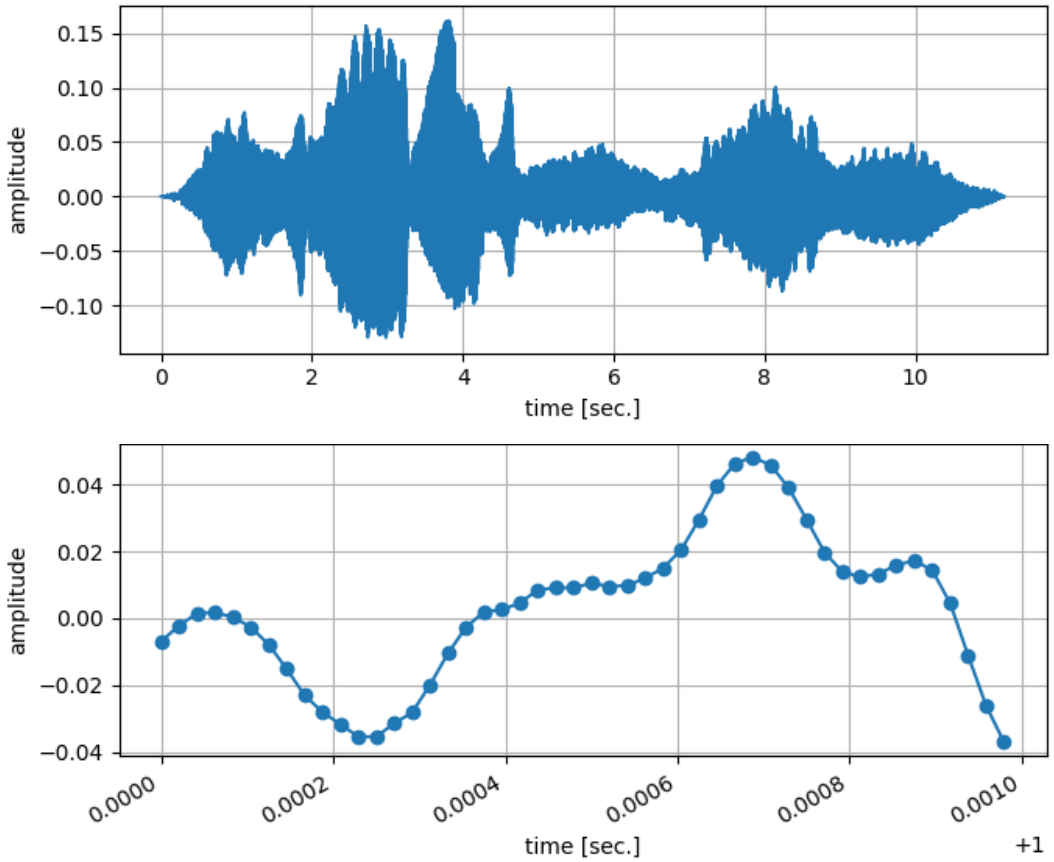
音の収録
–コンデンサーマイクの原理–

- 電源が必要（電源の切り忘れに注意！）
- フラットな周波数特性が得られやすい
- 温度や湿度の影響を受けやすい
  （マイクをたたいて音を確認するのは厳禁！）



音声ファイルを描画

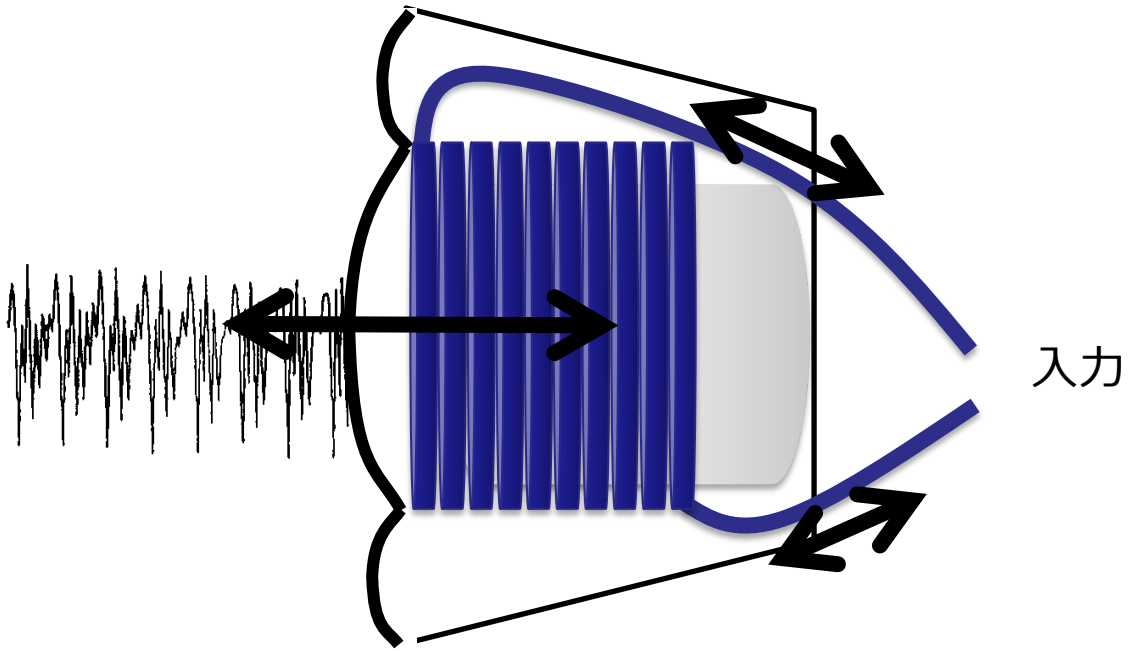
- X軸方向を経過時間に変換（サンプリングレートを使う）
- Y軸方向を理論的最大値で正規化（ビット深度を使う）
- 拡大してみよう



演習： SoundProcessing1.ipynb

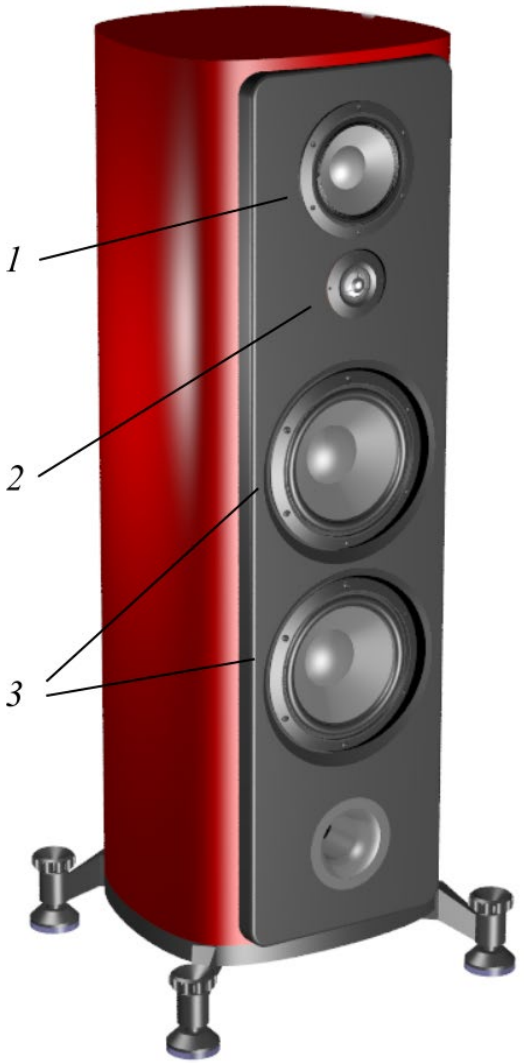
音の再生 –スピーカユニットの原理–

- ダイナミックマイクと同じ原理
- 振動板の前方と同じ音が後方にも放射されることに注意
  - 振動板の大きさに対し波長が十分長い音は前方と後方で
    うち消しあって音が発生しない→エンクロージャが必要



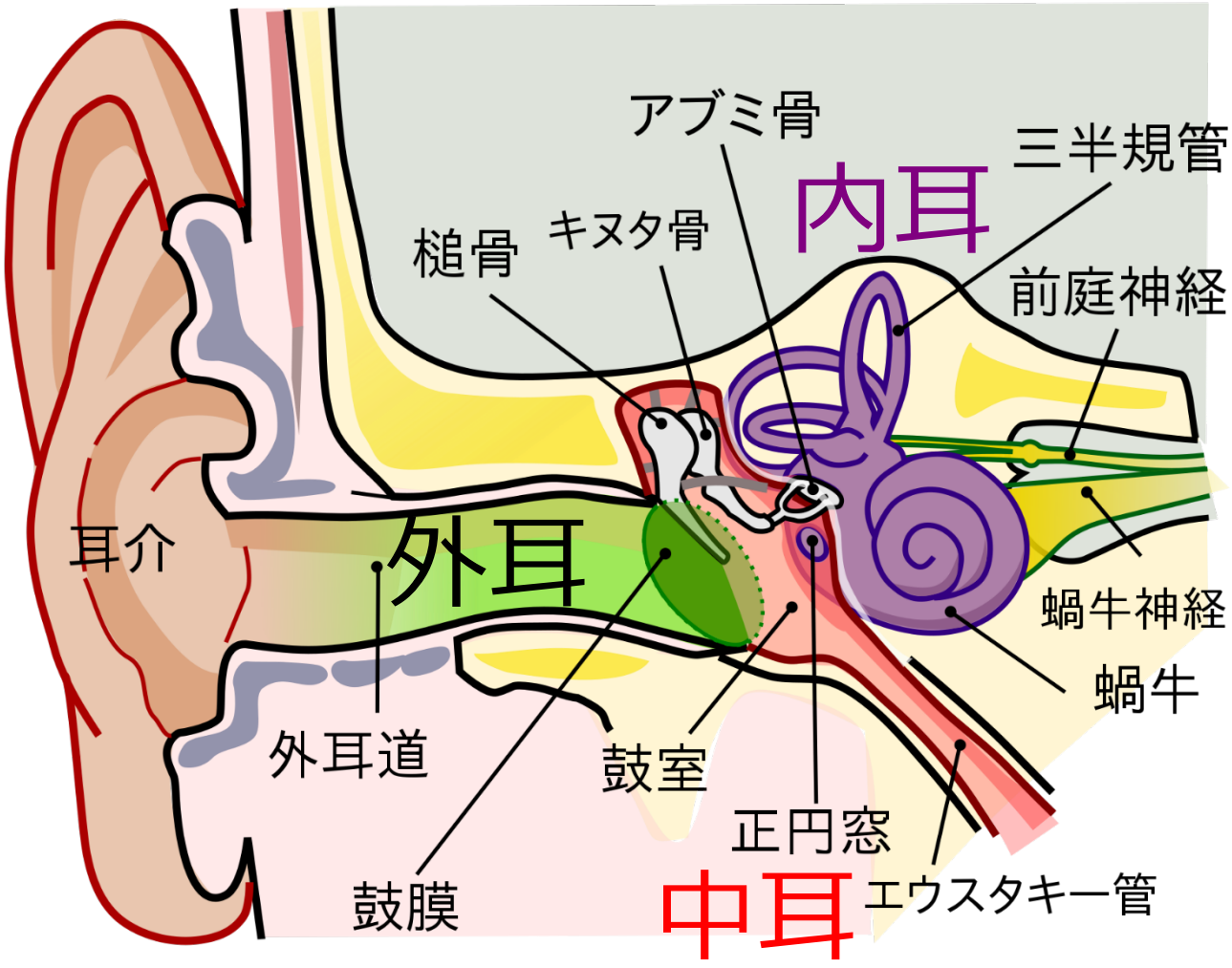
2way/3wayスピーカ

- 1つのスピーカユニットで全帯域の音を再生することは困難
- 2つ（2 way)もしくは3つ（3 way)のスピーカユニットで
周波数帯域を分担して再生
  1. Mid-range driver
  2. Tweeter (高音用)
  3. Woofers (低音用)



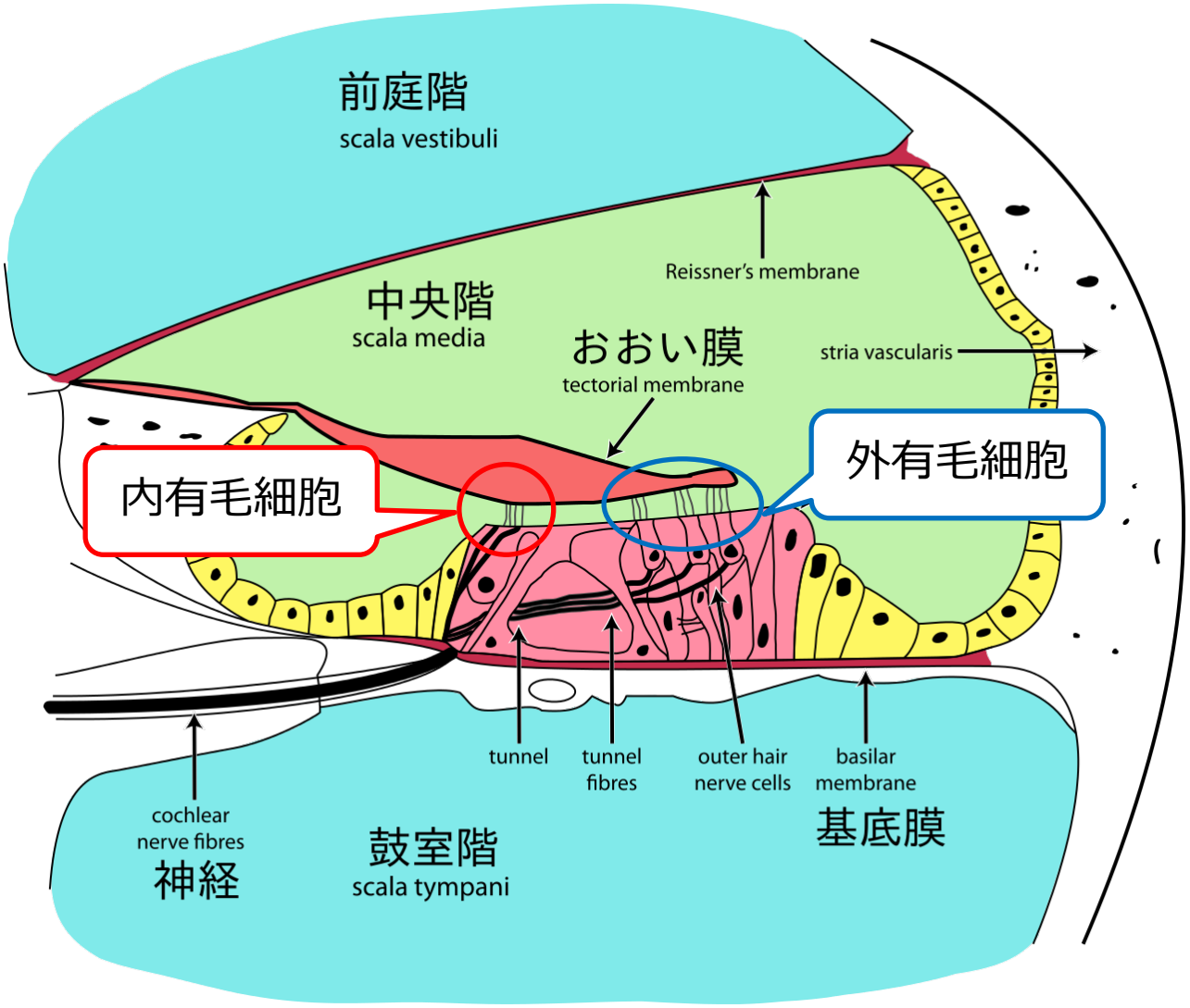
ref. (2020/06/09) File:Electrodynamic-loudspeaker.png, CC BY-SA 3.0, https://commons.wikimedia.org/wiki/File:Electrodynamic-loudspeaker.png

耳の構造



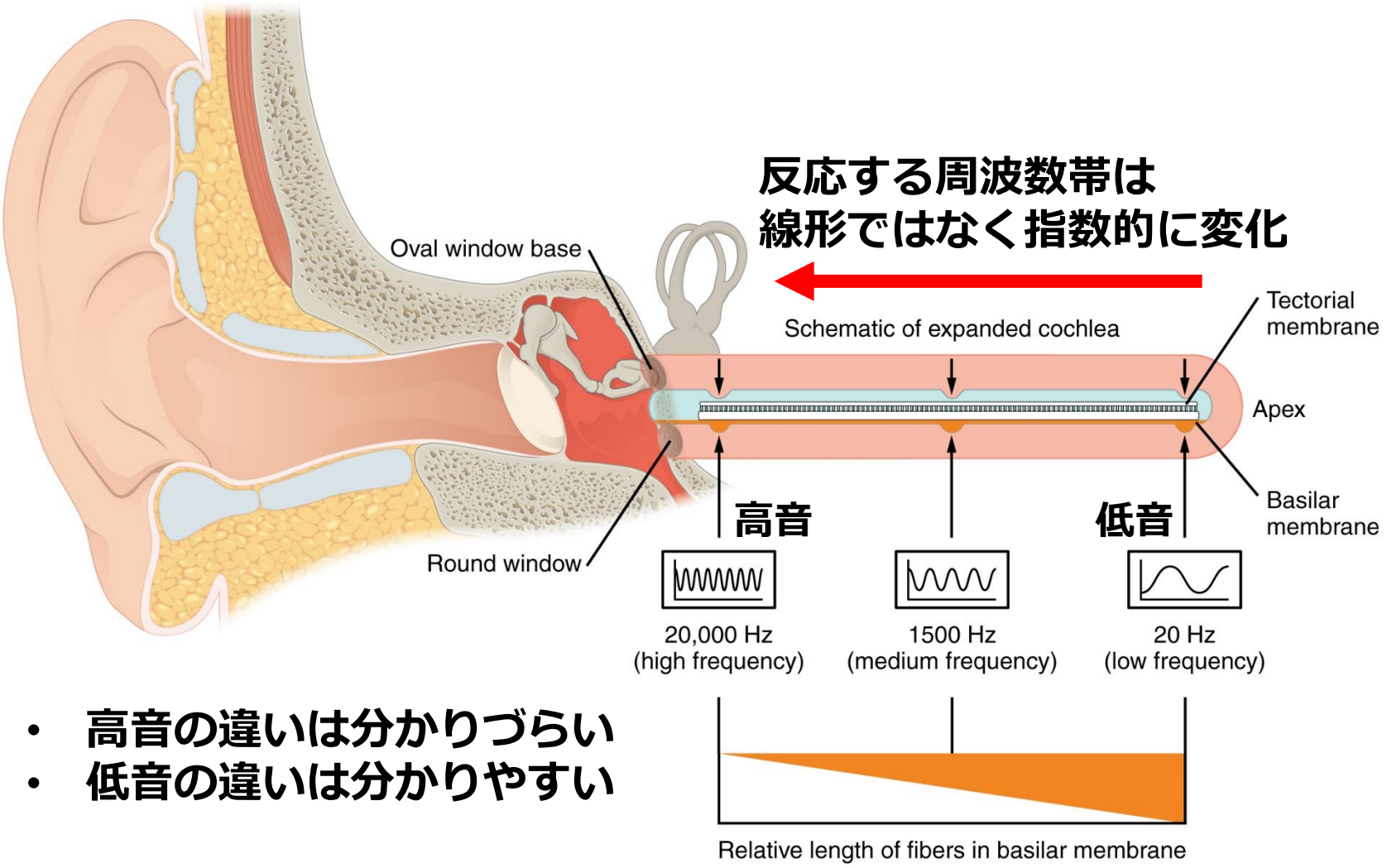
ref. https://commons.wikimedia.org/wiki/File:Anatomy_of_the_Human_Ear_ja_font.svg, CC 表示-継承 3.0

蝸牛の断面図



ref. https://commons.wikimedia.org/wiki/File:Cochlea-crosssection.png#/media/File:Cochlea-crosssection_jp.svg, CC BY-SA 3.0

蝸牛での周波数分析

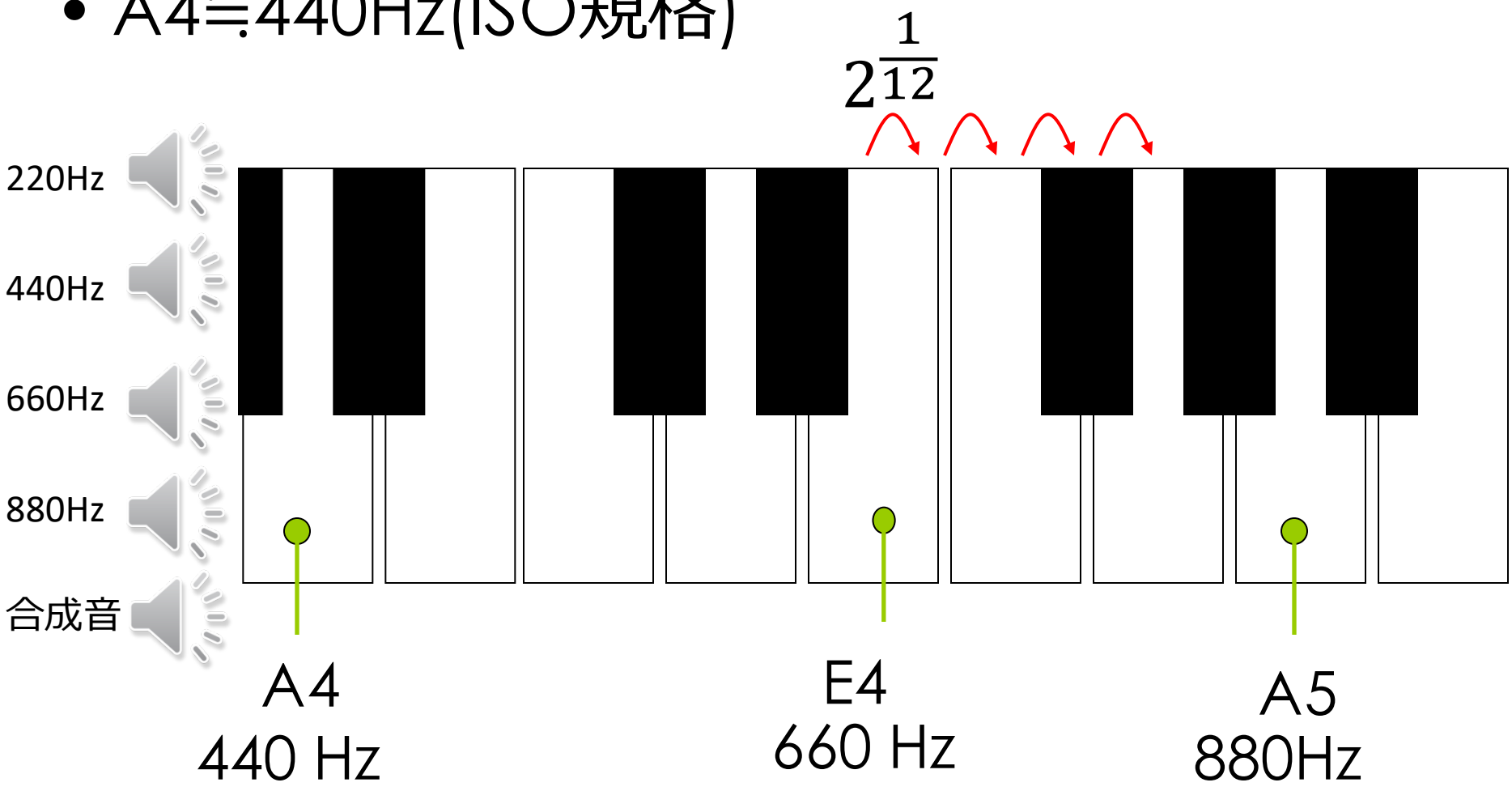


- 高音の違いは分かりづらい
- 低音の違いは分かりやすい

ref. https://commons.wikimedia.org/wiki/File:1408_Frequency_Coding_in_The_Cochlea.jpg, CC BY 4.0

音階の例

- A4≒440Hz(ISO規格)

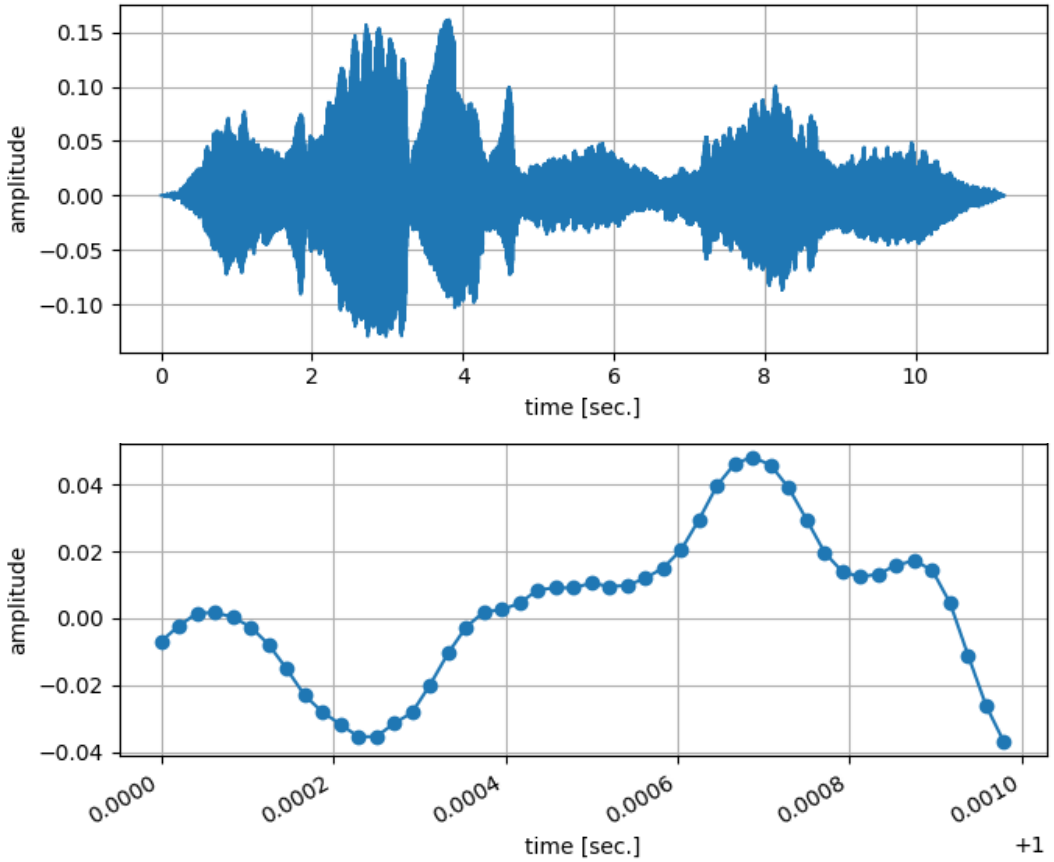


音の聞き比べ

- Pianoの単音（A3, A4, A5)
  - ref. (2020/06/09) University of Iowa Electronic Music Studios, http://theremin.music.uiowa.edu/MISpiano.html
- Shepard tone (無限音階)
  - ref. (2020/06/09) File:DescenteInfinie.ogg, https://en.wikipedia.org/wiki/File:DescenteInfinie.ogg

音声ファイルを描画

- X軸方向を経過時間に変換（サンプリングレートを使う）
- Y軸方向を理論的最大値で正規化（ビット深度を使う）
- 拡大してみよう



演習： SoundProcessing1.ipynb

アナログ-デジタル変換（AD変換）

標本化：サンプリングレート、フレームレート

- 時間的に連続するデータに対し、一定間隔ごとに記録する
- 単位時間あたり何回記録するか→サンプリングレート（単位はHz)

量子化：ビット深度、量子化ビット

- 1サンプルを何ビットで記録するか？
- 振幅方向の階調の粒度に影響
  例）1サンプルあたり16bitで表現するなら、2^{16}=65536階調
  ただし、波形は正と負があるため、正負それぞれの階調はその半分

音に限らず、AD変換する際は常に生じる問題！

標本化：
サンプリングレートの影響

演習：SoundProcessing2.ipynb >
1. サンプリングレートの影響

疑問：サンプリングレートはいくつにしたらいいの？

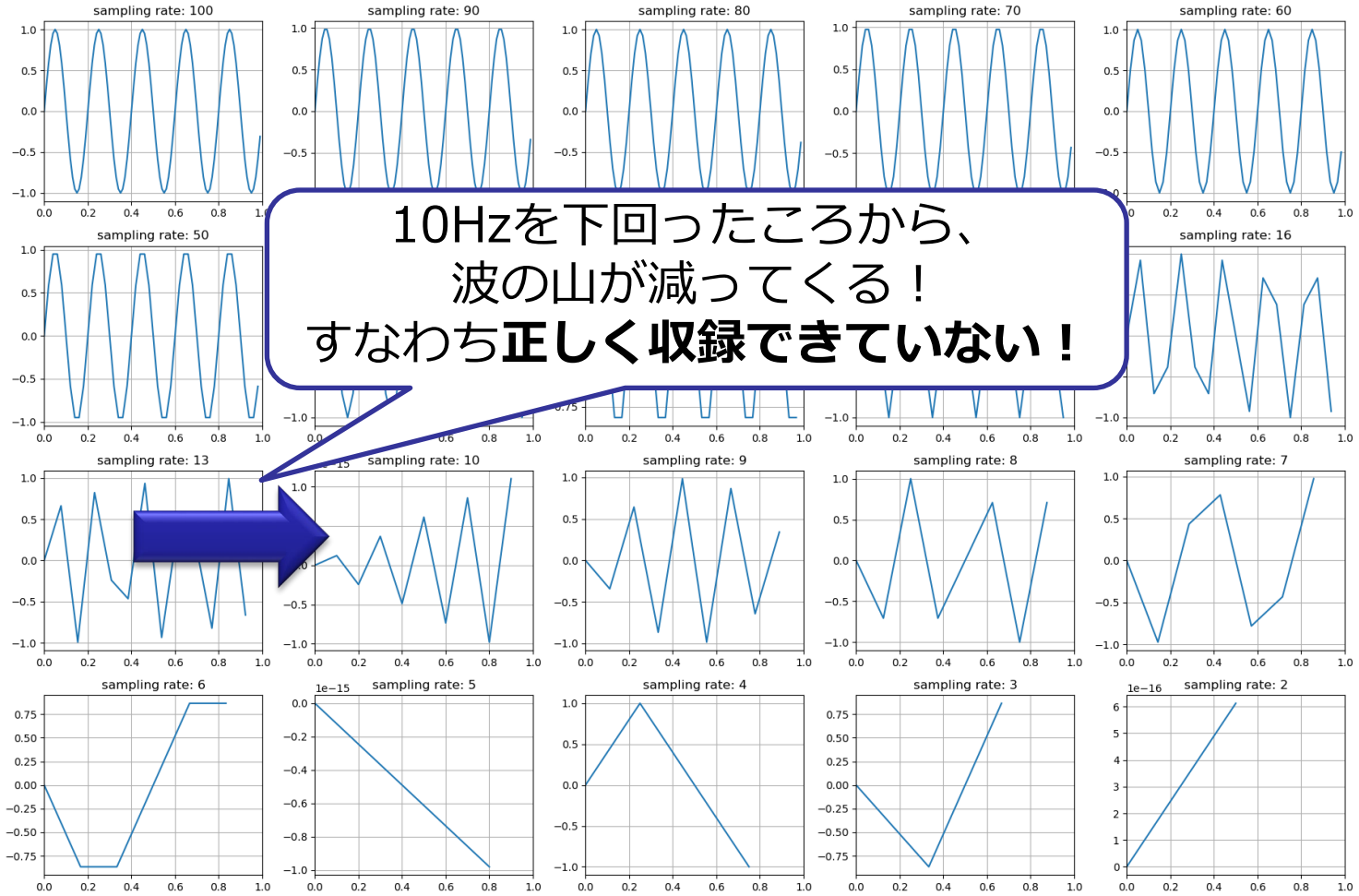
答え：

収録したい音の周波数帯域によって決まる
（どれくらい高い音＝高周波の音まで収録したいのか？）

- 4kHzまででいい
  → 8kHzで収録すればいい
- 人間の可聴帯域は高々24kHz程度と言われている
  → 48kHzで収録すればいい

音の周波数とサンプリングレートの影響

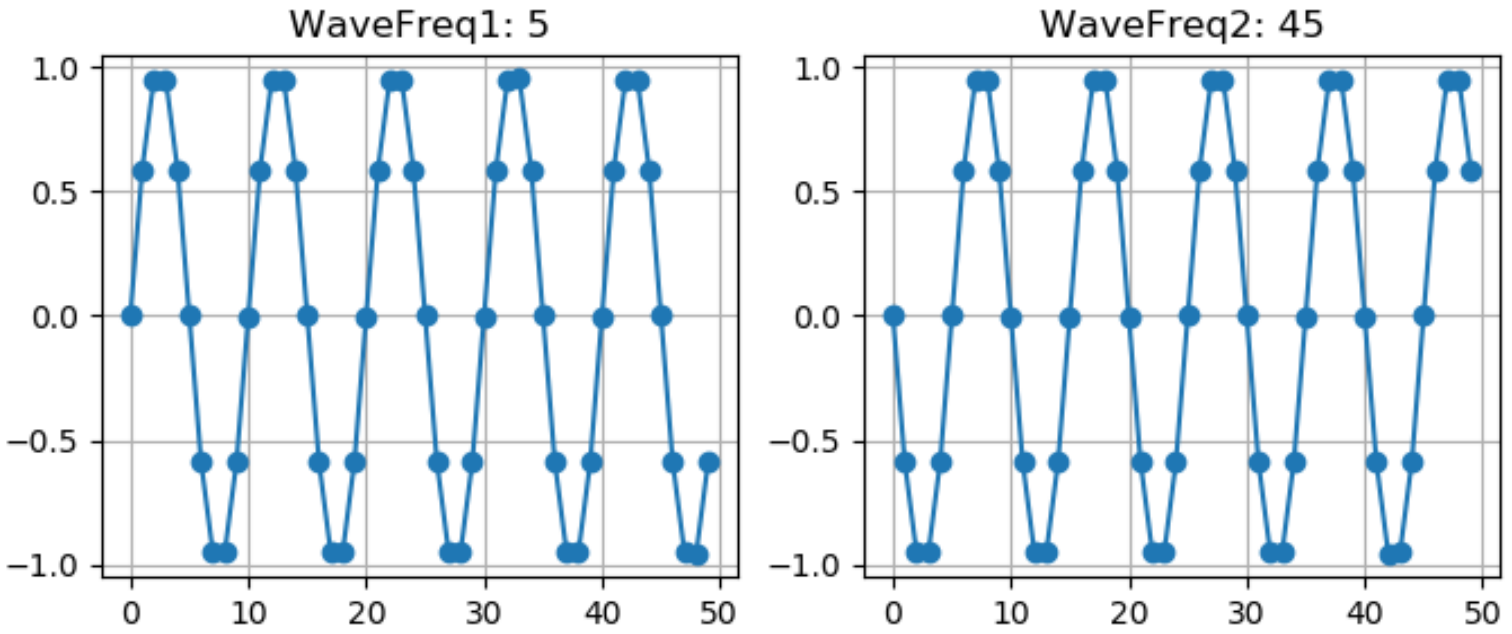
- すべて5Hzのサイン波＆1秒間
- サンプリングレートを下げていくとどうなるか？



演習： SoundProcessing2.ipynb > **1. サンプリングレートの影響**

エイリアシング現象の演習 1

50Hzで収録したとき、高さ5Hzのサイン波と45Hzのサイン波が同じ波形になる！

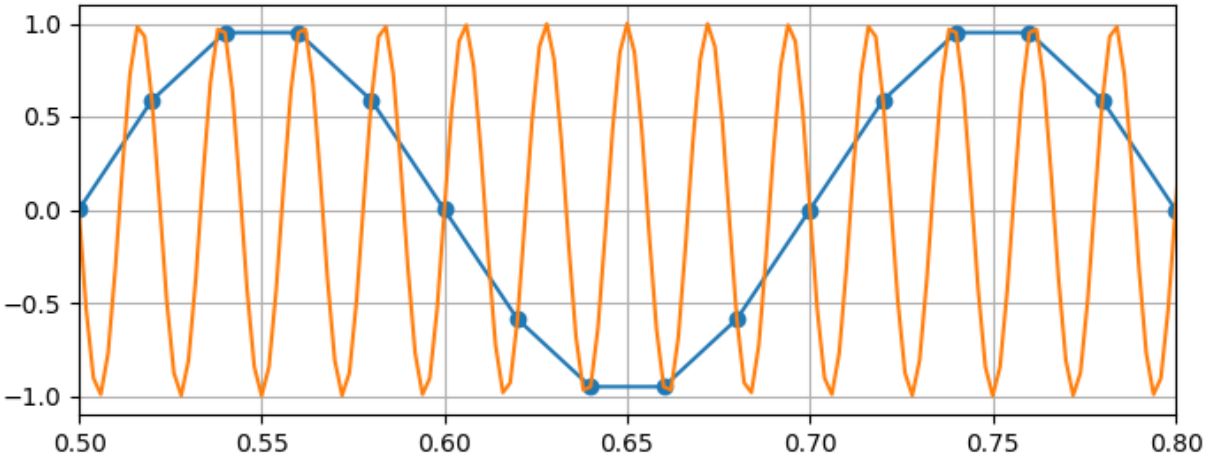


エイリアシング現象の演習２

なぜ45Hzの音が5Hzの音と同じ波形になってしまったのか？

- オレンジの波は、十分高い周波数で収録した45Hzの波
- 青の波は、サンプリングレート50Hzで収録した45Hzの波

サンプリングレートが45Hzだと、サンプルを取るタイミングが5Hzの音をサンプルしたときの位置とぴったり一致！！



45Hzの音が5Hzの音として記録される
→**エリアシングノイズ**あるいは**折り返し雑音**と呼ぶ

サンプリング定理

- サンプリングレートがFHzのとき f Hzの波は $(F-f)$ Hzの波と区別できない
- サンプリングレートがFHzのとき、$F/2$よりも高い周波数の波 f を収録してしまうと、 それは $(F-f)$の周波数の波と混ざって記録される
  → **折り返し雑音（aliasing noise)**と呼ぶ

逆に言えば・・・

収録対象に含まれている音（通常は様々な周波数の波が混ざった混合音）の最大周波数が f_{max} Hzである場合、これを記録するためには $F > 2f_{max}$ であるようなサンプリングレート Fで記録しなければならない！

→サンプリング定理

- サンプリングレートの1/2の周波数を**ナイキスト周波数**と呼ぶ

音の周波数

- 音の収録における注意点（まとめ）
  1. 収録したい音のうち最高周波数の2倍よりも高いサンプリングレートで収録する
  2. ナイキスト周波数よりも高い周波数成分は、**あらかじめアナログフィルタなどで除去**しておく
  3. ナイキスト周波数よりも高い周波数成分はエイリアシングノイズになる
- 人間の可聴(弁別)帯域はおよそ50Hz～20kHz
  - 健康診断でテストするのは1kHzと4kH
    4kHzが聞こえなくなると、音声が聞こえづらくる
  - 年齢が上がるにつれて高周波が聞こえなくなる
- サンプリングレート40kHzで収録すれば十分
  - CDのサンプリングレートは44.1kHz
  - mp3のサンプリングレートは32kHz, 44.1kHz, 48kHz

量子化：
ビット深度の影響

演習：SoundProcessing2.ipynb >
2. ビット深度の影響

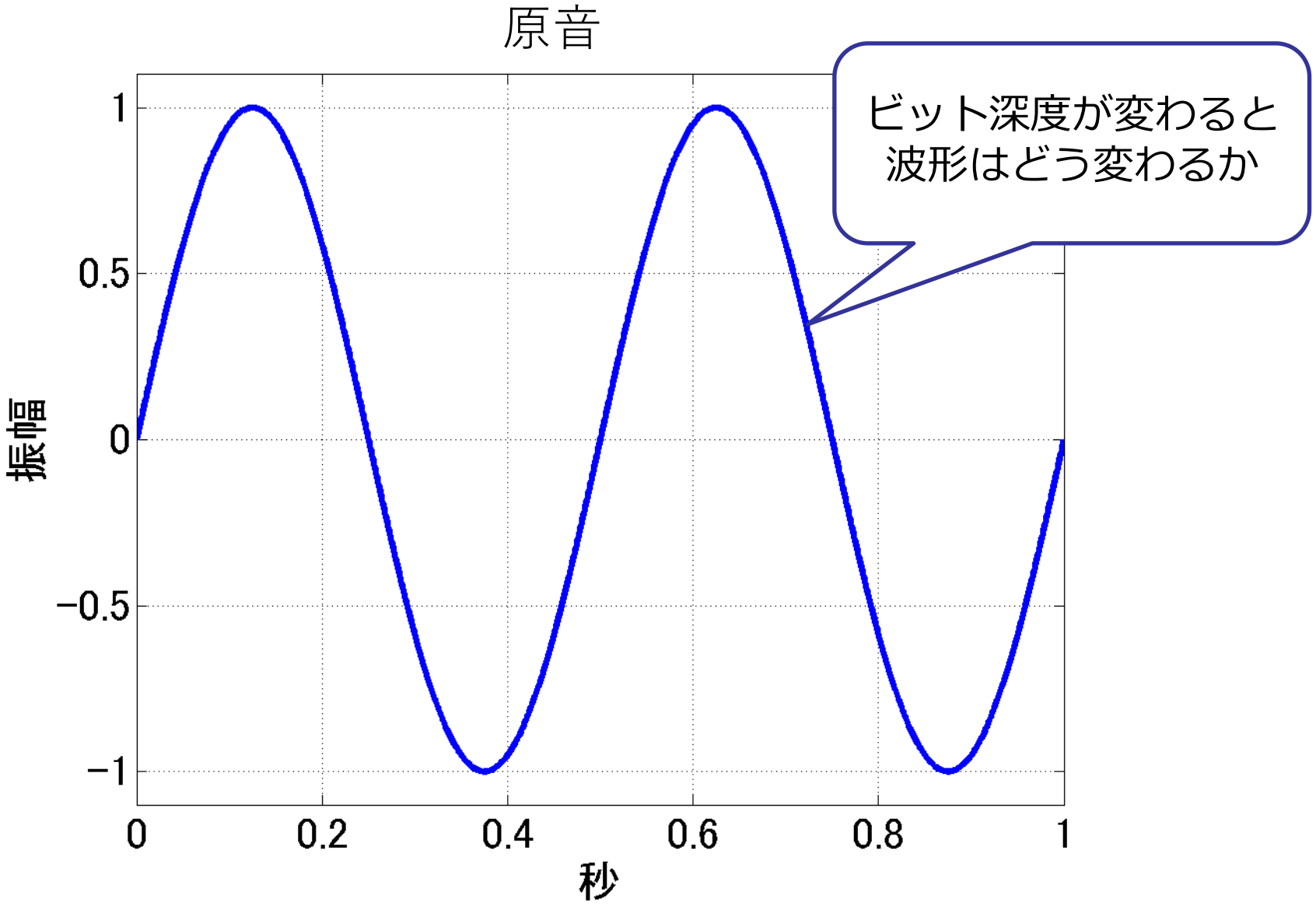
疑問：ビット深度はいくつにしたらいいの？

答え：

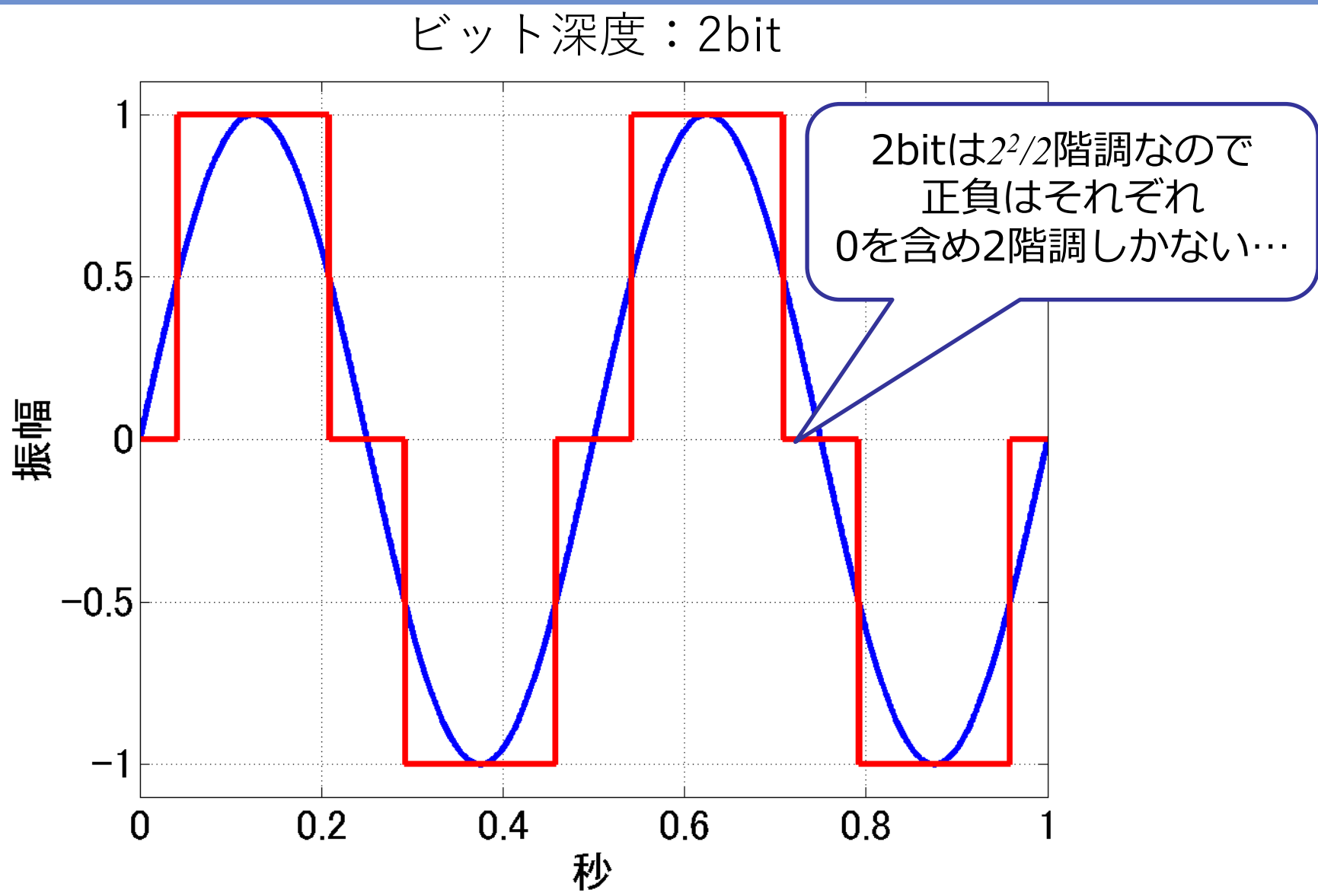
収録したい音の大きさのダイナミックレンジによる
（ダイナミックレンジとは、最小の音と最大の音の差）

- 音として収録されているのは相対的な波形
  （絶対的な音の大きさは通常記録されない）
- ビット深度が小さいと、音を大きくすると、本来は聞こえないほど小さな音であるはずのノイズが聞こえてしまう

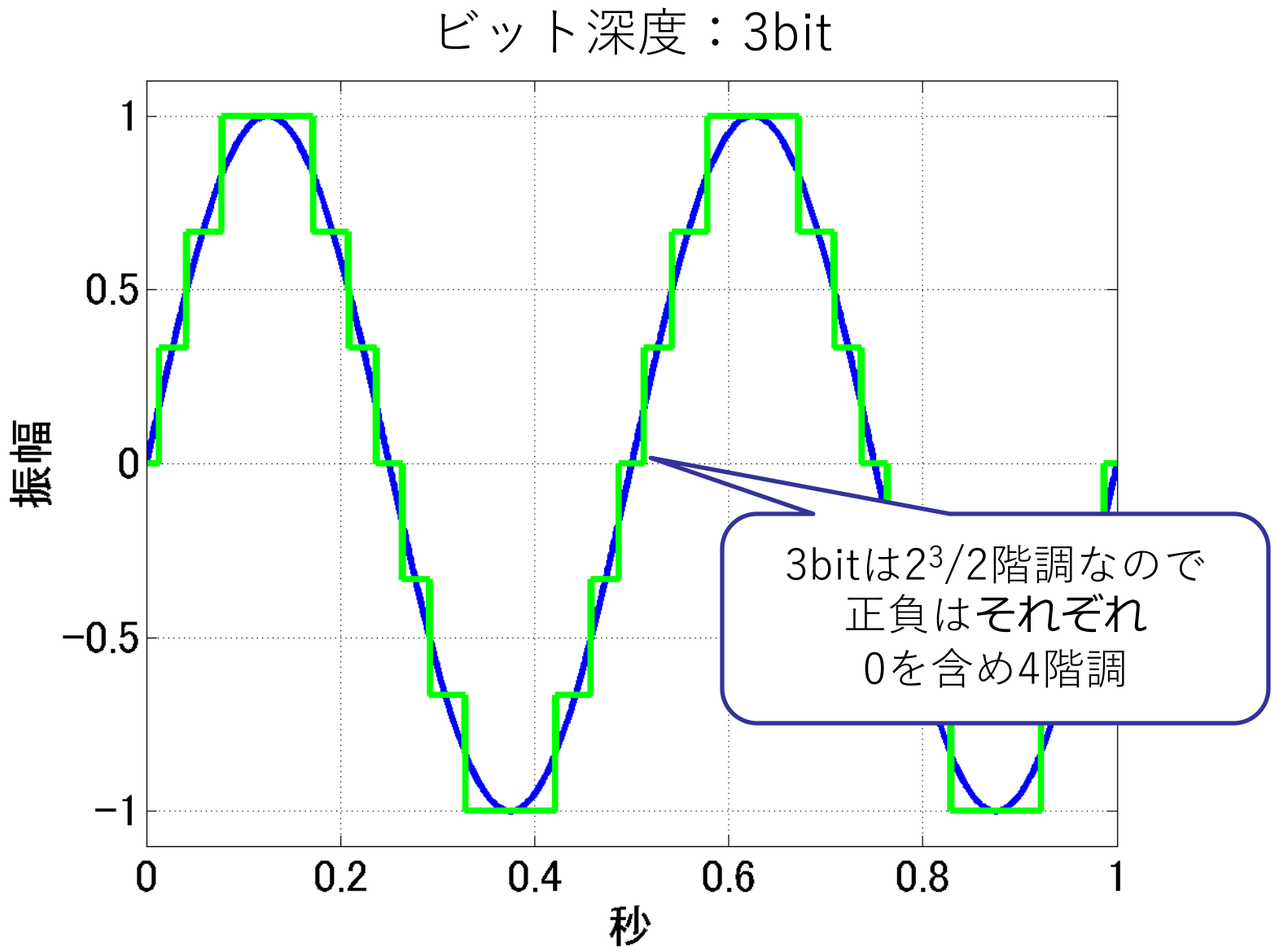
ビット深度が変わるとどう変わるか



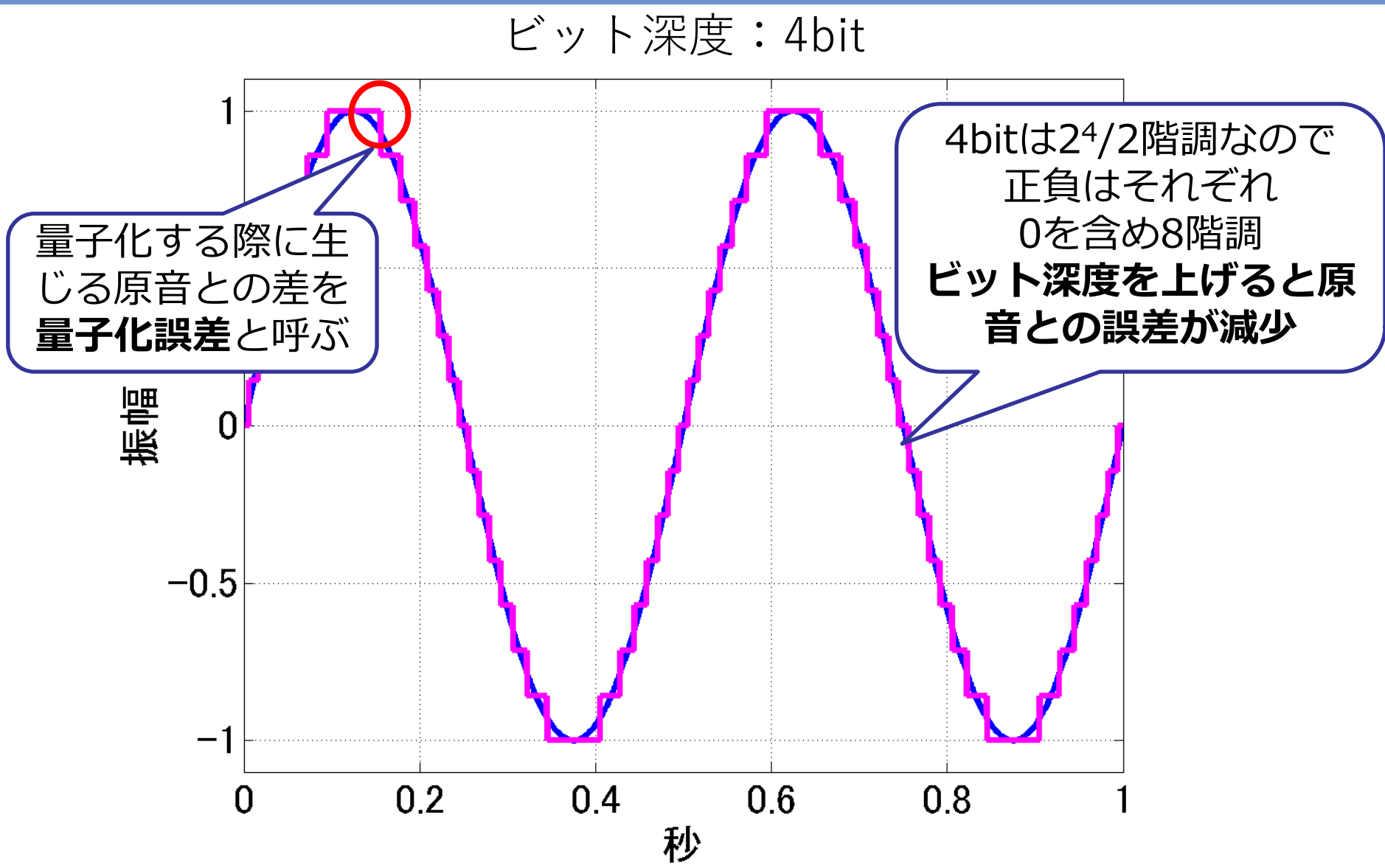
ビット深度が変わるとどう変わるか



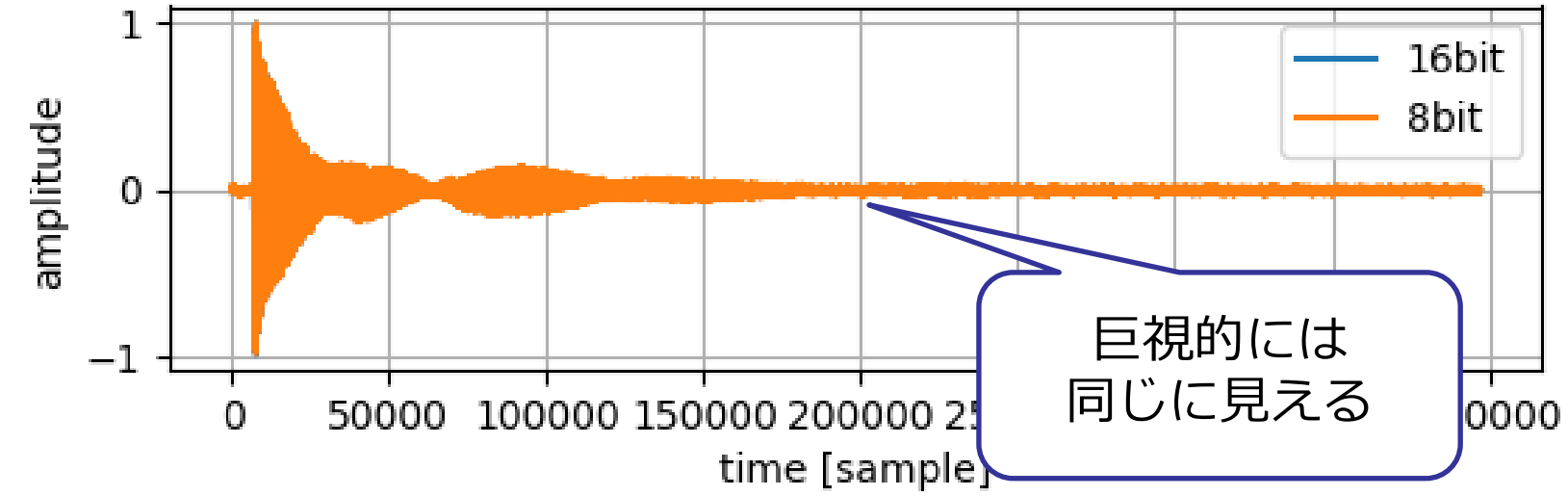
ビット深度が変わるとどう変わるか

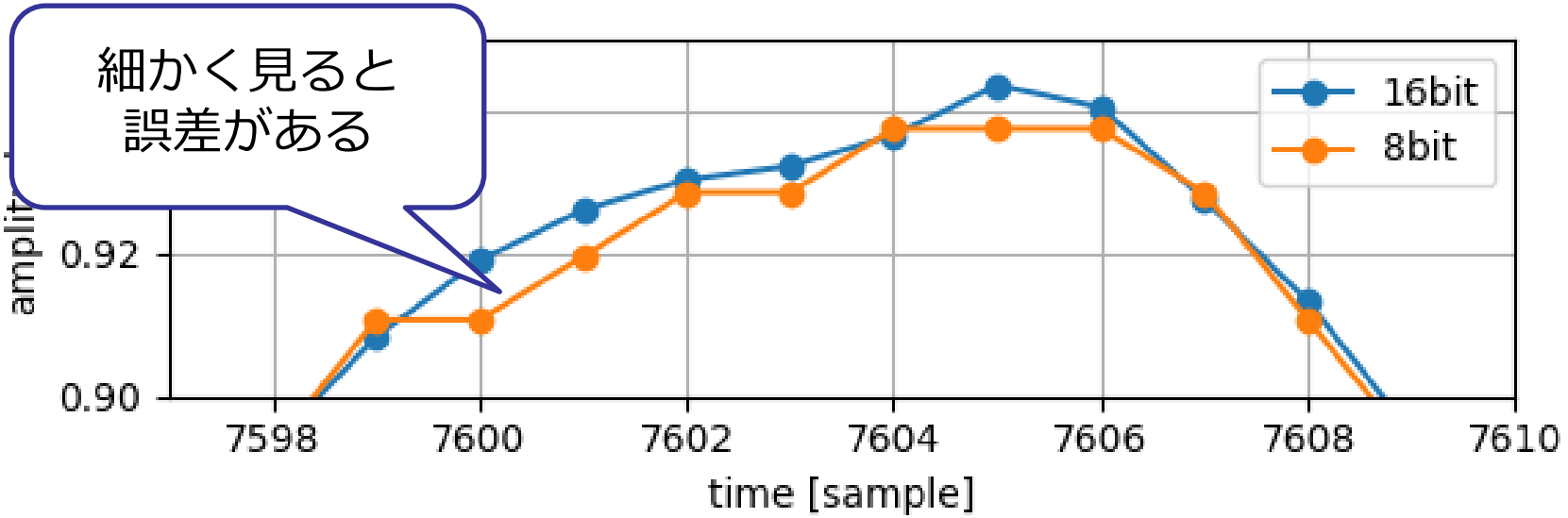


ビット深度が変わるとどう変わるか



実際の波形でビット深度が8-bitと16-bitの比較





演習： SoundProcessing2.ipynb > **2. ビット深度の影響**

ビット深度が浅いと何が起こるの？

- 本来の音と違う音が混じってくる（サーという雑音が聞こえる）→量子化雑音
- ダイナミックレンジ（最も小さい音と最も大きい音の差）が狭くなる

音の大きさ・音圧

記録されている音量は相対値

- 通常のマイクで収録できる音の大きさは相対値でしかない！
  - 収録の段階で振幅を調整するので、絶対的な値に意味がない
    （絶対音量＝音圧を記録したいなら校正された騒音計が必要）
- 相対的な信号の強度を表す単位: **dB (デシベル)**
  - 基準の信号と比較して、どれくらい大きいかによって大きさを表す

騒音計を使って
音圧レベルを測定

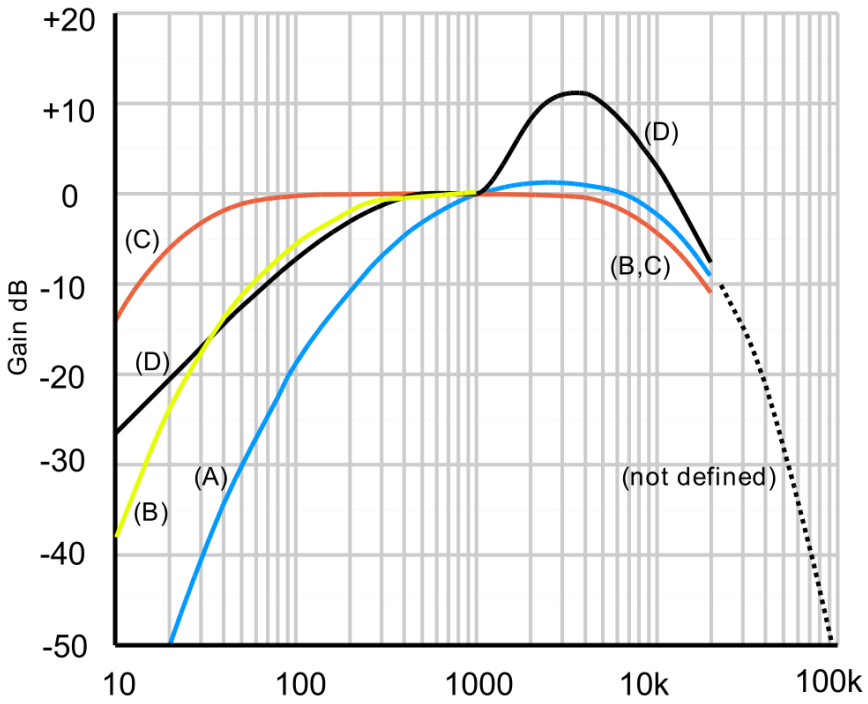
ref. https://en.wikipedia.org/wiki/Sound_level_meter#/media/File:Optimus_Sound_Level_Meter.jpg, CC BY-SA 3.0

信号の相対的な大きさの単位：デシベル (dB)

- 電圧・電流・音で使われる単位
  （ｄＢ）=20ｘ$\log_{10}$（対象の圧／基準圧）
- dBは相対値の単位
  - 振幅が元の2倍：$20\log_{10}2 = 6$ で6dB
  - 振幅が元の４倍：$20\log_{10}4 = 12$dB
  - 振幅が元の10倍：$20\log_{10}10 = 20$dB
  - 振幅が元の半分：$20\log_{10}1/2 = -6$dB
  - 振幅が元の1/4：$20\log_{10}1/4 = -12$dB
- 10dB上がると音の大きさは2倍になったように感じる（振幅値は約3.16倍）
- 音の場合、基準音圧は、通常の人の耳に聞こえる最小音=20μPaとする
- **通常のマイクは音圧を記録できないため、基準圧は1とする**

人間の聴覚による音の大きさ：A特性とC特性

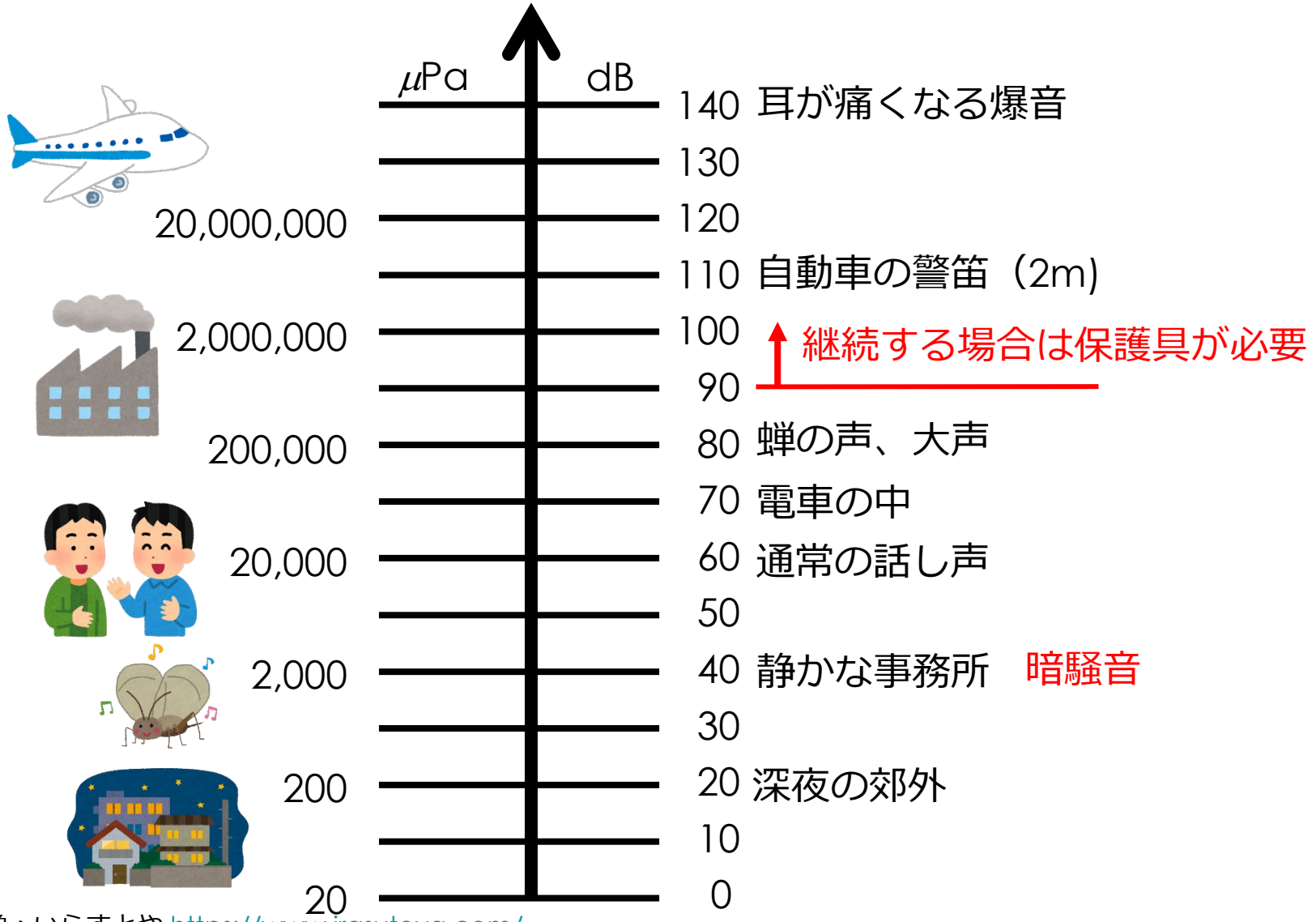
- 通常、騒音計ではA特性とC特性の2種類の測定方法がある
- 人間の聴覚は低音や高音は聞こえない
  → 低音・高音を平均的な聴力に合わせてカット
- 騒音（人間にとってうるさい音）レベルを測定する際は、通常、A特性を使う



A-weighting (blue), B (yellow), C (red), and D-weighting (blk)

ref. File:Acoustic weighting curves (1).svg, https://commons.wikimedia.org/wiki/File:Acoustic_weighting_curves_(1).svg, Public Domain

様々な環境音のA特性音圧レベル



ref. 画像：いらすとや https://www.irasutoya.com/

どの程度のダイナミックレンジが必要なの？

- ダイナミックレンジとは、人間の耳で感じられる最も小さい音から最も大きなおとまでの差
- 1-bitは6dBを表現
- 人間の聴覚は120dBのダイナミックレンジを聞き分けれられる

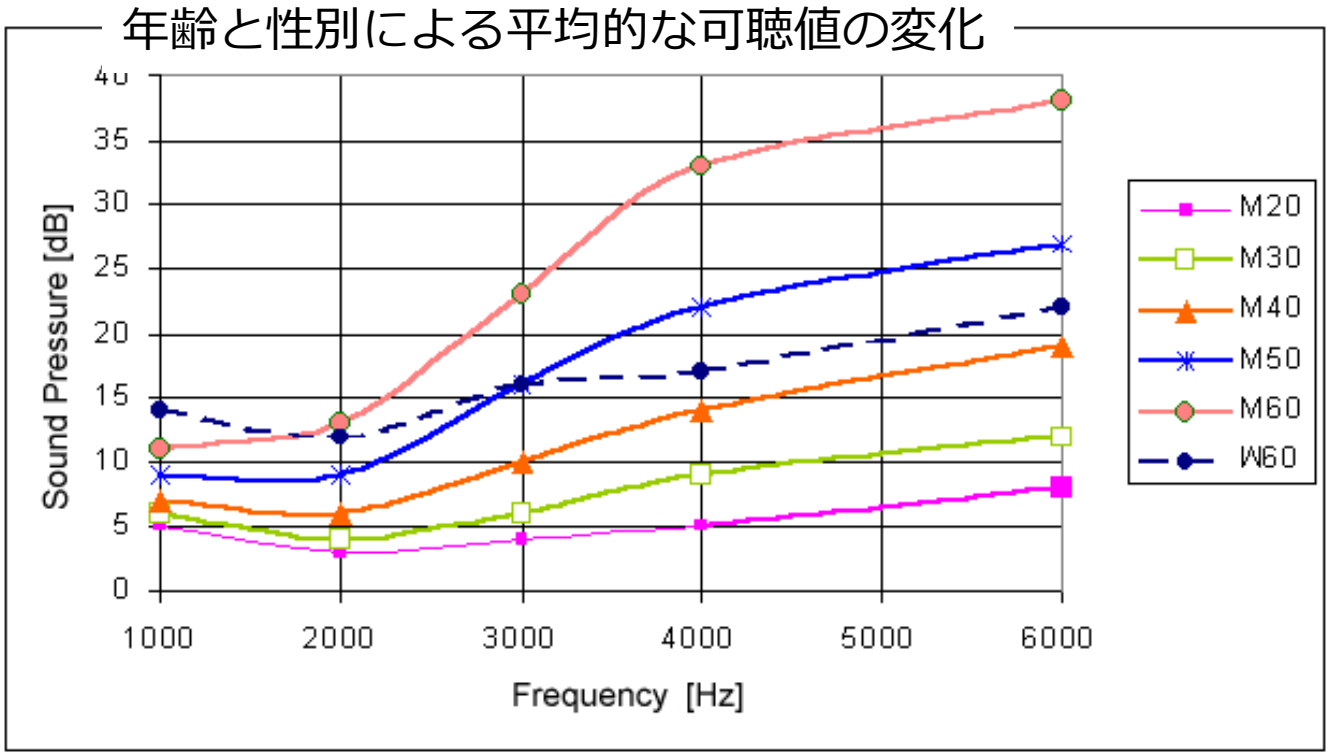
  →20-bitが必要（8の倍数とするので24-bit)

  - CDでは16-bitのダイナミックレンジは98.09dB
    →本当はちょっと足りない
  - DVDは24-bitまで可能、ダイナミックレンジは146.25dB
    →これなら人間は原音との差を知覚できないよう音量調節できるはず

年齢による聴力の衰え（男性・女性）

- 年齢が上がると高周波が聞こえなくなっていく
- 男性と比べると女性の聴力の低下は比較的穏やか

参考：母音を聞き取るためには4kHzまで、
子音を聴き取るためには8kHzまでの情報が必要

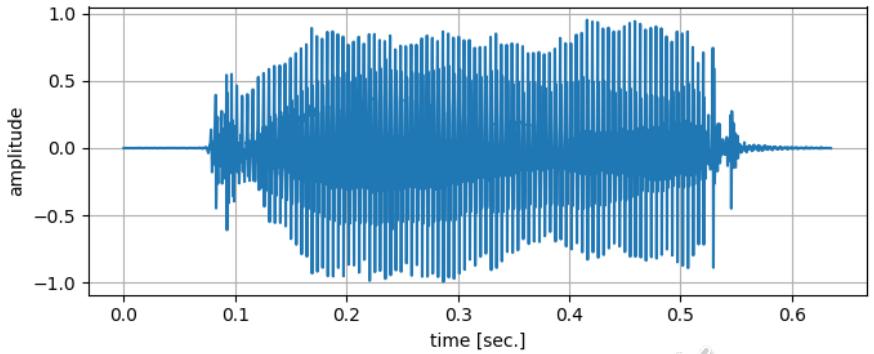


ref. https://commons.wikimedia.org/wiki/File:Ath-byage.png, CC 表示-継承 3.0

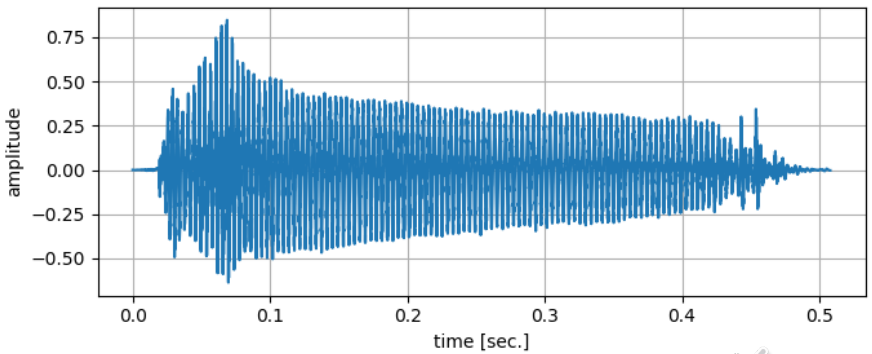
周波数分解

演習：SoundProcessing2.ipynb >
1. サンプリングレートの影響

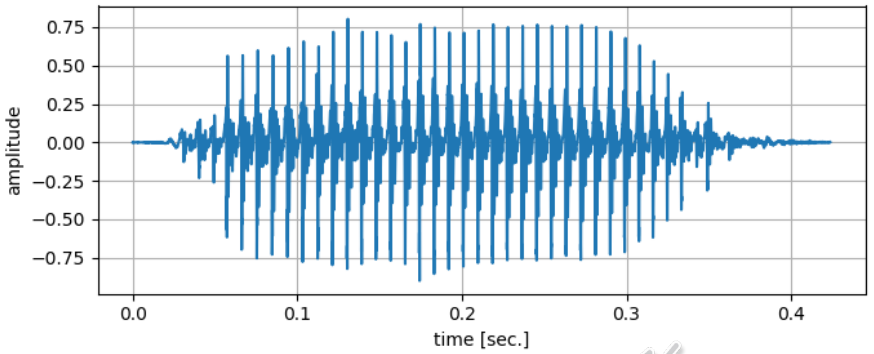
音声波形



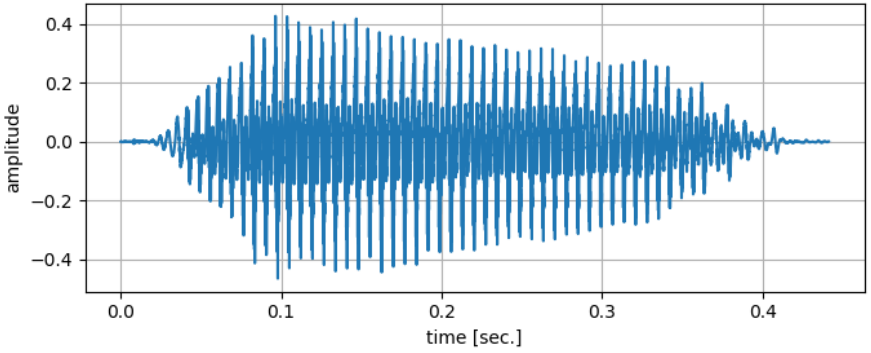
女性音声「あ」



女性音声「い」



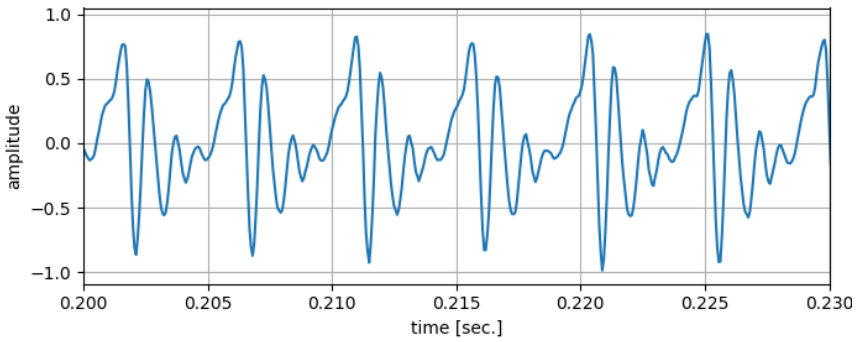
男性音声「あ」



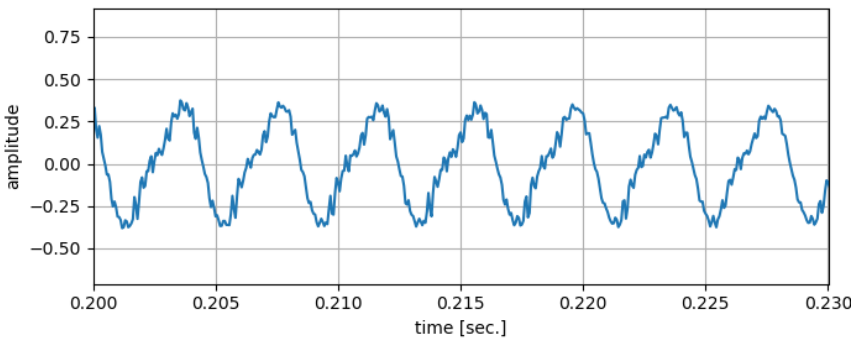
男性音声「い」

音声認識では左の波形を「あ」、右の波形を「い」と判別できます
どうやって分析したらいいでしょうか？

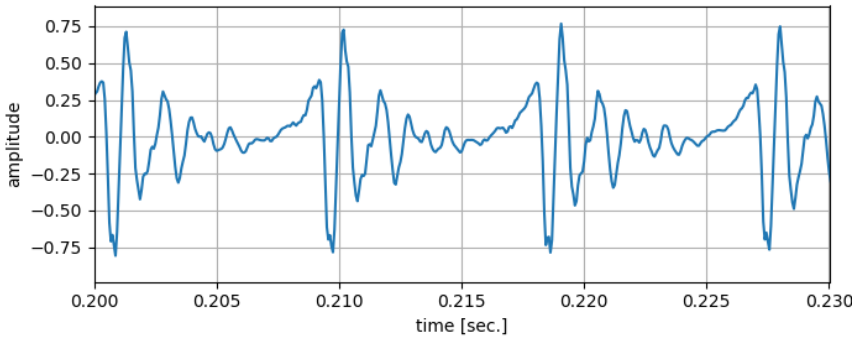
音声波形（拡大）



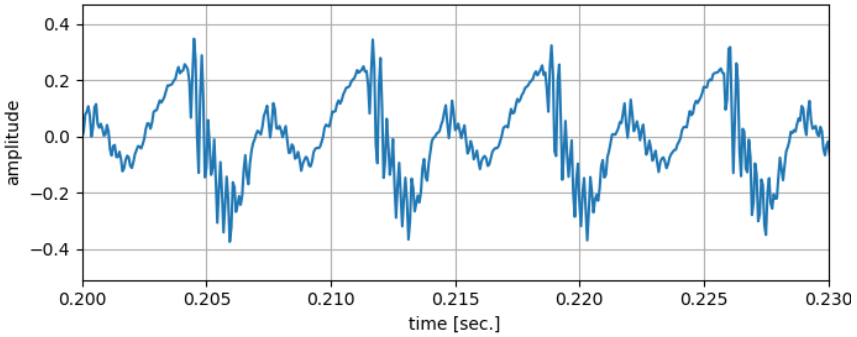
女性音声「あ」



女性音声「い」



男性音声「あ」



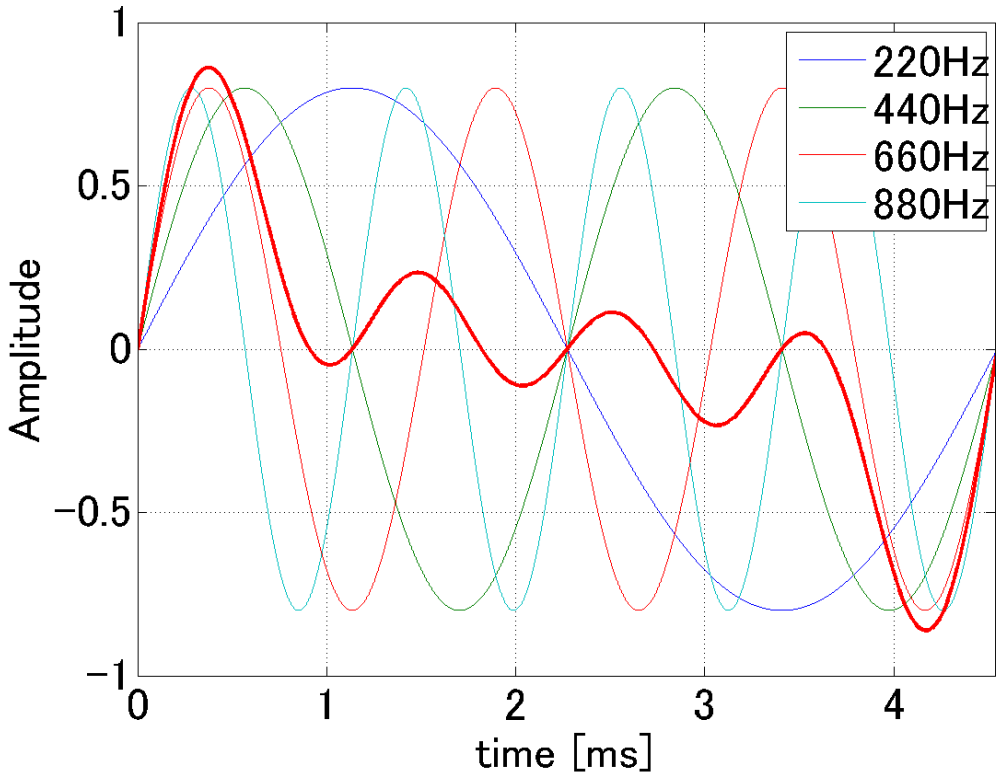
男性音声「い」

拡大するとなんとなく波形は似て見えますが…

音の高さごとに波を分解してみましょう！→ **周波数分解**

音は正弦波の重ね合わせと考えることができる

あらゆる音は振幅・位相・周波数（波長）によって
決まるサイン波を重ね合わせたものと考えることができる



220Hz

440Hz

660Hz

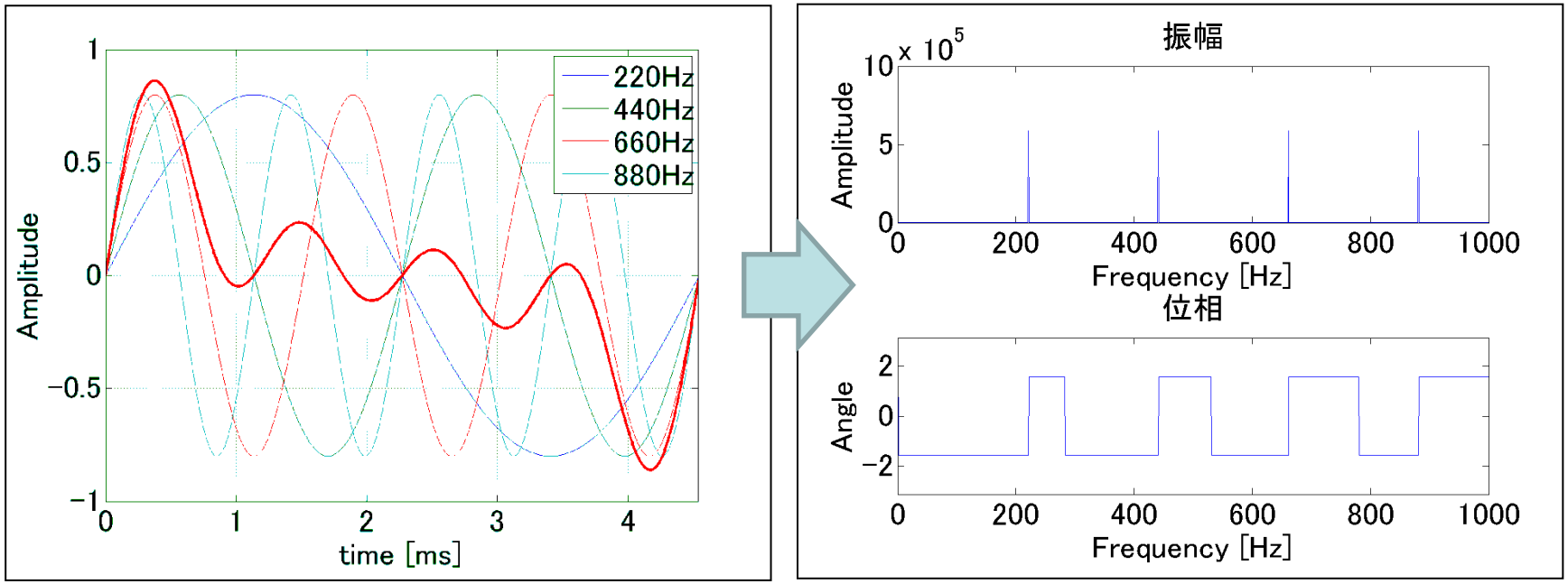
880Hz

合成音

周波数スペクトル

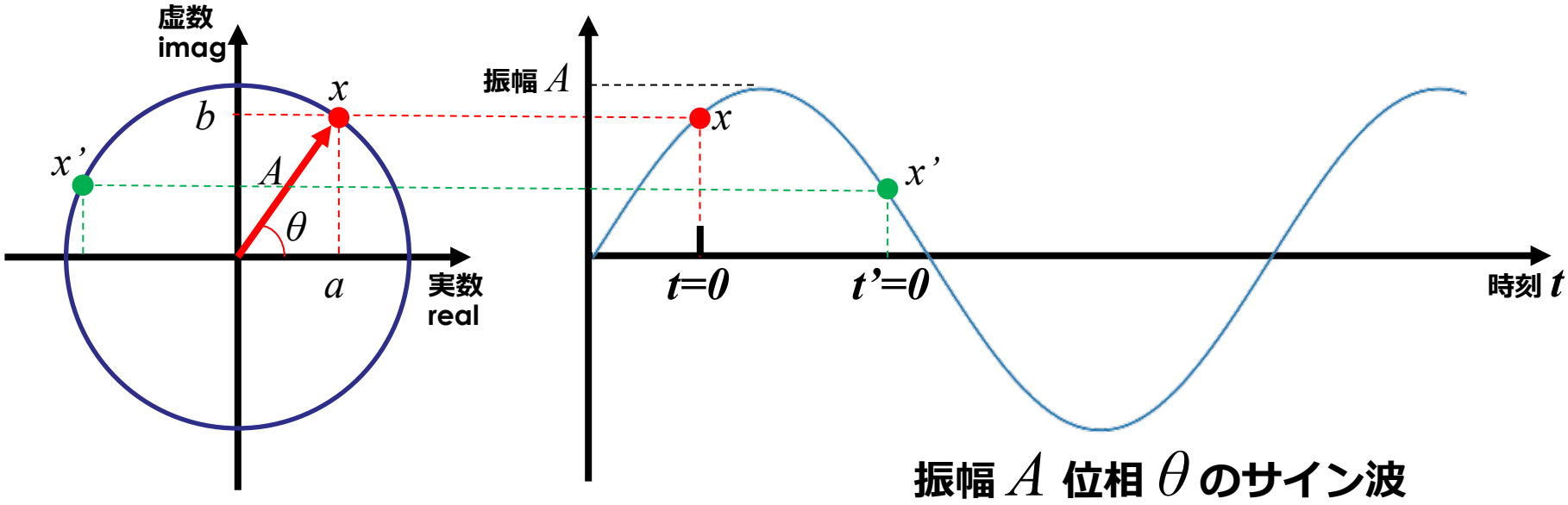
音を周波数ごとに分解して、その各周波数の振幅や位相で特徴を表現

→ **周波数スペクトル**と呼ぶ



周波数分解における振幅と位相

- ある純音は周波数・振幅・位相の組み合わせで一意に決まる
- 振幅と位相は複素数で表現されることがある
  - 下の図は振幅が A 、位相が θ の波形
  - 時刻$t=0$のときの波の値 $x = a + bi$
  - このとき、振幅は $A = \sqrt{a^2 + b^2}$ 、位相は$\theta = \tan^{-1} {}^{b}/_{a}$
    ここで$\tan^{-1}(x)$ とは$\tan(y) = x$となるような角度 yのこと

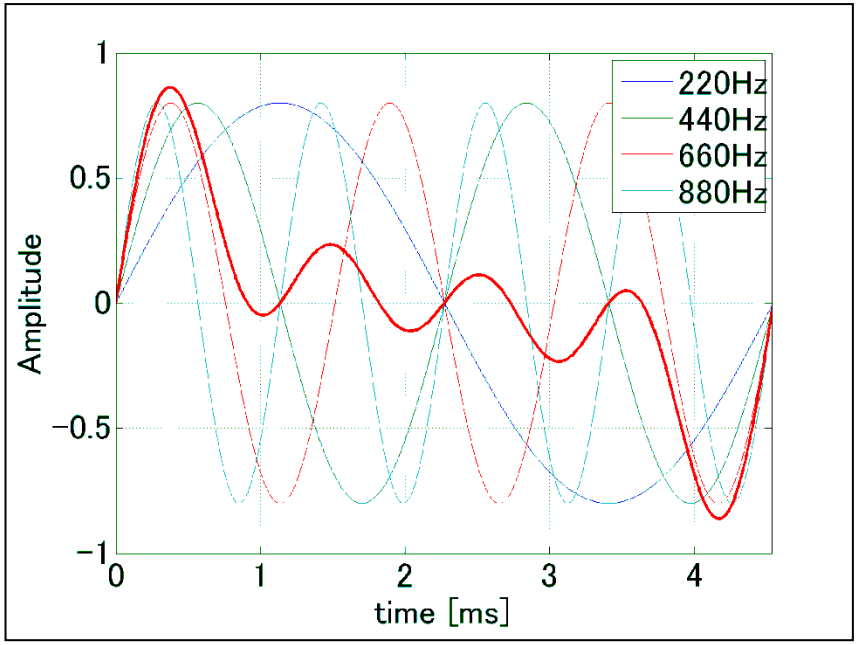


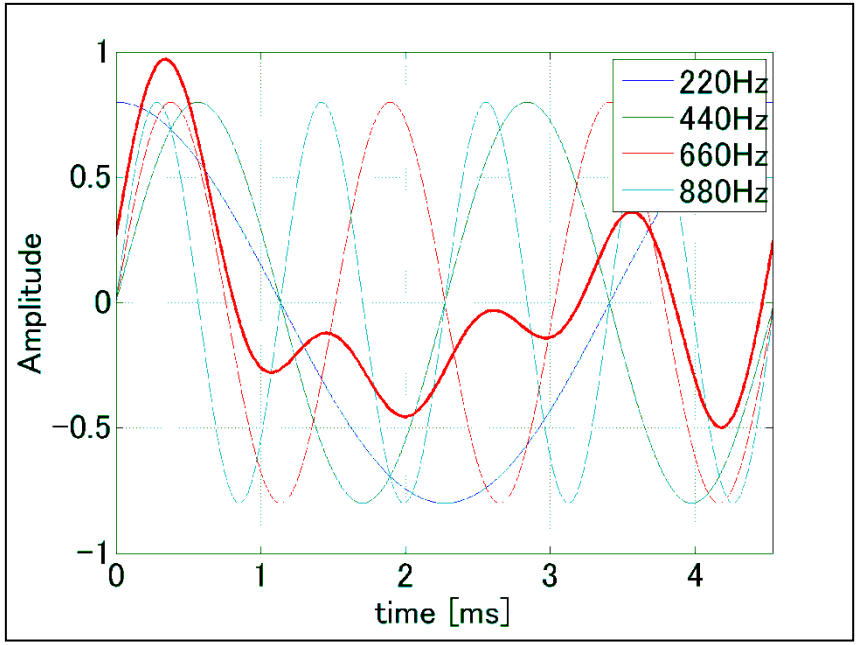
聴感における位相の影響

位相が違っても聴感には影響しない

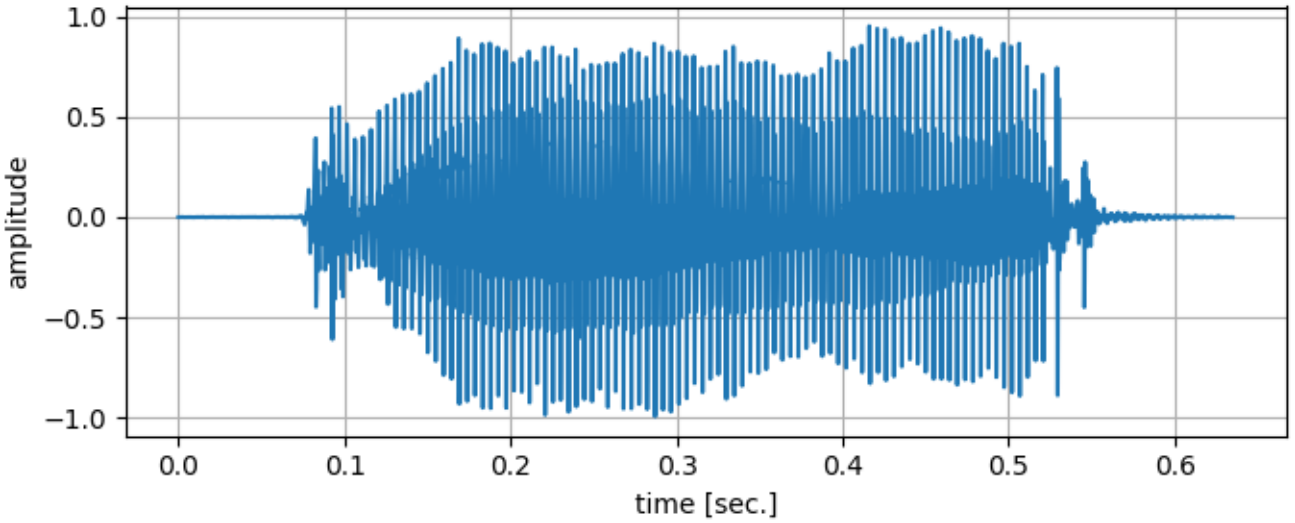
→ 聴覚で判別可能な音の特徴は振幅成分だけ見ればいい

220Hzの波の位相が
左のグラフより90度ずれている



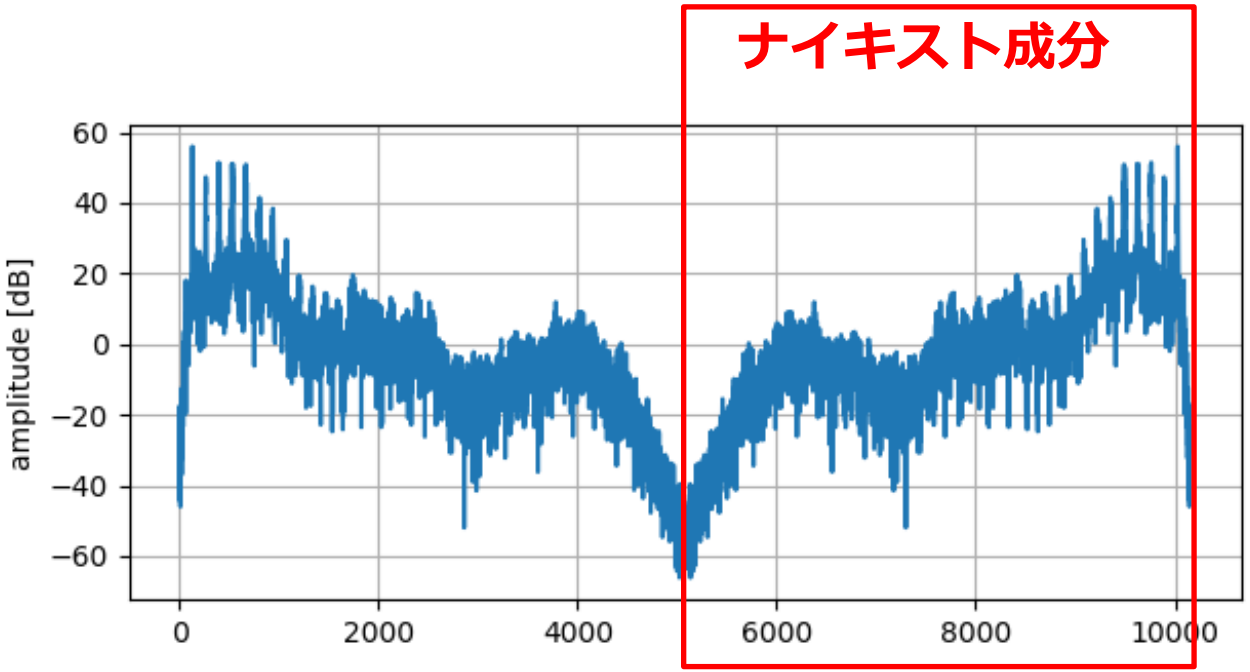


女性音声「あ」の波形



音声の周波数成分：ナイキスト成分の除去

- pythonではnumpyのfftモジュールを利用することで、高速フーリエ変換（Fast Fourier Transform, FFT）を実行できます
- 振幅成分は周波数成分をXとすると$20log_{10}(X)$でdB値に変換

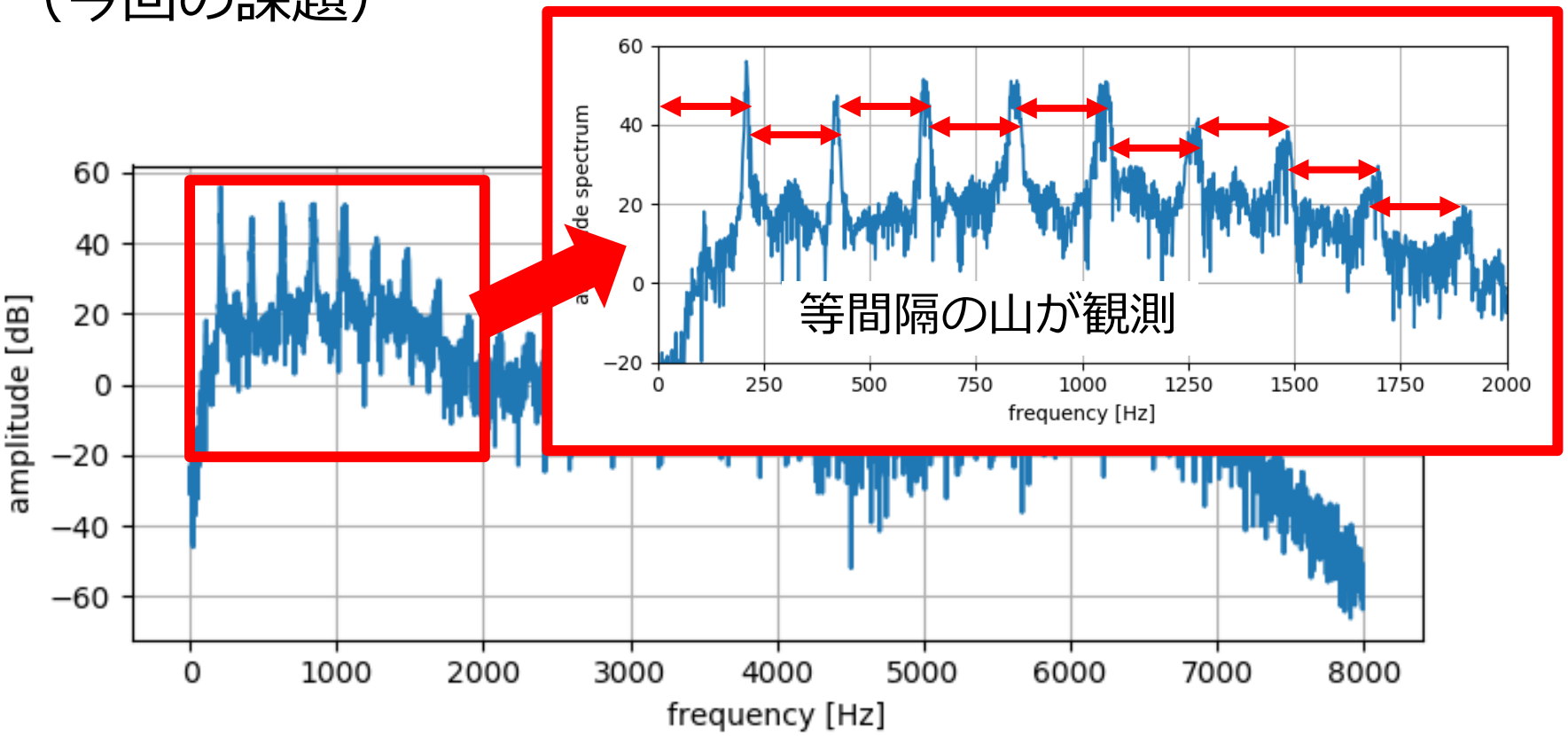


この音声ファイルのサンプリングレートは16kHz
→ 8kHz以上はナイキスト成分なので無視していい

演習： SoundProcessing3.ipynb > **2. 音声の周波数成分**

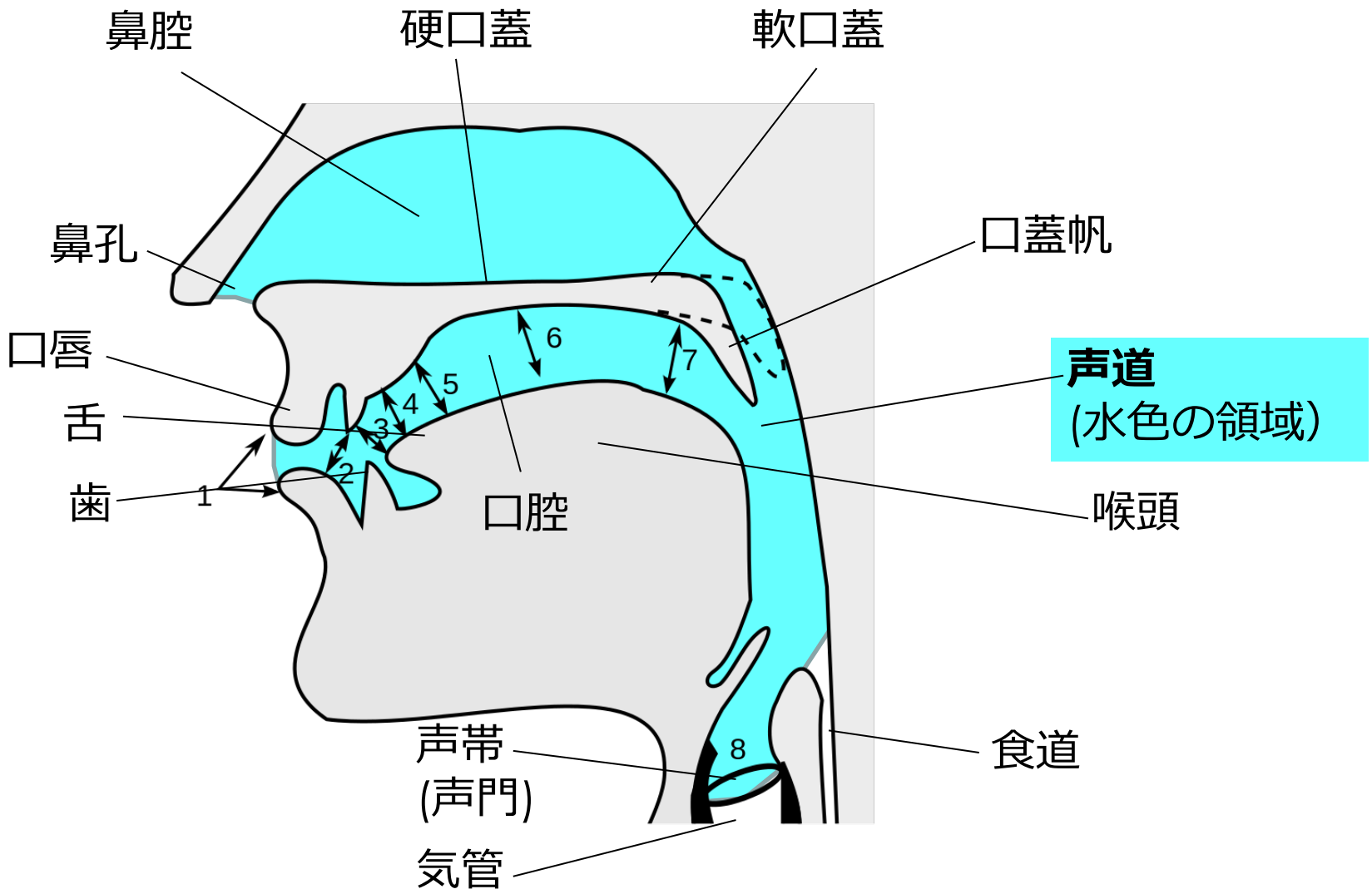
音声の周波数成分：音声の振幅成分

- 特に低周波領域に一定間隔の山が見える
  → 主に声帯で生じた倍音成分
- 楽器の単音でも同じ一定間隔の波が観測できます
  （今回の課題）



演習： SoundProcessing3.ipynb > **2. 音声の周波数成分**

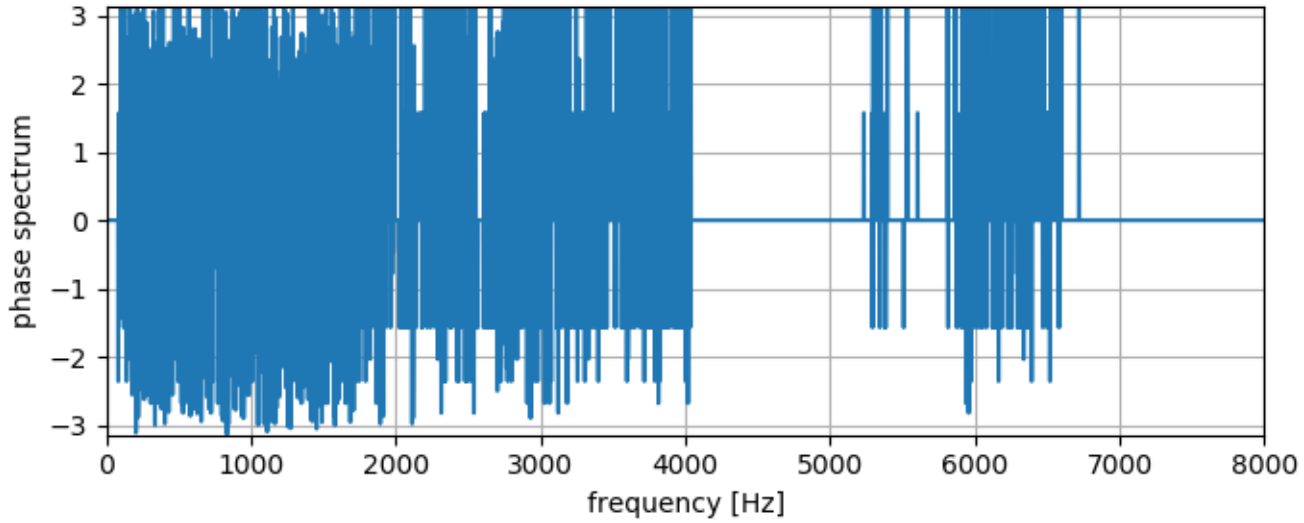
発音の差異を作り出す声道の変化



ref. https://commons.wikimedia.org/wiki/File:PlaceOfArticulation.svg, CC BY 3.0

音声の位相成分

- 見ても全然わかりません。。。
- でもフィルタリングには必要なので捨てないで！

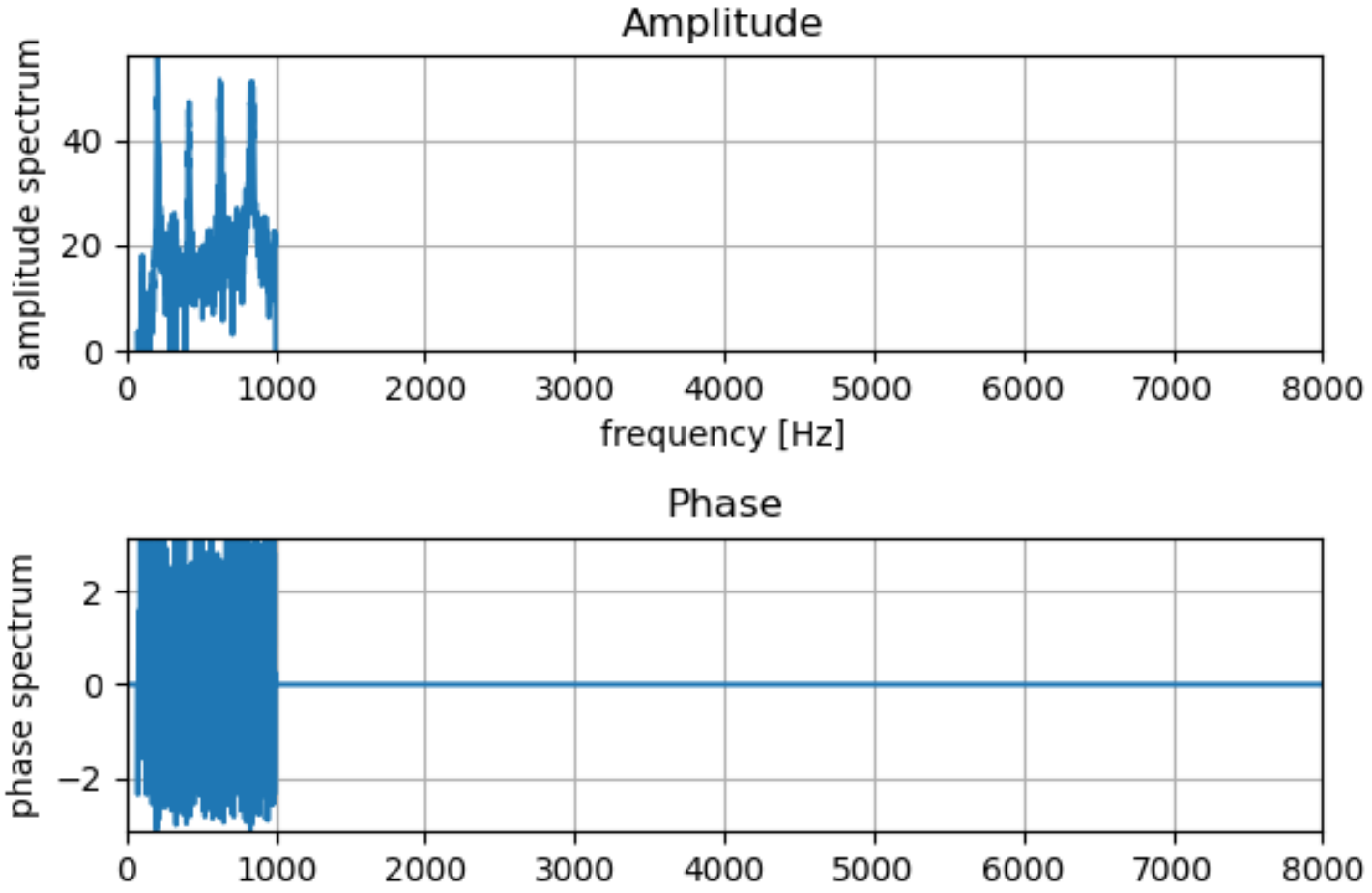


逆フーリエ変換

- フーリエ変換した値は逆フーリエ変換によって波形に戻すことができる
  - **このとき位相成分がないと元に戻りません！**
- 周波数空間で特定周波数成分を編集
  - サウンドエフェクトのイコライザとは、周波数ごとに大きくしたり小さくしたりする機能をイコライザと呼ぶ
  - 高周波を上げる→クリアな音に
  - 低周波を上げる→重厚な音に

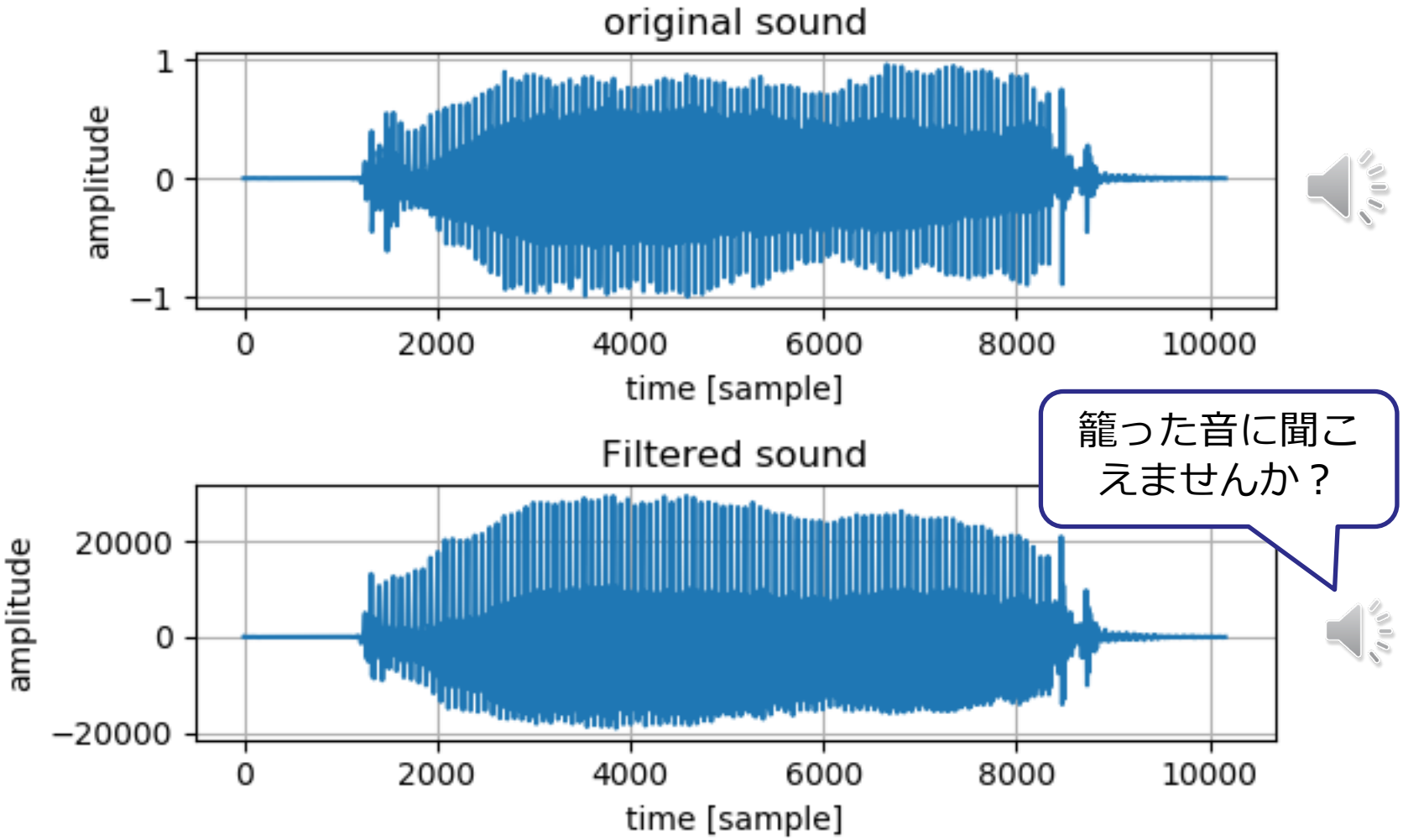
周波数フィルタリング

- 演習では1kHz以上の成分を除去



演習： SoundProcessing3.ipynb > **3. 周波数空間におけるフィルタリングと逆フーリエ変換**

逆フーリエ変換で波形に戻す



演習： SoundProcessing3.ipynb > **3. 周波数空間におけるフィルタリングと逆フーリエ変換**

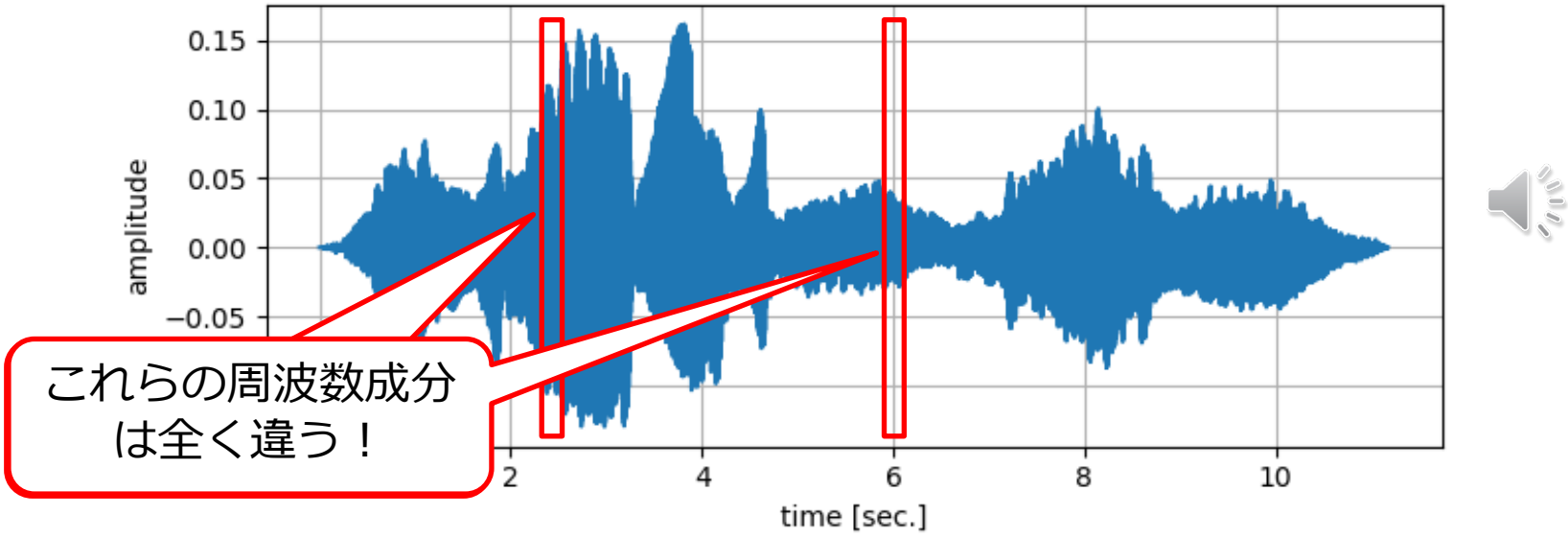
短時間フーリエ変換による スペクトログラム計算

演習：SoundProcessing4.ipynb

時間的に変化する音の周波数分解

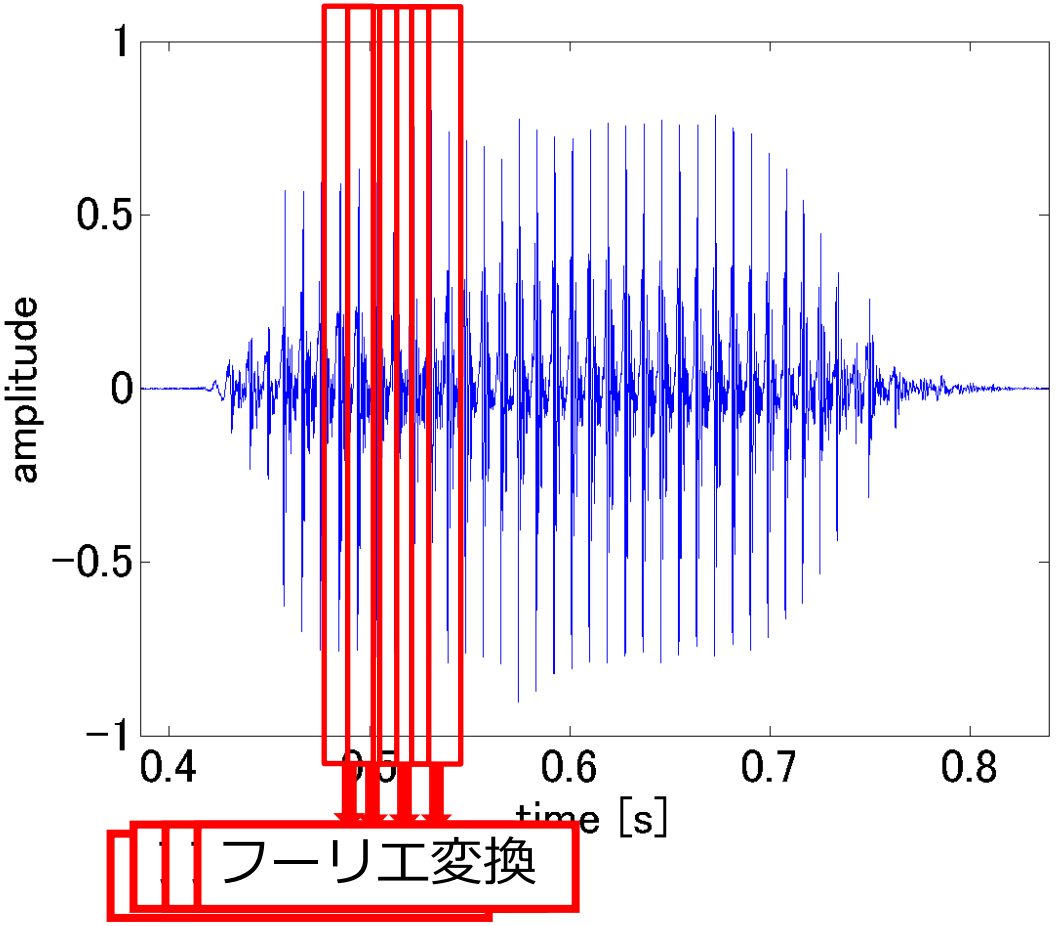
- 通常の音声や音楽は時間的に周波数成分が変化
- しかしフーリエ変換は
  - 与えられた信号に対し一気に周波数分解
  - ある瞬間の音（つまり1サンプル）では周波数分解できない

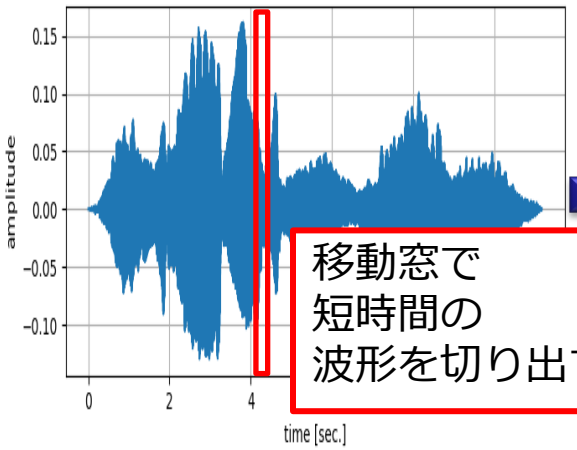
時間的に変化する音はどのように周波数分解したらいいか？



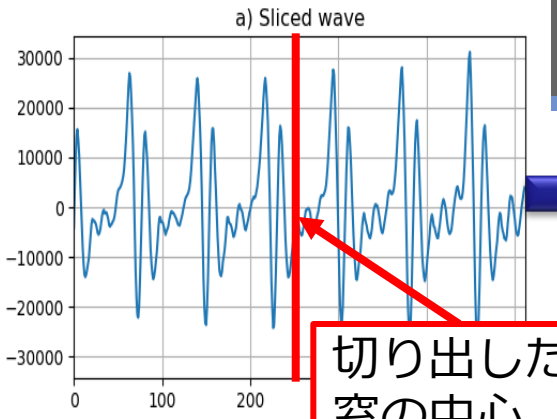
移動窓による短時間音声の切り出し

- 波形を短い区間で切り出して周波数分解
- 周波数成分を時間方向に並べればいい！

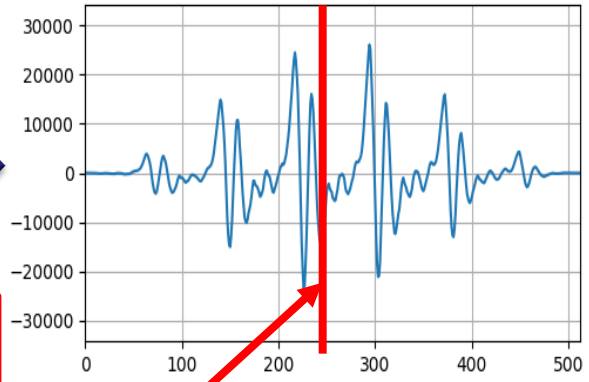




移動窓で
短時間の
波形を切り出す

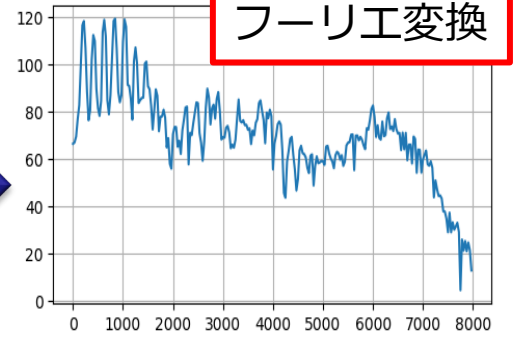


切り出した
窓の中心



その瞬間が一番重要で
その前後はあまり重要でないので
中心が大きくて、中心が離れるほど
小さくなるような重み（窓関数）をかける

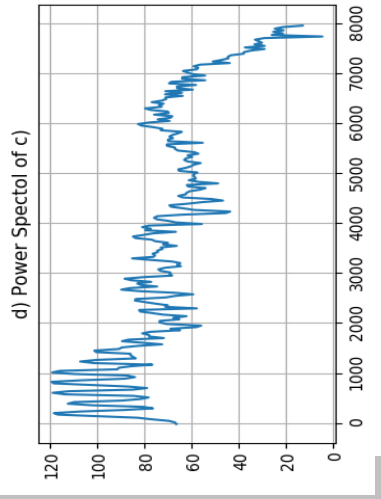
フーリエ変換

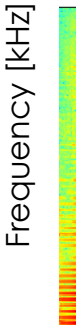


振幅の大きさを
色にして1つの帯にする

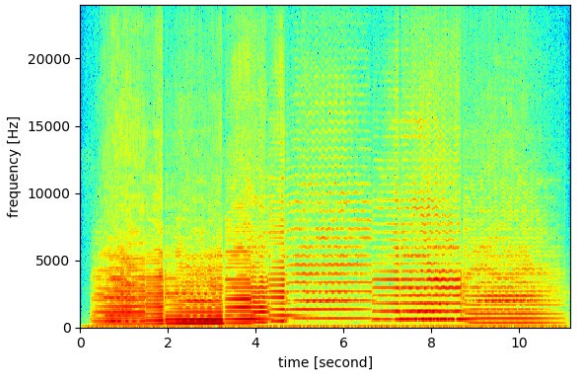
帯を時間方向に
並べる

90度回転

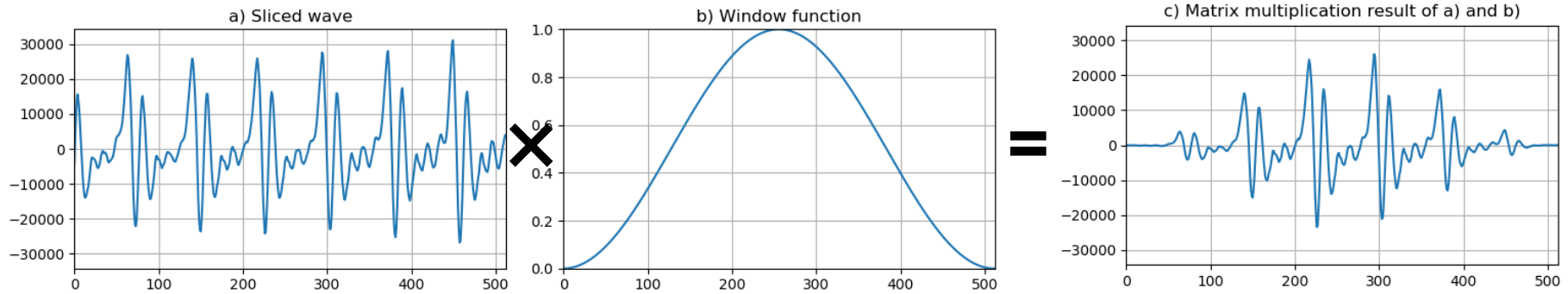




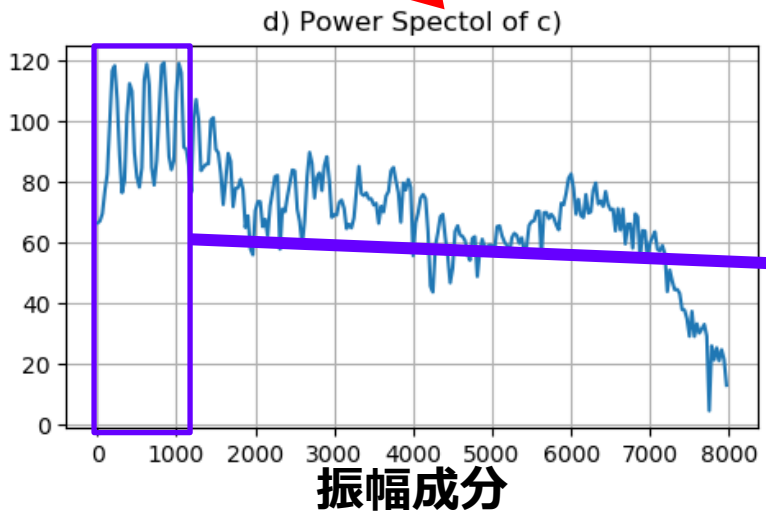
その瞬間の
周波数スペクトル

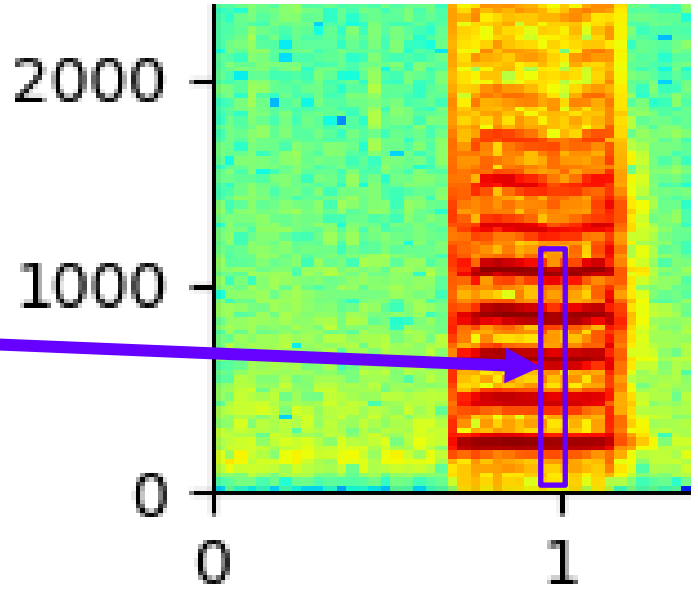


SoundProcessing4.ipynb：演習の解説



フーリエ変換

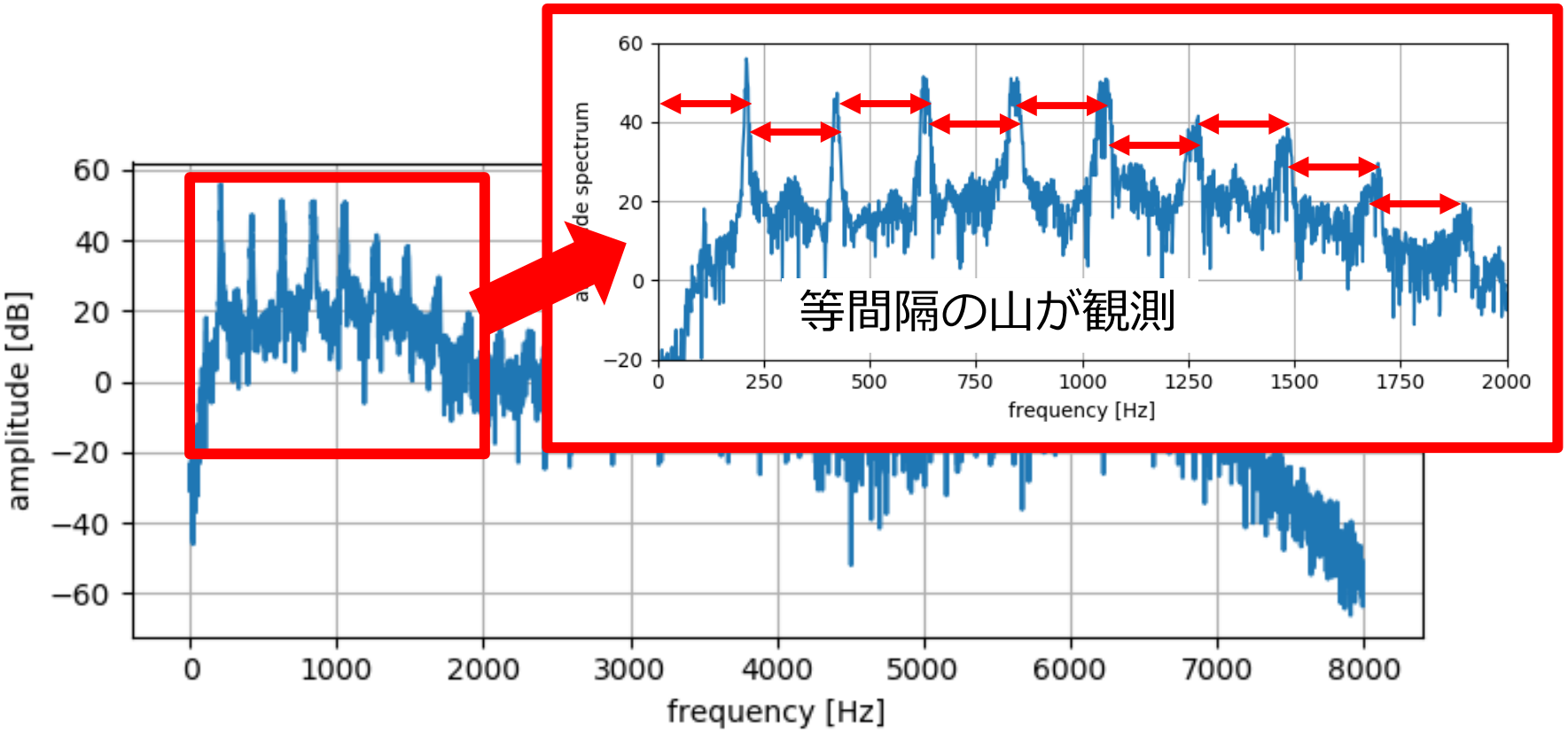




おまけ
スペクトル包絡

母音「あいうえお」を判別するには？

- 山谷にはあまり情報がない
- 山が全体としてどのような形で並んでいるかが重要！

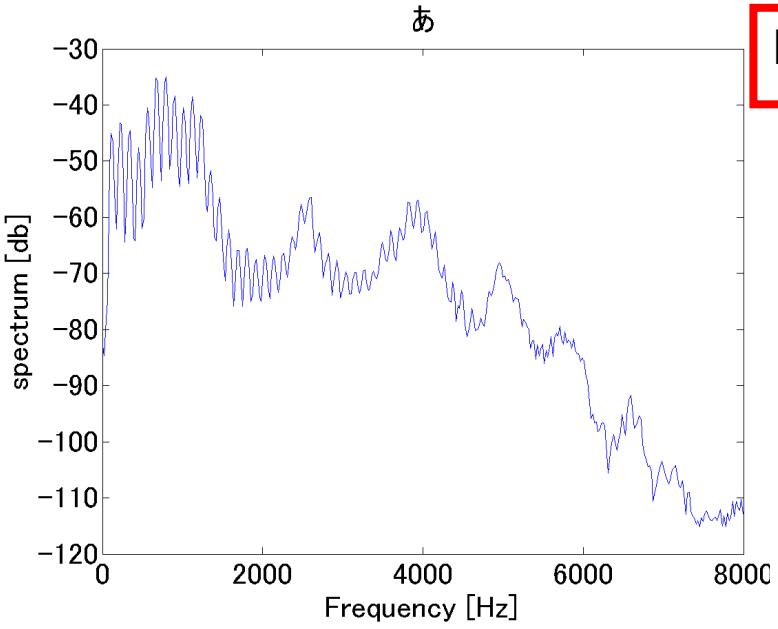


演習： SoundProcessing3.ipynb > **2. 音声の周波数成分**

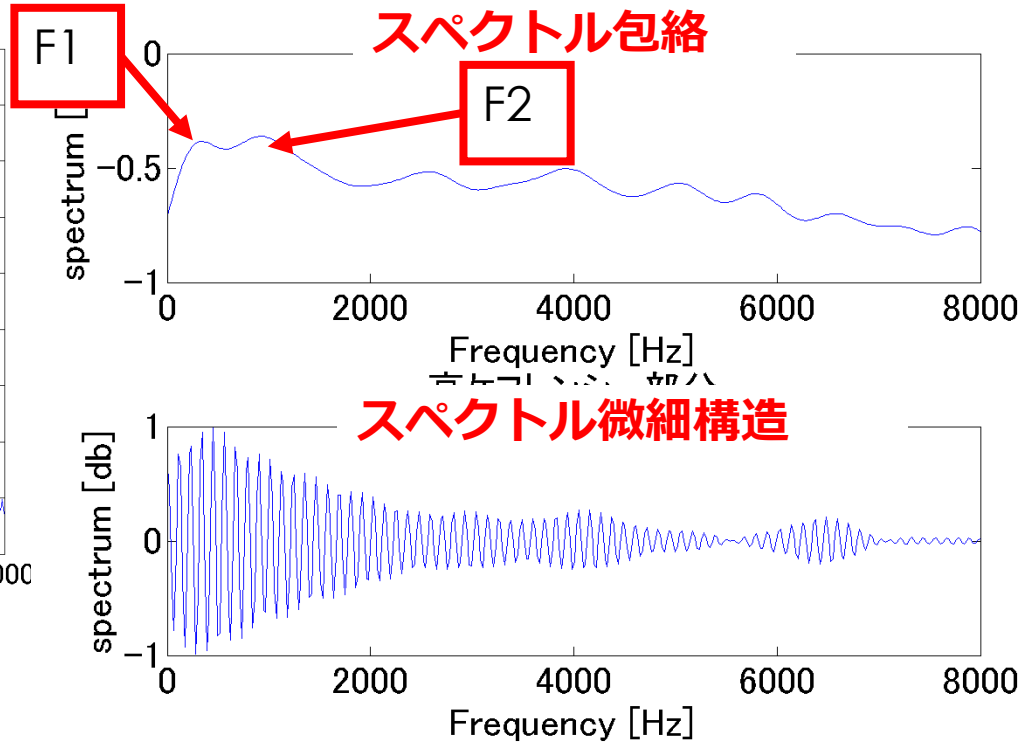
スペクトル包絡

- 周波数スペクトルをなだらかな山の成分（**スペクトル包絡**）と細かい山の成分（**スペクトル微細構造**）に分離
- 母音はスペクトル包絡を見れば判別できる！

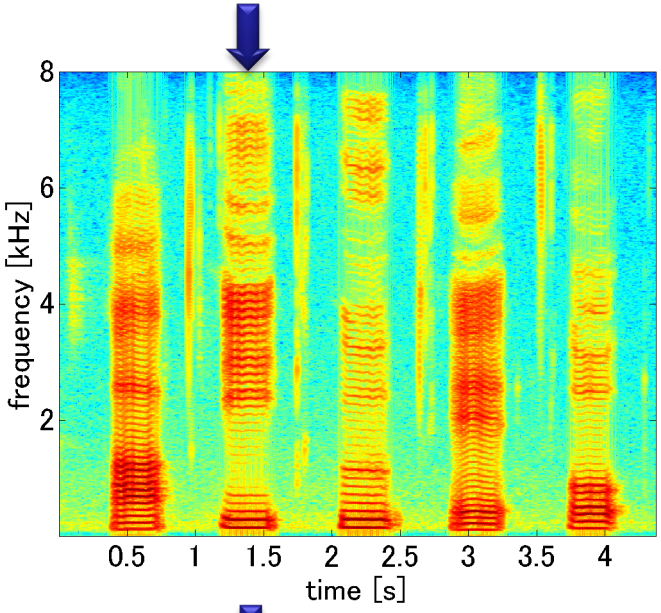
  スペクトル包絡の山の頂点を周波数の低い順に第１フォルマント(F1)、第2フォルマント(F2)と呼ぶ

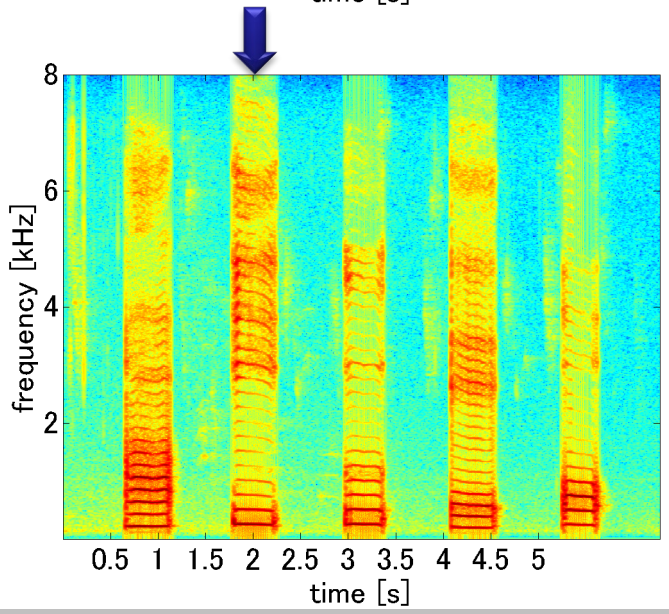


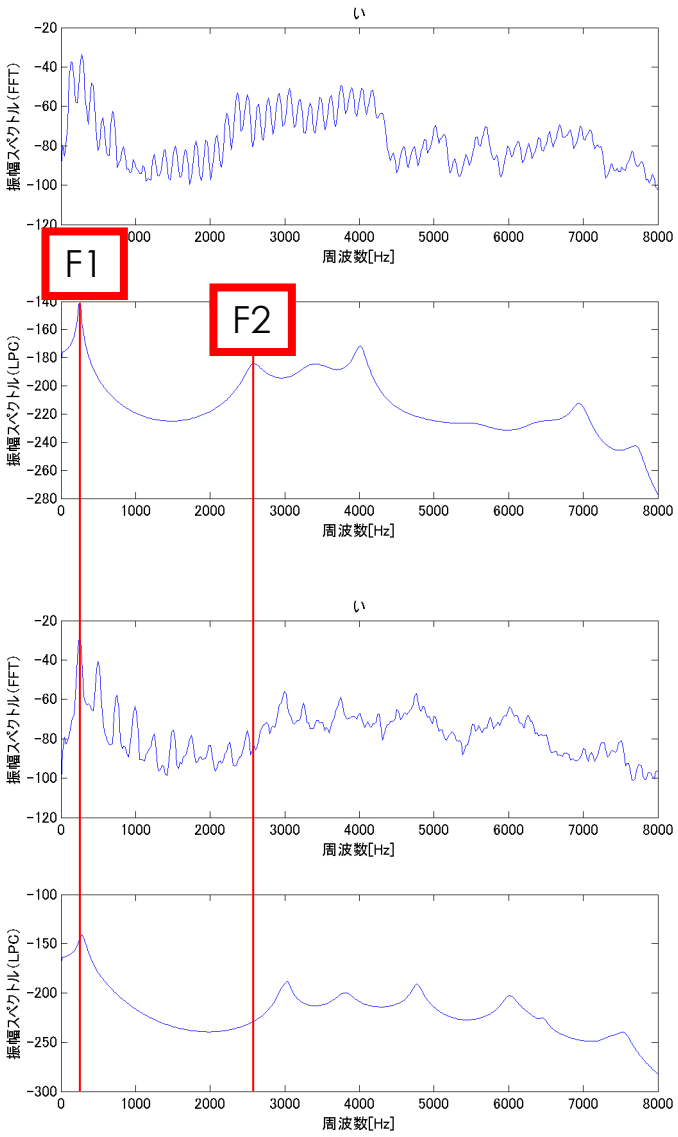
FFTによる周波数スペクトル



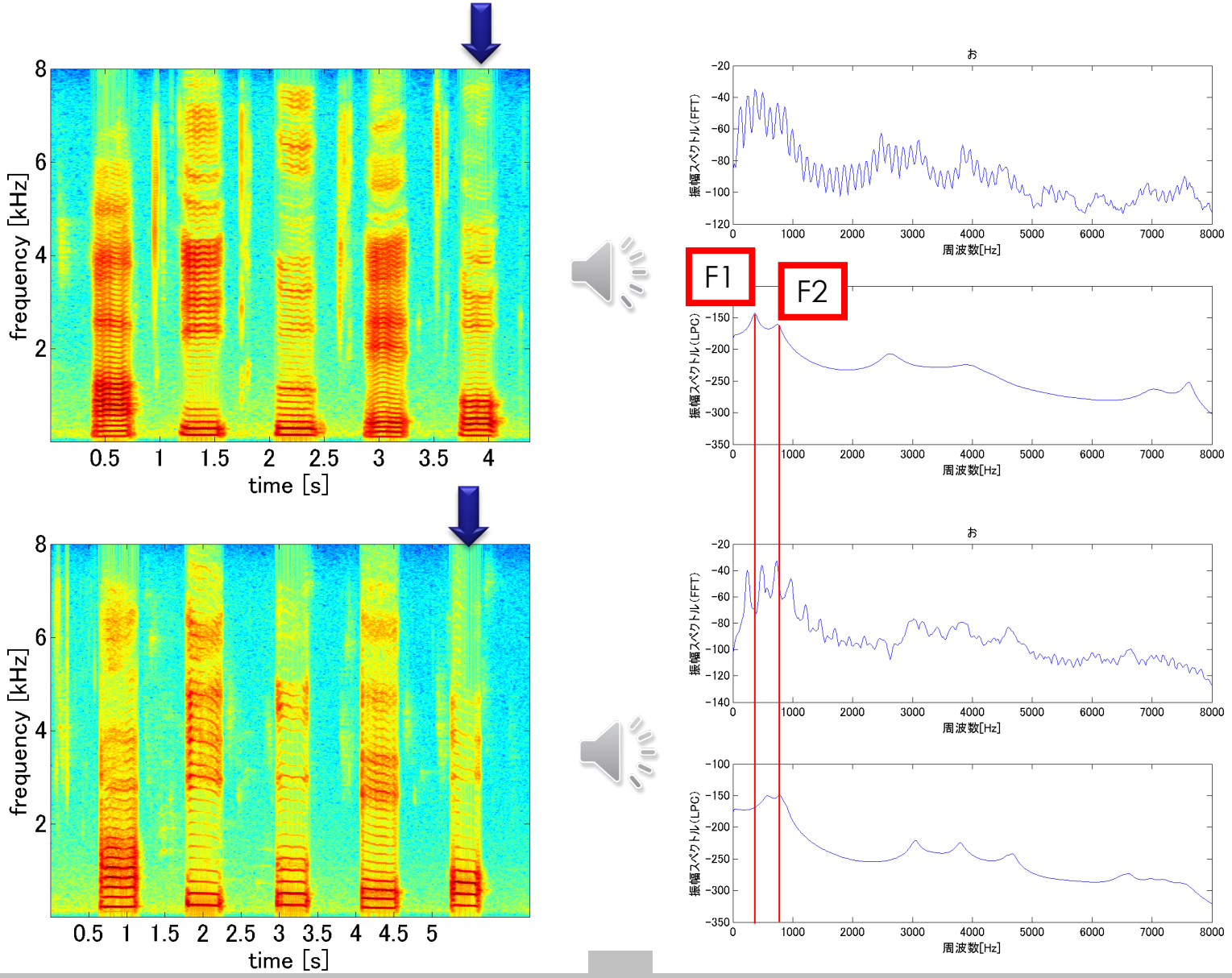
スペクトル包絡 “い” 男性 vs 女性



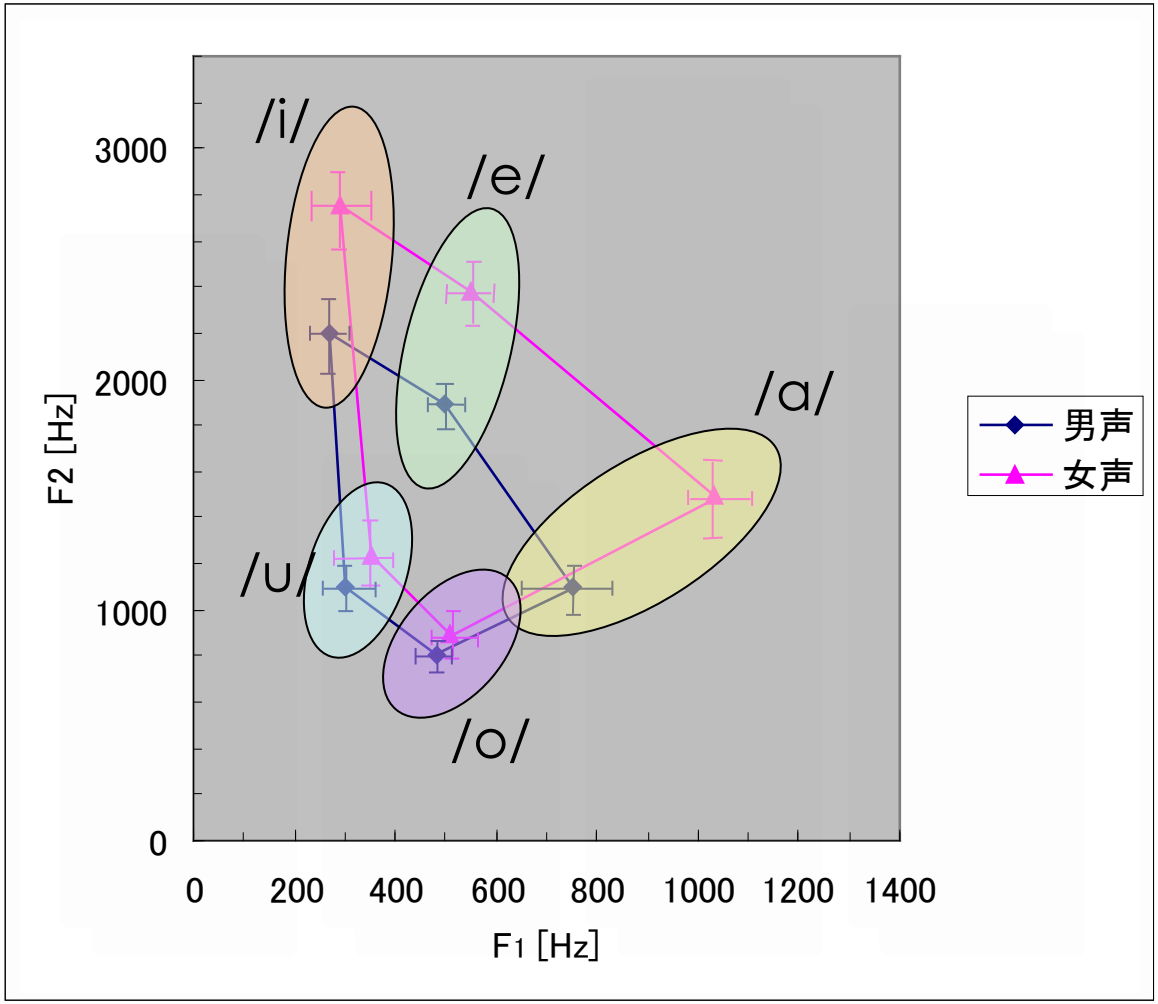




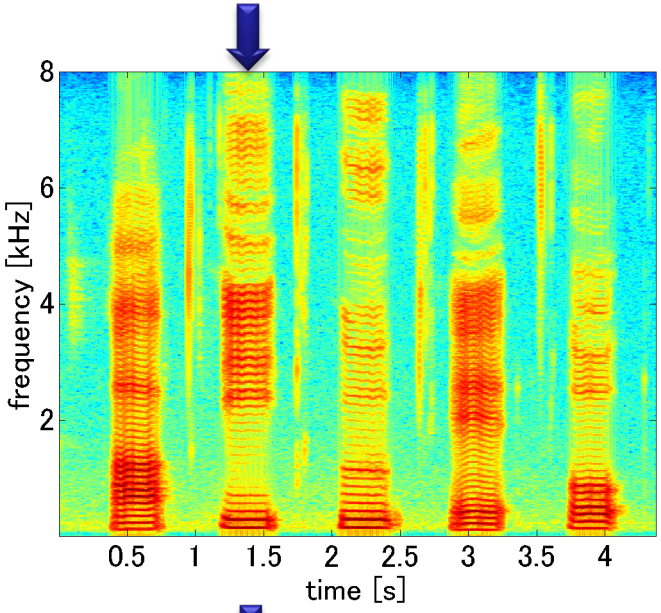
スペクトル包絡 “お” 男性 vs 女性

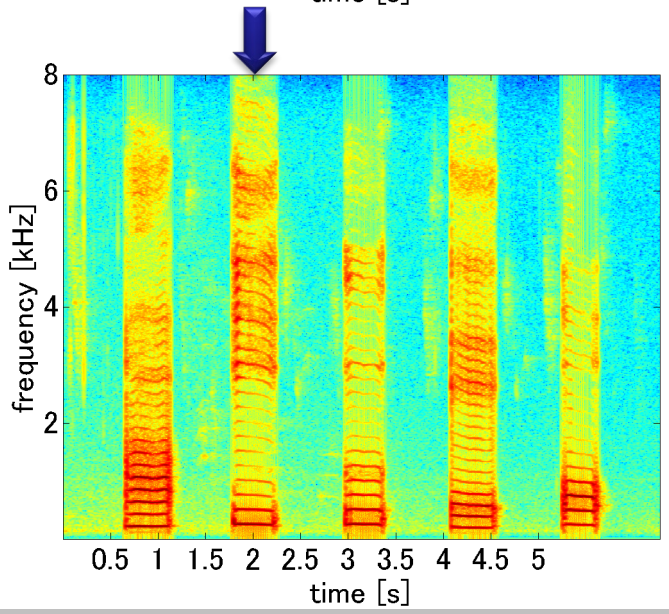


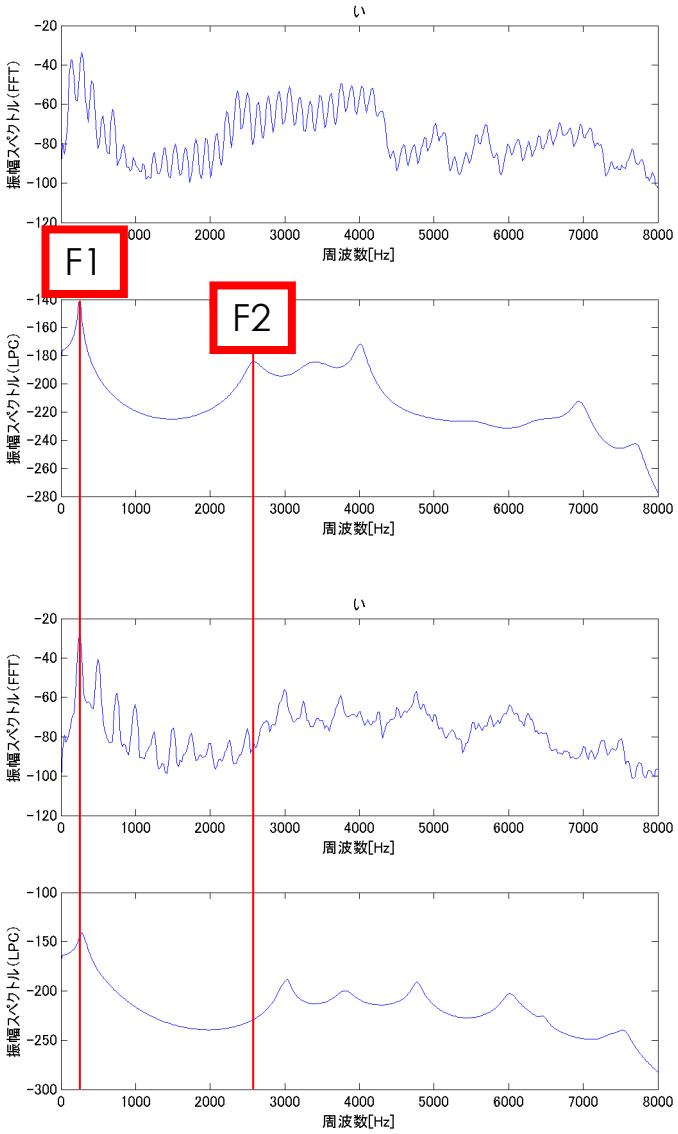
日本語母音の第1・第2フォルマント周波数の分布



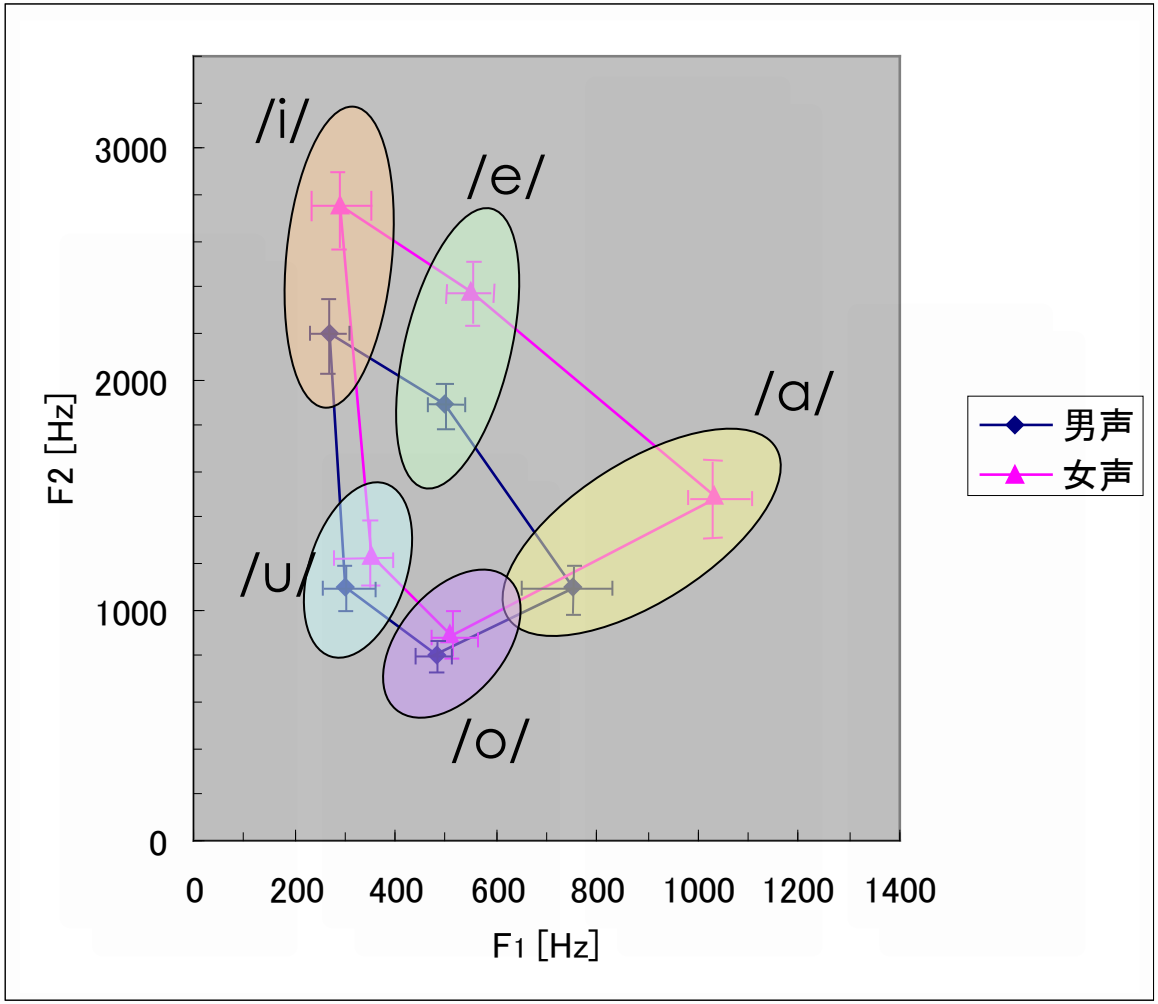
スペクトル包絡 “い” 男性 vs 女性



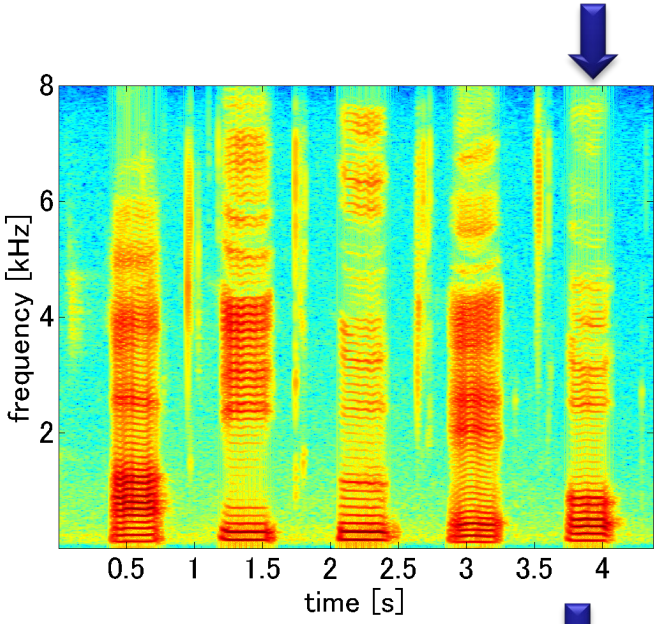


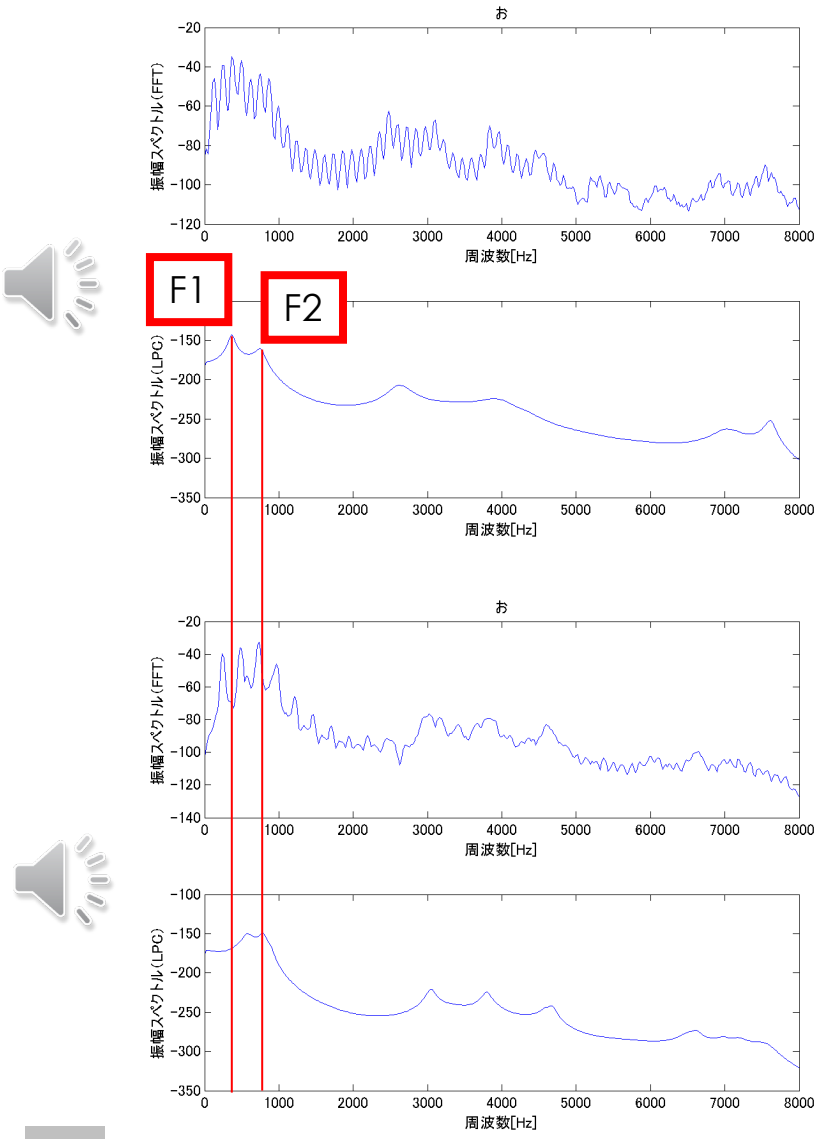


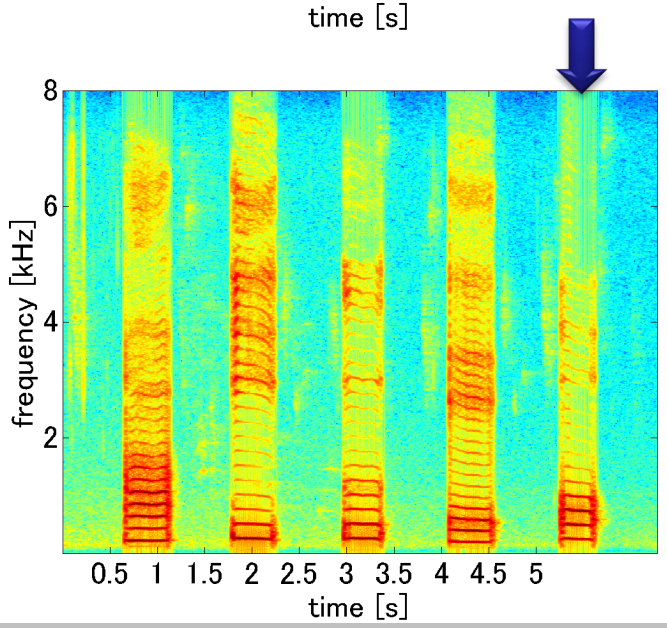
日本語母音の第1・第2フォルマント周波数の分布



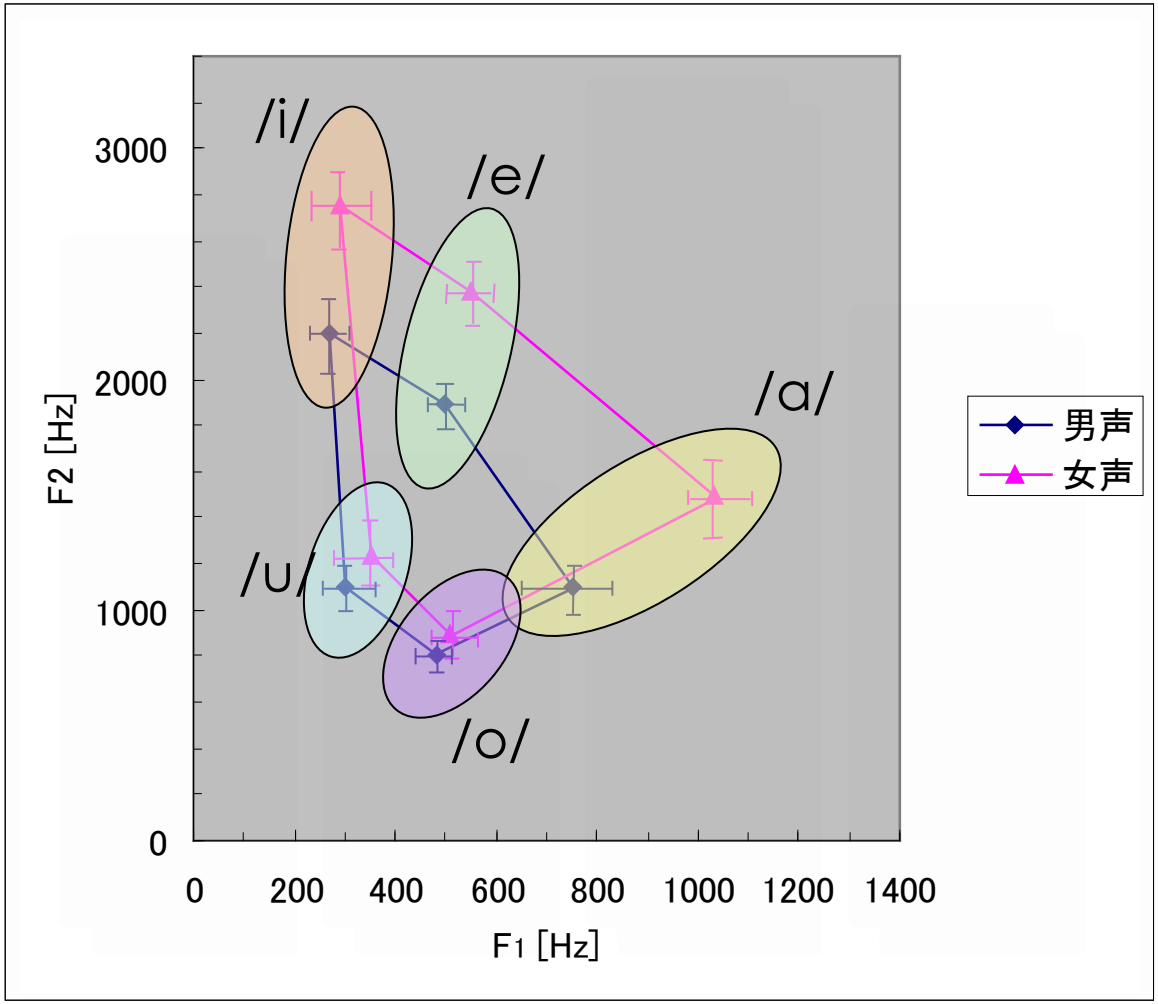
スペクトル包絡 “お” 男性 vs 女性



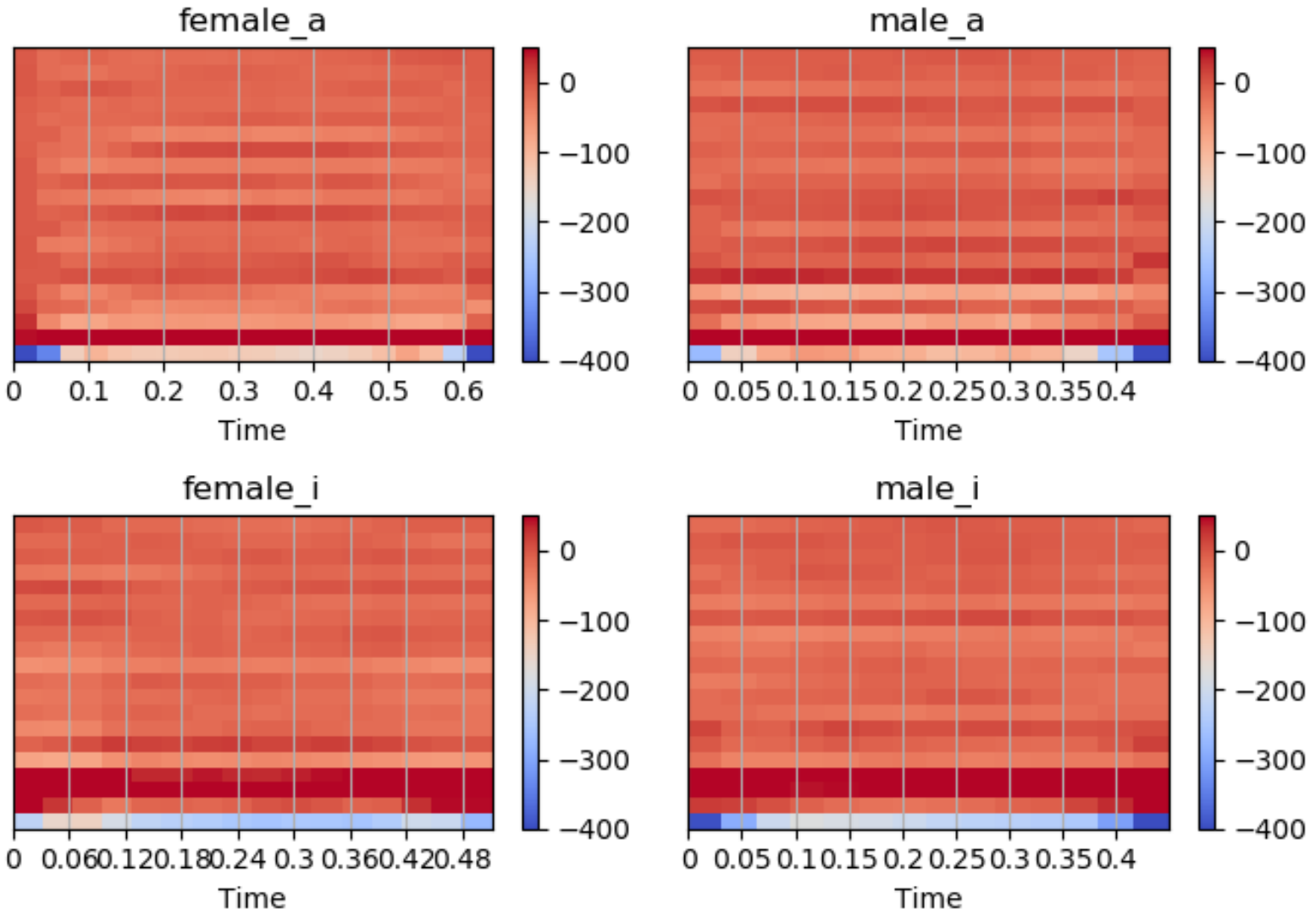




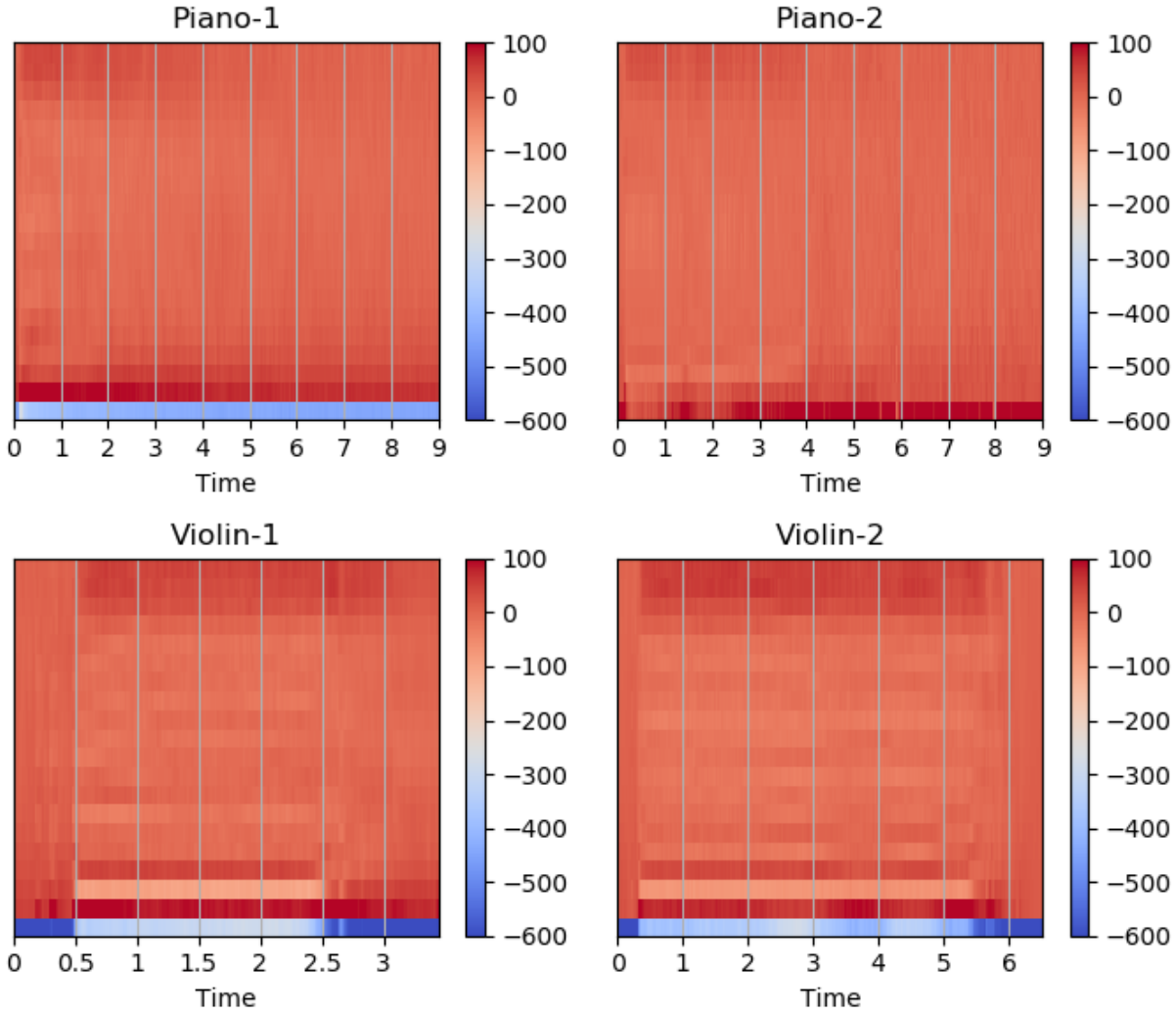
日本語母音の第1・第2フォルマント周波数の分布



音声のMFCC (男性音声・女性音声の「あ」と「い」)



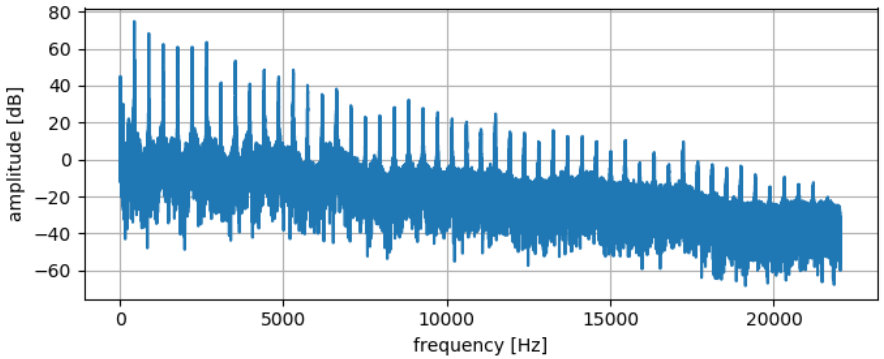
楽器の違い（A4=440Hz・Violin/Piano）



音の情報処理 課題：楽器の音を分析しよう

課題１：振幅スペクトルの描画

sound/ex-Vaiolin.wavを開いて、その振幅スペクトルを描画してください。



課題２：スペクトログラムの描画

ヴァイオリンの音楽sound/ex-Violin-music.wavのスペクトログラムを描画してください（**パラメータに注意**！）