直感重視の音声分析入門

@cuttlefish_math

2021 年 9 月 15 日

基幹理工学部 応用数理学科 2 年

アウトライン

- 自己紹介
- はじめに
- 準備編
  - 標本化
  - ベクトルの内積
- 音声分析（理論編）
  - 離散時間フーリエ変換
  - 短時間フーリエ変換
- 音声分析（実践編）
  - スペクトル包絡
- 補遺
  - 離散フーリエ変換
  - フーリエ変換

自己紹介

自己紹介

- @cuttlefish_math
- 基幹理工学部応用数理学科２年
- 応用数学（音声信号処理，数値計算など）が好き
- Web デザインたまにやってる
- よく同人イベント（コミケ，M3 など）に行っています

はじめに

はじめに

この発表では，音声分析の方法を「直感的に」解説する．ここでいう「音声分析」とは「音声から（使いやすい）パラメータを抽出すること」とする．

音声分析は，今日さまざまな分野において活用されている．たとえば，今日紹介する「スペクトル包絡」というパラメータは，携帯電話で重要な役割を演ずる[1]．また，カラオケの採点システムでは，音の高さを「基本周波数」というパラメータを推定することで算出していると考えられる[2]．

---

[1] このことについては [1] を参照．

[2] 厳密には，音の高さと基本周波数は完全には対応しない [6]．しかし，とても強い関係があるので，以降はあまり区別せず扱う．

はじめに

音声分析の方法はいろいろとあるが，それらは大きく 2 つに分けられる．

- 音声が数理モデルにしたがうと仮定し，モデルのパラメータを推定する
- 音声に数理モデルを仮定せず，普遍的な手法でパラメータを作成する

今回扱うのは主に後者の手法である（最後に示す「線形予測分析」だけは前者）．

準備編

準備編：標本化

コンピュータで音声信号を扱うには，まず標本化という操作が必要になる．標本化とは，絶え間なく流れてくる音声信号を時間が Δt だけ経つごとに記録し続けて，連続信号を離散信号に変換する操作である．
要するに，連続信号 $x(t)$ から，離散信号 $x[n] = x(n\Delta t)$ を得る操作が標本化である．

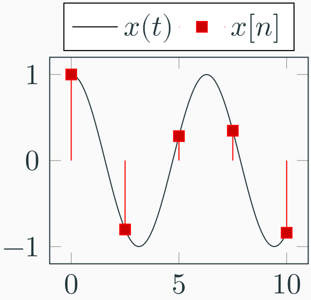

図 1: 標本化の様子

実ベクトル $\vec{x} = (x_0, x_1, x_2)$ と $\vec{y} = (y_0, y_1, y_2)$ の内積（ドット積）を

$$\vec{x} \cdot \vec{y} = x_0 y_0 + x_1 y_1 + x_2 y_2$$

と定義する．このとき，2 つのベクトル $\vec{x}$ と $\vec{y}$ がなす角を θ とすると

$$|\vec{x} \cdot \vec{y}| = ||\vec{x}||\vec{y}| \cos \theta| \leq |\vec{x}||\vec{y}|$$

が成立する．

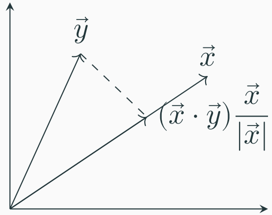

図 2: 内積となす角の関係

ドット積を拡張して，$\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ の標準内積を

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T} \overline{\mathbf{y}} = x_0 \overline{y_0} + \cdots + x_{N-1} \overline{y_{N-1}}$$

と定義する[3]．$y_0, \ldots, y_{N-1}$ だけ共役を取っているのは一見不自然かもしれないが，こうすると $\mathbf{y} = \mathbf{x}$ のとき

$$\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{n=0}^{N-1} x_n \overline{x_n}} = \sqrt{\sum_{n=0}^{N-1} |x_n|^2}$$

という，3 次元実ベクトルで成り立っていた関係が保たれる．左辺の量を $\mathbf{x}$ のノルム（長さ）といい，$\|\mathbf{x}\|$ と表す．

[3] N 個の複素数の組の全体集合を $\mathbb{C}^N$ と書く．

準備編：ベクトルの内積

ドット積と同様，$\mathbf{a}, \mathbf{b} \in \mathbb{C}^N$ なら

$$|\langle \mathbf{a}, \mathbf{b} \rangle| \leq \|\mathbf{a}\| \|\mathbf{b}\|$$

が成立する（シュワルツの不等式）.

等号が成り立つのは「$\mathbf{a} = t\mathbf{b}$ を満たす $t \in \mathbb{C}$ が存在する」とき，言い換えると「$\mathbf{a}$ と $\mathbf{b}$ が平行なとき」である.

よって，$\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{C}^N$ が $k = \|\mathbf{b}_1\| = \|\mathbf{b}_2\|$ を満たせば

$$|\langle \mathbf{a}, \mathbf{b}_i \rangle| \leq \|\mathbf{a}\|\|\mathbf{b}_i\| = k\|\mathbf{a}\| \quad (i = 1, 2)$$

である．等号が成り立つのは，やはり $\mathbf{a} \mathbin{/\!/} \mathbf{b}_i$ のときである．

ということは，もし $|\langle \mathbf{a}, \mathbf{b}_1 \rangle| < |\langle \mathbf{a}, \mathbf{b}_2 \rangle|$ なら，$\mathbf{b}_1$ よりも $\mathbf{b}_2$ のほうが「等号成立に近い」，言い換えると「$\mathbf{a}$ と平行に近い」と考えられる．つまり，ノルムが等しい複数のベクトル間では，内積は $\mathbf{a}$ とどれだけ平行に近いかを表す指標になる．

音声分析（理論編）

音声分析（理論編）：離散時間フーリエ変換

一般に，関数を分析するには「扱いやすい関数で（近似）表現する」のが有効である．この一番有名な例は，おそらくテイラー展開

$$f(x) = \frac{f^{(0)}(c)}{0!}(x-c)^0 + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n + \cdots$$

であろう．

音声信号を分析するにあたっては，周期関数で音声信号を表現できると良い．これは，音声の周期は概ね音の高さと対応するためである[4]．そこで，ここでは「周期関数を用いて音声信号を表現する」方法を考える．

[4]詳しくは 1 日目の発表「音と音の数学的関係」を参照．

周期関数で一番有名なのは，三角関数 $\cos(\omega t)$ と $\sin(\omega t)$ であろう．2 つを個々に考えてもよいのだが，オイラーの公式

$$e^{i\omega t} = \cos(\omega t) + i\sin(\omega t)$$

を使うとひとまとめにできて，数学的見通しが良くなる．

つまり，分析対称である離散信号 $x[n]$ と，なんらかの意味で「よく似た」$e^{i\Omega n}$ という離散信号を見つける手法があれば，とても好ましいと言える．

音声分析（理論編）：離散時間フーリエ変換

そのために，2 つのベクトル

$$\mathbf{x} = \begin{bmatrix} x[-N] & \cdots & x[N] \end{bmatrix}^\mathsf{T}$$
$$\mathbf{w}_\Omega = \begin{bmatrix} e^{i\Omega(-N)} & \cdots & e^{i\Omega N} \end{bmatrix}^\mathsf{T}$$

について，標準内積を計算する．Ω の値によらず $\|\mathbf{w}_\Omega\| = \sqrt{2N+1}$ だから

$$X_N(\Omega) = \langle \mathbf{x}, \mathbf{w}_\Omega \rangle$$

の絶対値が大きいほど，$\mathbf{x}$ は $\mathbf{w}_\Omega$ と平行に近い．

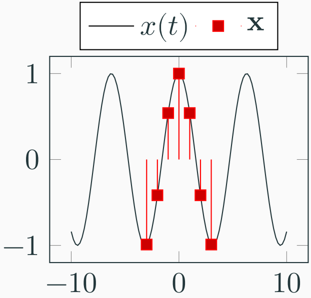

図 3: $x(t)$ と $\mathbf{x}$（$N = 3$）

したがって，$X_N(\Omega)$ は $\mathbf{x}$ と $\mathbf{w}_\Omega$ の近さを測る 1 つの指標になる．
実際に $X_N(\Omega)$ を計算すると

$$X_N(\Omega) = \langle \mathbf{x}, \mathbf{w}_\Omega \rangle = \sum_{n=-N}^{N} x[n]e^{-i\Omega n}$$

となる．すべての時刻における $x[n]$ の値を計算に含めるには，
極限

$$X(\Omega) = \lim_{N\to\infty} X_N(\Omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\Omega n}$$

をとればよい．$X(\Omega)$ を $x[n]$ の離散時間フーリエ変換という．

実は，$X(\Omega)$ から $x[n]$ は

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\Omega) e^{i\Omega n} \, \mathrm{d}\Omega \tag{1}$$

で復元できる．式 (1) を

$$x[n] = \lim_{\Delta\Omega \to +0} \left[\frac{1}{2\pi} \sum_{\substack{k=0,\pm1,\pm2,\ldots \\ -\pi \le k\Delta\Omega \le \pi}} X(k\Delta\Omega) e^{i(k\Delta\Omega)n} \Delta\Omega \right]$$

と書くと，$|X(\Omega)|$ は $x[n]$ を $e^{i\Omega n}$ に関する和（の極限）で書いたときの，各 $e^{i\Omega n}$ の重みを表していると分かる．

離散時間フーリエ変換

$$X(\Omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\Omega n}$$

は，確かに $x[n]$ と $e^{i\Omega n}$ との近さを測る良いツールである．

しかし，現在と1時間前の信号の値をまったく同じ重みで評価し，近さを測ることは，はたして妥当なのだろうか？　　たいてい，音声信号を分析するときは，もっと局所的な様子を調べたいはずだ．

音声分析（理論編）：短時間フーリエ変換

ある時刻 k 周辺における近さを求めるため，次のような形をした「重み」$w[n-k]$ を $x[n]$ に掛けてやる．

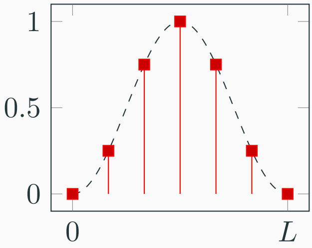

図 4: $w[n]$ のグラフ

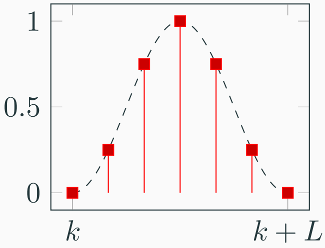

図 5: $w[n-k]$ のグラフ

こうすることで，n が k に近いとき，$x[n]$ の値が $X(\Omega)$ の値に強く影響するようにできる．

さらに，$n < k$ or $k + L \leq n$ のとき $w[n-k] = 0$ なら

$$X(\Omega) = \sum_{n=-\infty}^{\infty} w[n-k]x[n]e^{-i\Omega n} = \sum_{n=k}^{k+L-1} w[n-k]x[n]e^{-i\Omega n}$$

となるから，$\sum$ は無限和ではなくなり，$X(\Omega)$ を計算できるようにもなる．

ただし，$X(\Omega)$ は注目する時刻 k の関数にもなっている．そこで

$$\mathrm{STFT}\, x(k,\Omega) = \sum_{n=-\infty}^{\infty} w[n-k]x[n]e^{-i\Omega n}$$

と，あらためて置きなおす．$\mathrm{STFT}\, x(k,\Omega)$ を $x[n]$ の短時間フーリエ変換という[5]．また，$w[n]$ を窓関数という．

ここまでの議論によれば，$|\mathrm{STFT}\, x(k,\Omega)|$ は時刻 k 周辺において信号 $x[n]$ が $e^{i\Omega n}$ にどれだけ近いかを表している．

[5] 短時間フーリエ変換の定義にはいくつかの流儀がある．詳しくは [5, 7] を参照．

音声分析（実践編）

音声分析（実践編）: スペクトル包絡

いよいよお待ちかね（？），実際に音声を短時間フーリエ変換で分析してみよう．

簡単のため，今回は分析する音声を母音（あ・い・う・え・お）に限る．

音声分析（実践編）: スペクトル包絡

以下に，MATLAB® で短時間フーリエ変換を計算・図示するコード の例を示す（2 行目以降の改行は行が溢れたためのもの）.

```
[y,fs] = audioread("ファイルのパス")
stft(y,fs,"Window",hanning(2048),"OverlapLength
  ",1024,"FFTLength",2048,"FrequencyRange","onesided
  ")
```

なお，以下では時刻 k と角周波数 Ω を，標本化する前の時刻と角周波数に換算して表示する[6]．

また，短時間フーリエ変換の絶対値は多くの場合，デシベル $x\,\mathrm{dB} = 20\log_{10} x$ を用いて表示する（「片対数グラフで表示する」とも言える）．

[6]標本化後の角周波数は正規化角周波数と呼ばれる．詳しくは [4] を参照．

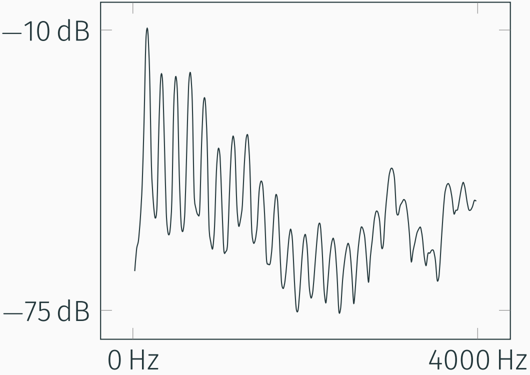

図 6: ある時刻 k における「あ」の $20\log_{10}|\mathrm{STFT}\,x(k,\Omega)|$ の様子

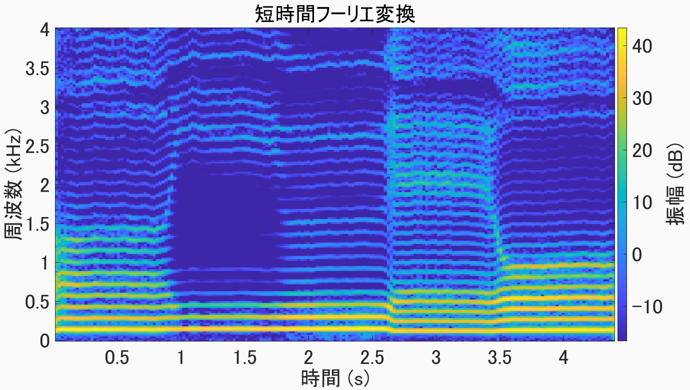

図 7:「あいうえお」の $20\log_{10}|\mathrm{STFT}\,x(k,\Omega)|$ の様子

音声分析（実践編）：スペクトル包絡

さきほどの図から，ある時刻 k における短時間フーリエ変換は「緩やかで非周期的な変動」と「激しく周期的な変動」を併せ持つことが分かる．前者のことをスペクトル包絡という．

後者は声の高さに起因する成分だから，前者は声から「声の高さ」の影響を取り除いたときに残る成分である．より大雑把に言うと，スペクトル包絡は声の「音色」に相当するパラメータである．

音声分析（実践編）：スペクトル包絡

それでは，スペクトル包絡はどうやって算出すればよいのだろう？　有名かつ古典的な手法として，次の 2 つが知られている．

- 線形予測分析
- ケプストラム法

大雑把に 2 つの概要を述べて，この発表を終わりとすることにしよう[7]．線形予測分析では，音声が p 次の自己回帰過程

$$x[n] = \sum_{i=1}^{p} \phi_i x[n-i] + \epsilon_n \quad (\epsilon_n \text{ はホワイトノイズ})$$

にしたがうと仮定し，$\phi_1, \ldots, \phi_p$ を推定することで，信号のスペクトル包絡を推定する．

一方，ケプストラム法では，fig. 6 のグラフを平滑化する（山と谷を均す）ことでスペクトル包絡を求める．

[7] [3, 2] により詳しい記述がある．

補遺

離散フーリエ変換
ここでは線形代数の応用として，離散フーリエ変換について紹介する（短時間フーリエ変換は「離散フーリエ変換を繰り返し行うこと」と解釈できる）.

Definition
有限長の複素数列 $\mathbf{x} = [x_0 \quad \cdots \quad x_{N-1}]^\mathsf{T}$ に対し，$\mathbf{x}$ の離散フーリエ変換 $\mathbf{X} = [X_0 \quad \cdots \quad X_{N-1}]^\mathsf{T} \in \mathbb{C}^N$ を

$$X_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k e^{-i(n\Delta\Omega)k} \quad (\Delta\Omega = 2\pi/N)$$

により定義する.

補遺：離散フーリエ変換

以下では，離散フーリエ変換の意味づけを与える．信号 $\{x_n\}$ が周期 N の周期数列なら，$\{x_n\}$ はベクトル $\mathbf{x} = \begin{bmatrix} x_0 & \cdots & x_{N-1} \end{bmatrix}^\mathsf{T}$ と 1 対 1 に対応する．つまり，$\{x_n\}$ について調べるには，$\mathbf{x}$ について調べれば十分である．

さて

$$\mathbf{w}_n = \frac{1}{\sqrt{N}} \begin{bmatrix} e^{i(n\Delta\Omega)0} & \cdots & e^{i(n\Delta\Omega)(N-1)} \end{bmatrix}^\mathsf{T} \quad (\Delta\Omega = 2\pi/N)$$

とおくと，$\mathcal{B} = \{\mathbf{w}_0, \dots, \mathbf{w}_{N-1}\}$ は $\mathbb{C}^N$ の正規直交基底になる．ということは，$\mathbf{x} = X_0\mathbf{w}_0 + \cdots + X_{N-1}\mathbf{w}_{N-1}$ を満たす $X_0, \dots, X_{N-1} \in \mathbb{C}$ は一意に定まる．

補遺：離散フーリエ変換

X_n を求めよう．$\mathscr{B}$ は正規直交基底だから

$$
\begin{aligned}
\langle \mathbf{x}, \mathbf{w}_n \rangle &= X_0 \langle \mathbf{w}_0, \mathbf{w}_n \rangle + \cdots + X_{N-1} \langle \mathbf{w}_{N-1}, \mathbf{w}_n \rangle \\
&= \cancelto{0}{0X_0 + \ldots} + 1X_n \cancelto{0}{+ \cdots + 0X_{N-1}}
\end{aligned}
$$

である．$\langle \mathbf{x}, \mathbf{w}_n \rangle$ を $\sum$ によって書けば，次のようになる．

$$
X_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k e^{-i(n\Delta\Omega)k}
$$

これは離散フーリエ変換に他ならない．つまり，**離散フーリエ変換はベクトルの基底を取り換える操作**と捉えられる．

フーリエ変換
$\omega \in \mathbb{R}$ を固定する．離散信号 $e^{i\Omega n}$ は，連続信号 $e^{i\omega t}$ を標本化して得られたものだとする．

$x[n] = x(n\Delta t)$ という式を思い出すと，$e^{i\Omega n} = e^{i\omega n\Delta t}$ でなければならない．したがって，Ω と ω の間には $\Omega = \omega\Delta t$ という関係がある．

補遺：フーリエ変換

$$X(\omega\Delta t) = \sum_{n=-\infty}^{\infty} x(n\Delta t) e^{-i(\omega\Delta t)n}$$

であるから，Δt を掛けて極限をとれば次のようになる．

$$\begin{aligned}
&X(\omega\Delta t)\Delta t \\
&= \sum_{n=-\infty}^{\infty} x(n\Delta t) e^{-i\omega n\Delta t}\Delta t \\
&\to \underbrace{\int_{-\infty}^{\infty} x(t) e^{-i\omega t}\,\mathrm{d}t}_{\mathcal{F}\,x(\omega)\ \text{とおく}} \quad (\Delta t \to +0)
\end{aligned}$$

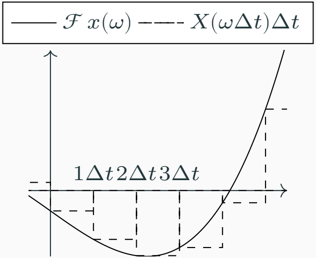

図 8: $X(\omega\Delta t)\Delta t \to \mathcal{F}\,x(\omega)$ $(\Delta t \to +0)$ のイメージ図

補遺：フーリエ変換

Definition (フーリエ変換)
関数 $x(t)$ に対し

$$\mathscr{F}\,x(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}\,\mathrm{d}t$$

で定義される関数 $\mathscr{F}\,x(\omega)$ を，$x(t)$ のフーリエ変換という．

各 ω について，$\mathscr{F}\,x(\omega)$ は $x(t)$ に含まれる $e^{i\omega t}$ という関数の「振幅」と「位相」を表している．

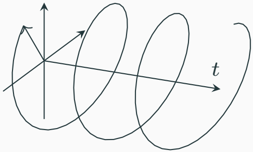

図 9: $Ae^{i(\omega t+\phi)}$ のグラフ（時間軸と直交する平面は複素平面）

参考文献

参考文献 i

[1] 守谷健弘, 鎌本優, 原田登, 杉浦亮介.
音声音響符号化技術の進展.
電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 10, No. 4, pp. 246–256, 2016.

[2] 森勢将雅.
音声分析合成.
コロナ社, 東京, 2018.

[3] 高道慎之介.
やさしく音声分析法を学ぶ：ケプストラム分析と lpc 分析.
`https://www.slideshare.net/ShinnosukeTakamichi/lpc-49065650`, 2015.
最終閲覧：2021 年 9 月 13 日.

[4] 鏡慎吾.
やる夫で学ぶディジタル信号処理.
http://www.ic.is.tohoku.ac.jp/~swk/lecture/yaruodsp/main.html, n.d.
最終閲覧：2021 年 9 月 13 日.

[5] 矢田部浩平, 升山義紀, 草野翼, 及川靖広.
位相変換による複素スペクトログラムの表現.
日本音響学会誌, Vol. 75, No. 3, pp. 147–155, 2019.

[6] 柏野牧夫.
音のなんでもコーナー q and a.
https://acoustics.jp/qanda/answer/101.html, n.d.
最終閲覧：2021 年 9 月 13 日.

[7] 小野順貴.
短時間フーリエ変換の基礎と応用.
日本音響学会誌, Vol. 72, No. 12, pp. 764–769, 2016.