奈良先端大 音情報処理論第2回 (2017/11/12)

音声の特徴抽出 (DFT, LPC, ケプストラム分析)

東京大学 情報理工学系研究科 特任助教
高道 慎之介

自己紹介

名前・所属

高道 慎之介 (たかみち しんのすけ)

東京大学 大学院情報理工学系研究科 特任助教

NAISTとの関わり

2011/04: 知能コミュニケーション研究室 (中村 哲教授) 1期生

2016/03: 博士課程修了

研究分野

電気音響・音像定位

音声信号処理

音声合成・変換

言語教育

本講義の目的

音声の特徴とは何か，それをどう定量化するか

デジタル信号処理の基礎

特徴抽出の前準備

音声とは

音声の生成過程，包絡成分，微細構造

音声の特徴抽出

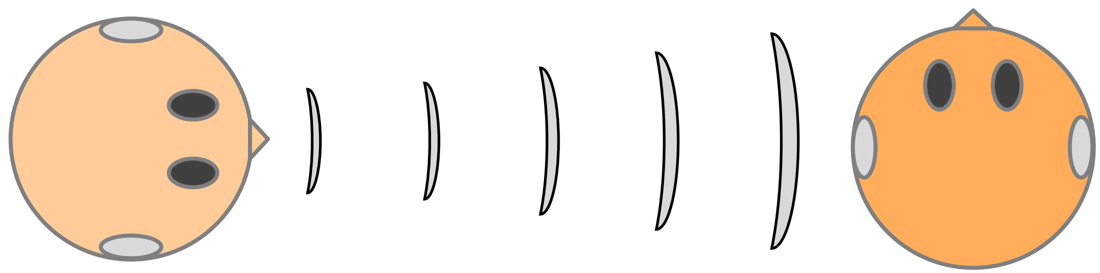
ケプストラム分析，LPC分析

デジタル信号処理の基礎

アナログ／デジタル変換による音声信号の取り込み

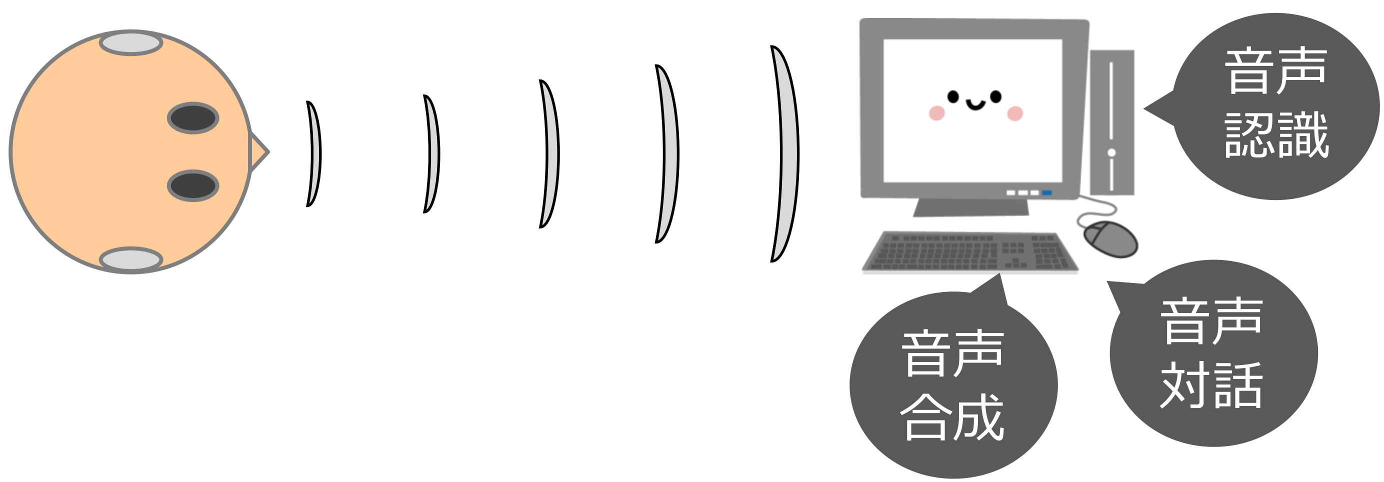
我々はどうやって音声コミュニケーションを行う？

口から発せられた原音声信号が，空気中を伝播して耳に到達



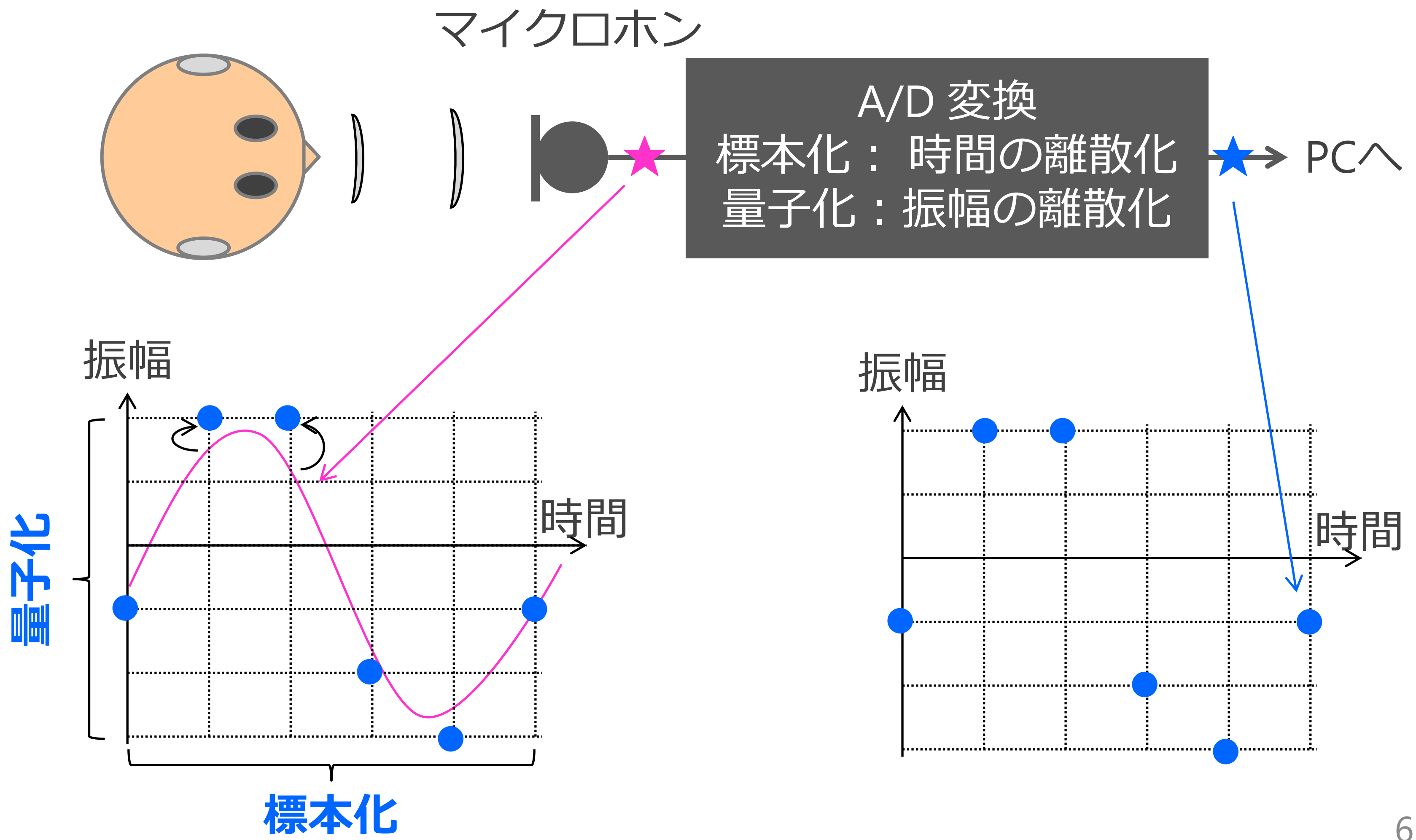
この一方をデジタル計算機に置き換えたら？

音声信号をデジタル信号に変えて処理 → **アナログ／デジタル変換**



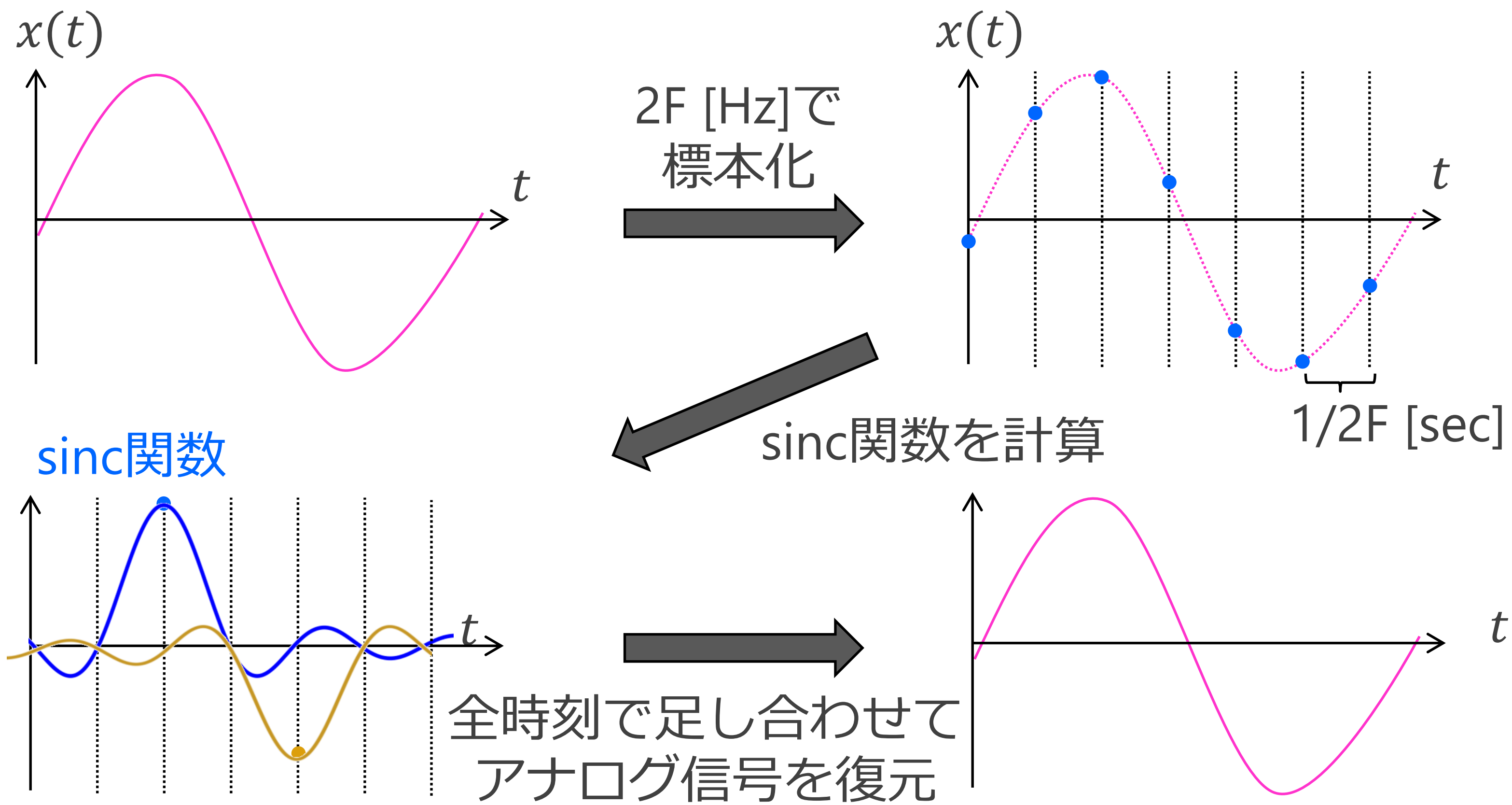
アナログ／デジタル変換（A/D変換）

原音声信号 (アナログ) を，計算機で扱えるデジタル信号へ



標本化定理 (sampling theorem)

原信号の最大周波数が F [Hz]であるとき，2F [Hz] 以上で標本化すれば，原信号を完全復元できる！



* sinc関数： デジタル→アナログ復元のための関数

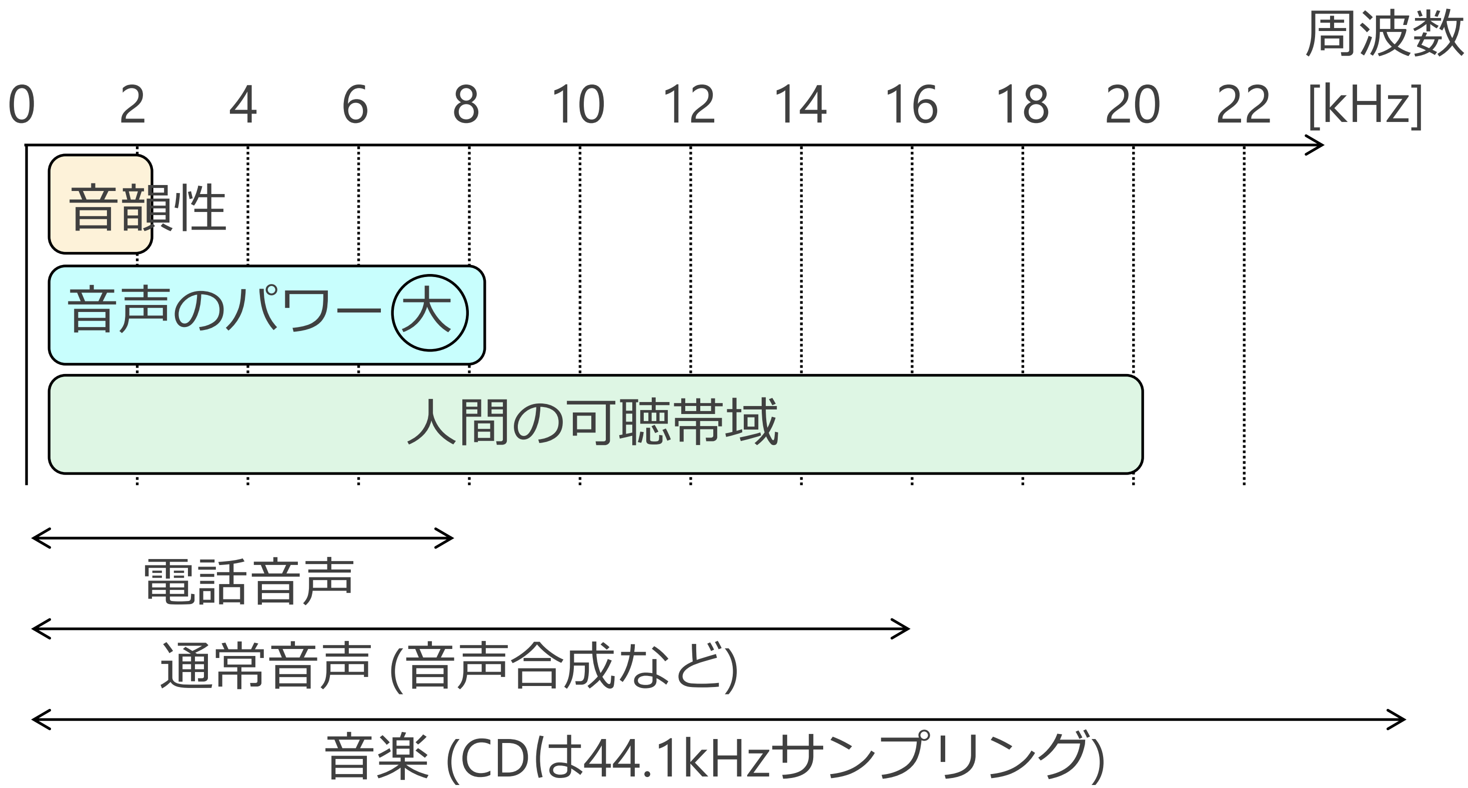
音声処理で用いられる標本化

必要な情報に応じて標本化周波数を変化

標本化周波数高→ 多くの情報を保存できるが，データサイズ大

必要な帯域の2倍以上の標本化周波数を使用

例えば…



離散フーリエ変換・z変換

A/D変換した後の音声特徴量抽出

- **離散フーリエ変換**： ケプストラム分析
- **z変換**： LPC分析

離散フーリエ変換 (Discrete Fourier Transform: DFT)

- デジタル信号を「時間とともに振動する波」の和で表現
- フーリエ変換の離散版

z変換 (z-transform)

- デジタル信号を「時間とともに増加・減衰しながら振動する波」の和で表現
- ラプラス変換の離散版

離散フーリエ変換・z変換

A/D変換した後の音声特徴量抽出

- **離散フーリエ変換**： ケプストラム分析
- **z変換**： LPC分析

離散フーリエ変換 (Discrete Fourier Transform: DFT)

- デジタル信号を「時間とともに振動する波」の和で表現
- フーリエ変換の離散版

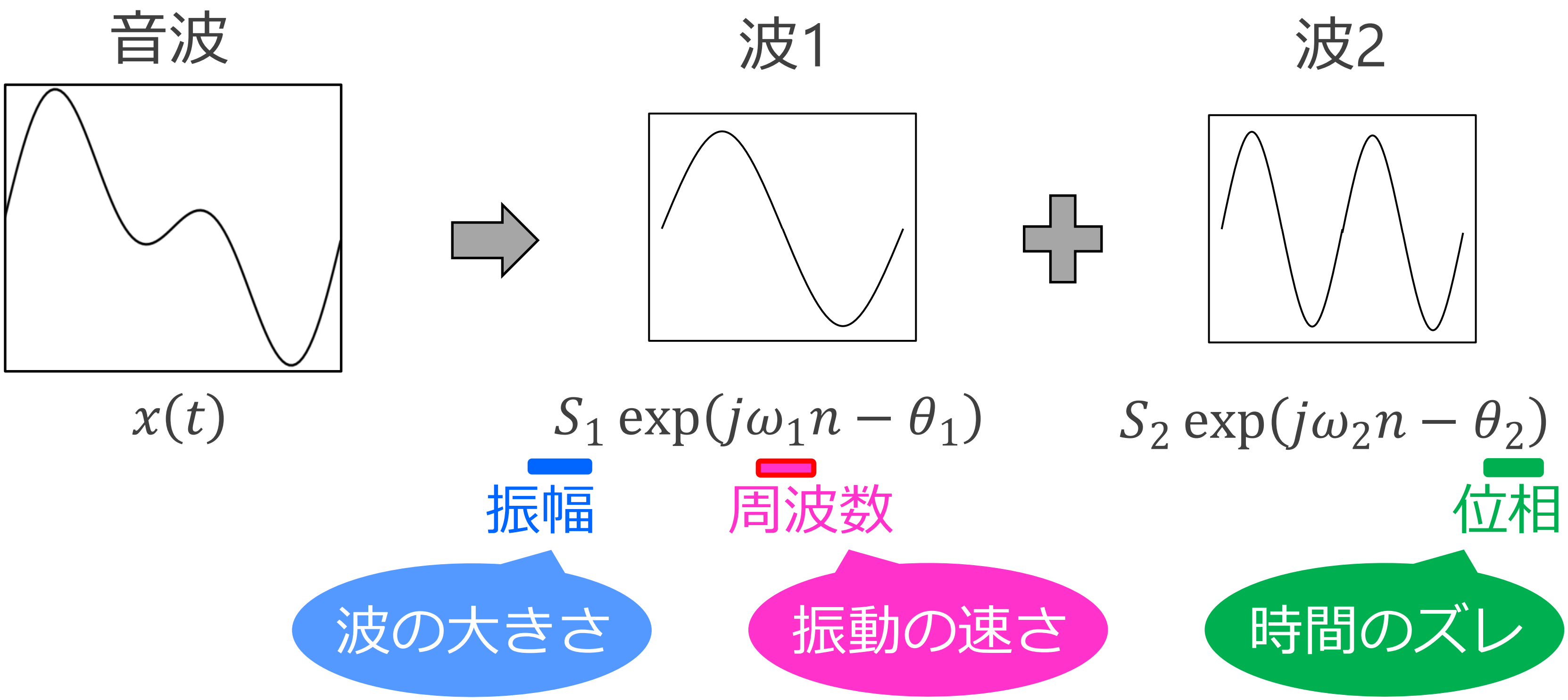
z変換 (z-transform)

- デジタル信号を「時間とともに増加・減衰しながら振動する波」の和で表現
- ラプラス変換の離散版

フーリエ変換 (直感的なイメージ)

フーリエ変換

連続時間の波を 振動する波 $\exp(j\omega t)$ の要素で
表現する方法

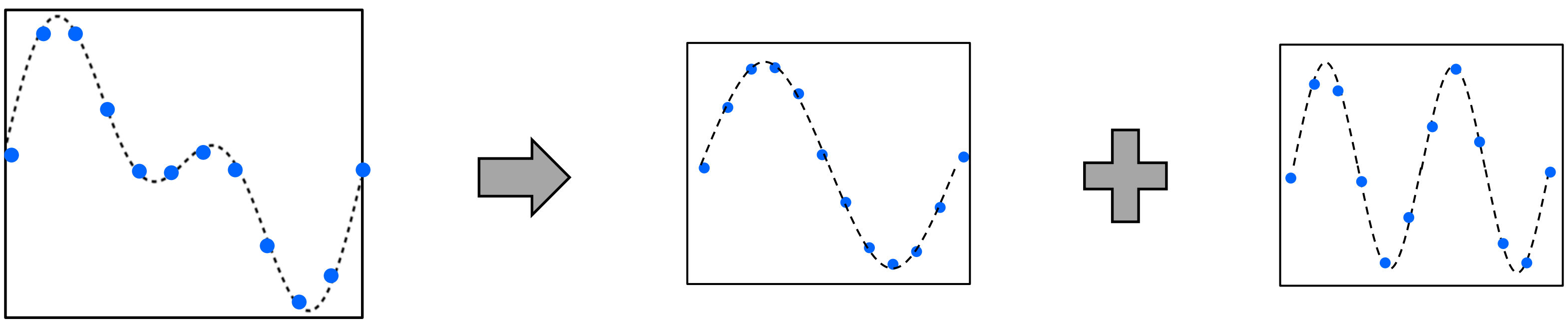


離散フーリエ変換 (直感的なイメージ)

離散フーリエ変換

離散時間の波を 振動する波 $\exp(jkn)$ の要素で表現する方法

音波　　波1　　波2



$x(n)$　　$S_1 \exp(jk_1 n - \theta_1)$　　$S_2 \exp(jk_2 n - \theta_2)$

振幅：波の大きさ

周波数：振動の速さ

位相：時間のズレ

離散フーリエ変換の定義

変数定義

離散時間信号 $\boldsymbol{x} = [x(0), x(1) \cdots, x(n), \cdots, x(N-1)]$ ($x(n)$は実数)

周波数特性 $\boldsymbol{X} = [X(0), X(1) \cdots, X(k), \cdots, X(N-1)]$ ($X(k)$は複素数)

ただし，nは時間インデックス，kは周波数インデックス

時間領域から周波数領域へ（正変換）

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}$$

周波数領域から時間領域へ (逆変換)

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\frac{j2\pi kn}{N}}$$

$X(k)$は複素数なので，極座標形式に直せば，前のページに対応

離散フーリエ変換・z変換

A/D変換した後の音声特徴量抽出

- **離散フーリエ変換**： ケプストラム分析
- **z変換**： LPC分析

離散フーリエ変換 (Discrete Fourier Transform: DFT)

- デジタル信号を「時間とともに振動する波」の和で表現
- フーリエ変換の離散版

z変換 (z-transform)

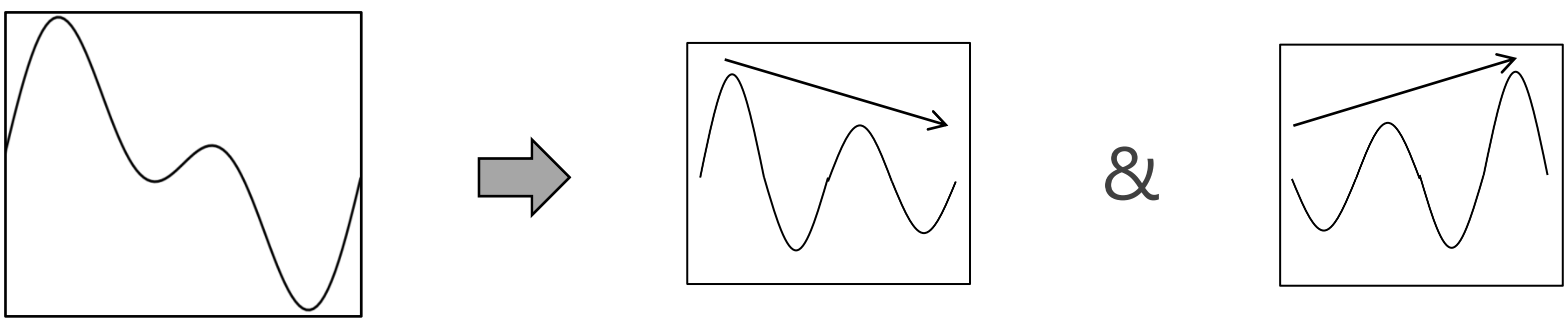
- デジタル信号を「時間とともに増加・減衰しながら振動する波」の和で表現
- ラプラス変換の離散版

Z変換の前準備：
フーリエ変換からラプラス変換へ

ラプラス変換

連続時間の波を増加・減衰しながら振動する波
$\exp\{(\sigma + j\omega)t\}$ の要素で表現する方法

音波　　　　　　　　波1　　　　　　　　波2

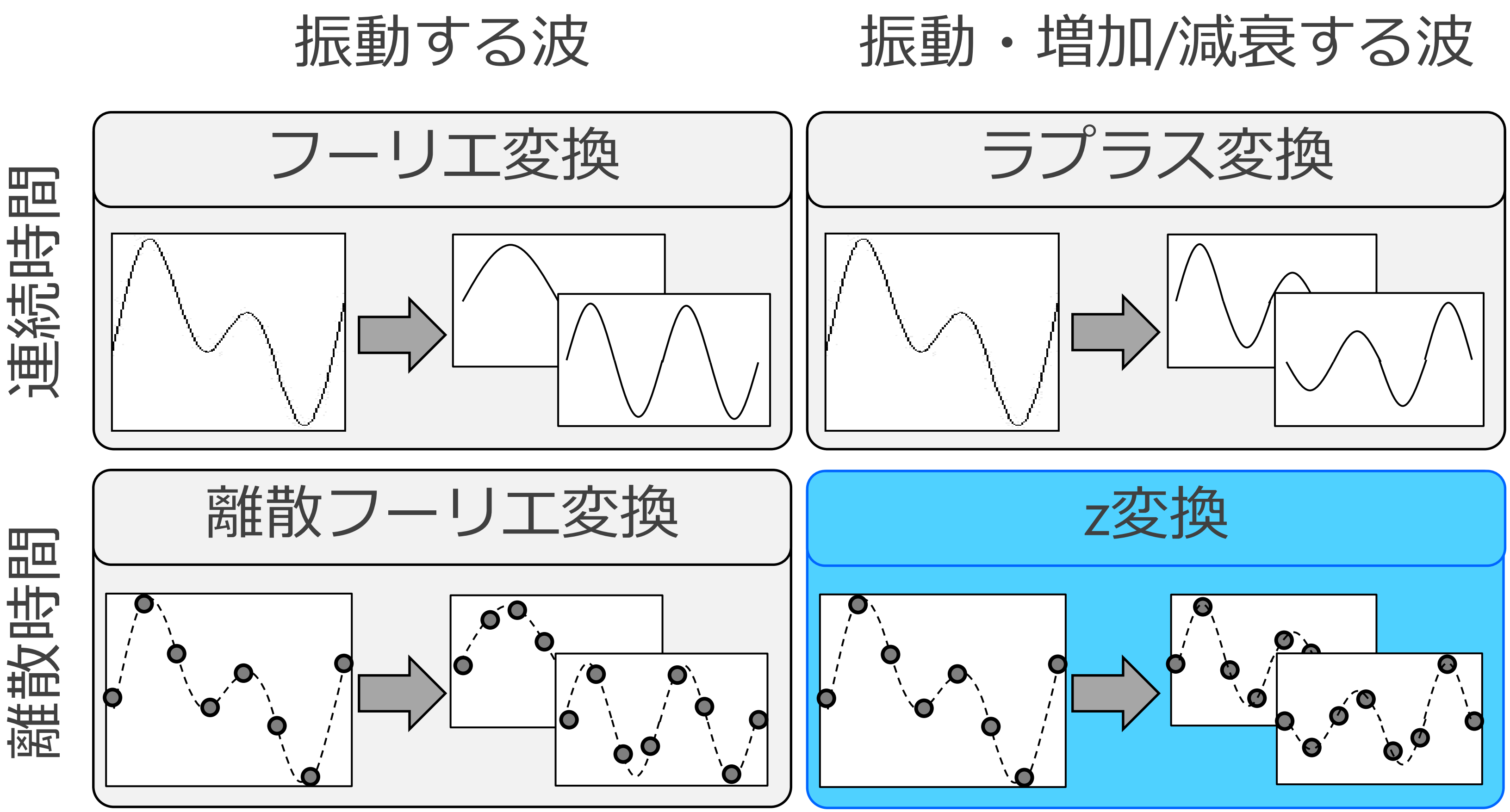


$x(t)$　　$A_1 \exp\big((\sigma_1 + j\omega_1)t - \theta_1\big)$　　$A_2 \exp\big((\sigma_2 + j\omega_2)t - \theta_2\big)$

増減の周波数　　振動の周波数　　位相

増減の速さ　　振動の速さ

各変換法の関係性

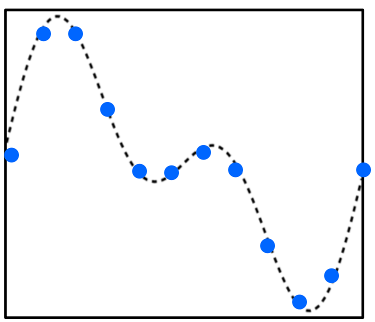


z変換

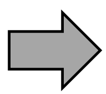
z変換

離散時間の波を増加・減衰しながら振動する波
$z = \exp\{(\sigma + jk)n\}$ の要素で表現する方法

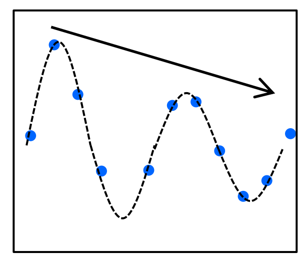
音波



$x(n)$

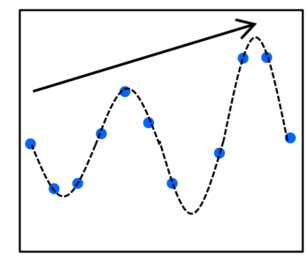


波1



&

波2



$A_1 \exp\big((\sigma_1 + jk_1)n - \theta_1\big)$ $A_2 \exp\big((\sigma_2 + jk_2)n - \theta_2\big)$

増減の周波数 振動の周波数 位相

z変換 (数式)

変数定義

離散時間信号 $\boldsymbol{x} = [x(0), x(1) \cdots, x(n), \cdots, x(N-1)]$ ($x(n)$は実数)

周波数特性 $X(z)$

ただし，nは時間インデックス，zは複素数

時間領域から周波数領域へ（正変換）

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$$

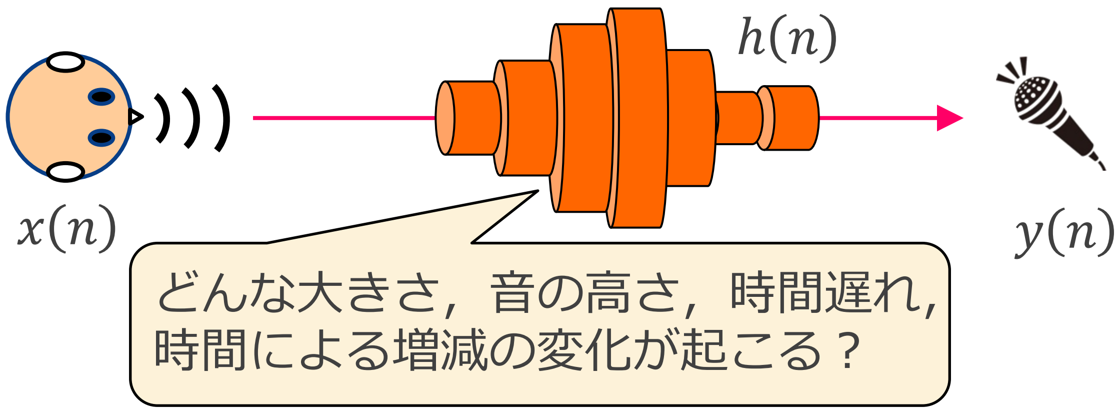
周波数領域から時間領域へ (逆変換)

$$x(n) = \frac{1}{2\pi j} \oint_c X(z) z^{n-1} dz$$

波zが増減と振動を表し，
$X(z)$は，その振幅・位相を表す．

伝達特性

z変換を使うと，経路の伝達特性が分かる！



経路の応答 $h(n)$ **のz変換** $H(z)$ **が，経路の伝達特性を表す！**

$$y(n) = h(n) * x(n)$$

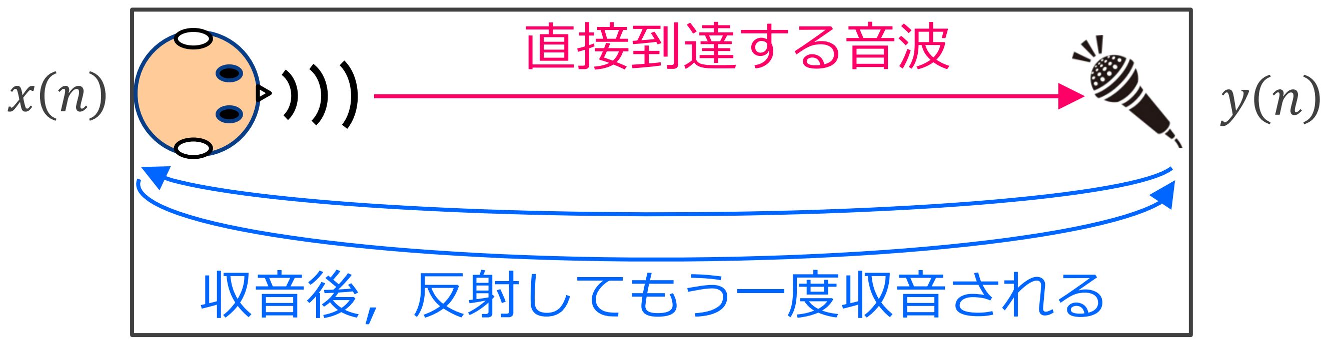
(∗は畳み込み)
z変換で畳み込み演算は掛け算へ

$$Y(z) = H(z)X(z)$$

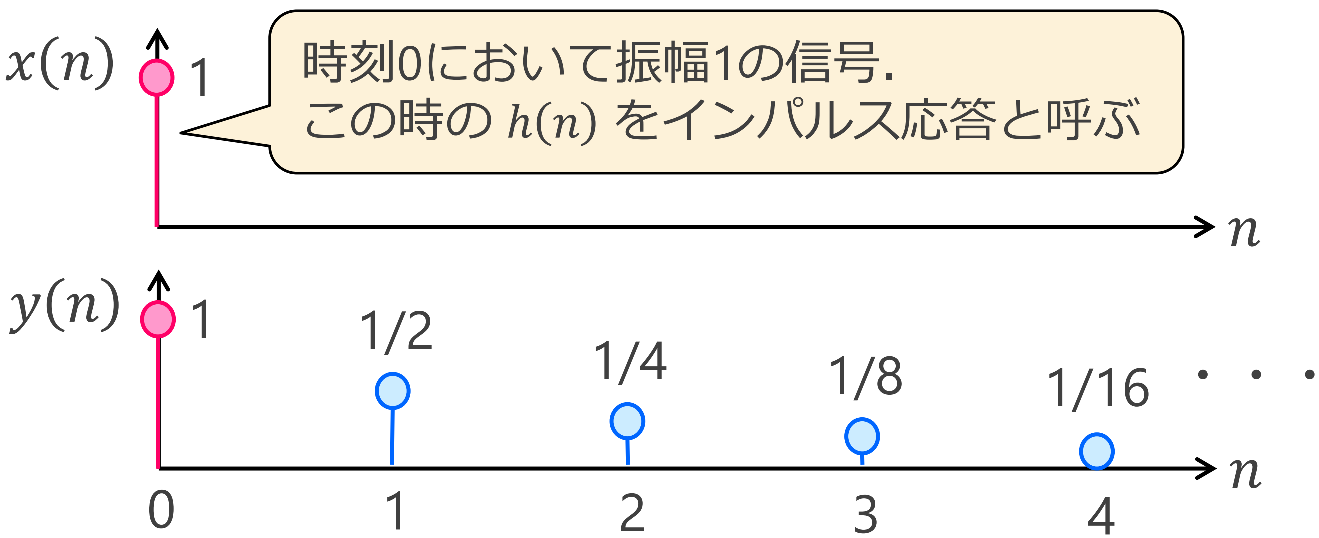
$$H(z) = \frac{Y(z)}{X(z)}$$

z変換を用いたシステム伝達特性

以下のような部屋(音響管)で音を鳴らす



次の音が得られた．音源からマイクロホンへの伝達特性は？



音源からマイクロホンへの伝達特性

$x(n)$ と $y(n)$ を数式で表すと・・・

$$x(n) = \delta(n)$$

$$y(n) = \delta(n) + \frac{1}{2}\delta(n-1) + \frac{1}{4}\delta(n-2)\cdots$$

$$\delta(n) = \begin{cases} 1 & (n = 0) \\ 0 & (n \neq 0) \end{cases}$$

z変換すると・・・

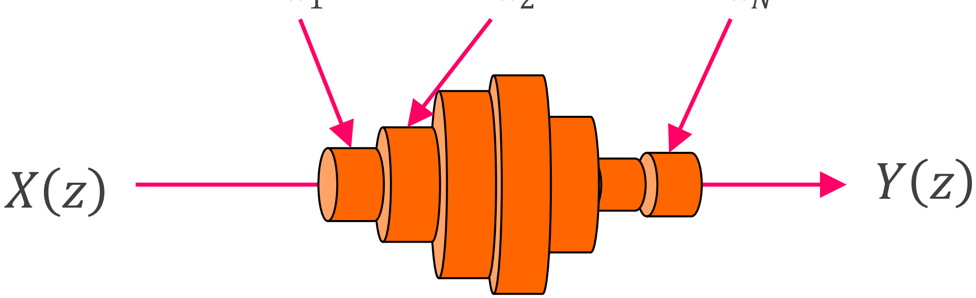
$$X(z) = 1, Y(z) = 1 + \frac{1}{2}z^{-1} + \frac{1}{4}z^{-2}\cdots = \frac{1}{1-\frac{1}{2}z^{-1}}$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1-\frac{1}{2}z^{-1}}$$

単一の共振特性をもつときの伝達特性

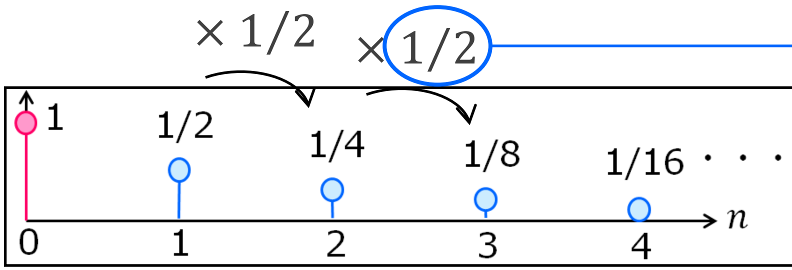
これを踏まえると，複数の共振特性を持った音響管も記述できる

$$H(z) = \frac{1}{1-a_1z^{-1}} \cdot \frac{1}{1-a_2z^{-1}} \cdots \frac{1}{1-a_Nz^{-1}}$$



システムの安定性

時間信号の挙動と伝達特性の関係を考える



自己回帰 (AR) モデル

$$H(z) = \frac{1}{1 - \frac{1}{2} z^{-1}}$$

絶対値が**1より小さい**と，時間とともに**0に収束**＝**安定**
絶対値が**1より大きい**と，時間とともに**無限大に発散**＝**不安定**

時間信号をARモデルで表現する場合，安定性の補償が必要

安定性を保障できない → (例えば) ハウリングを起こす

安定性を保障した分析法 → **LPC分析** (後述)

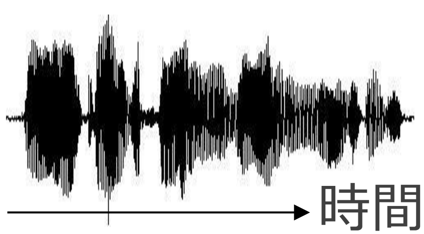
音声とは

音声の生成過程

音色の付与

口や舌を動かして，
音色をつける！

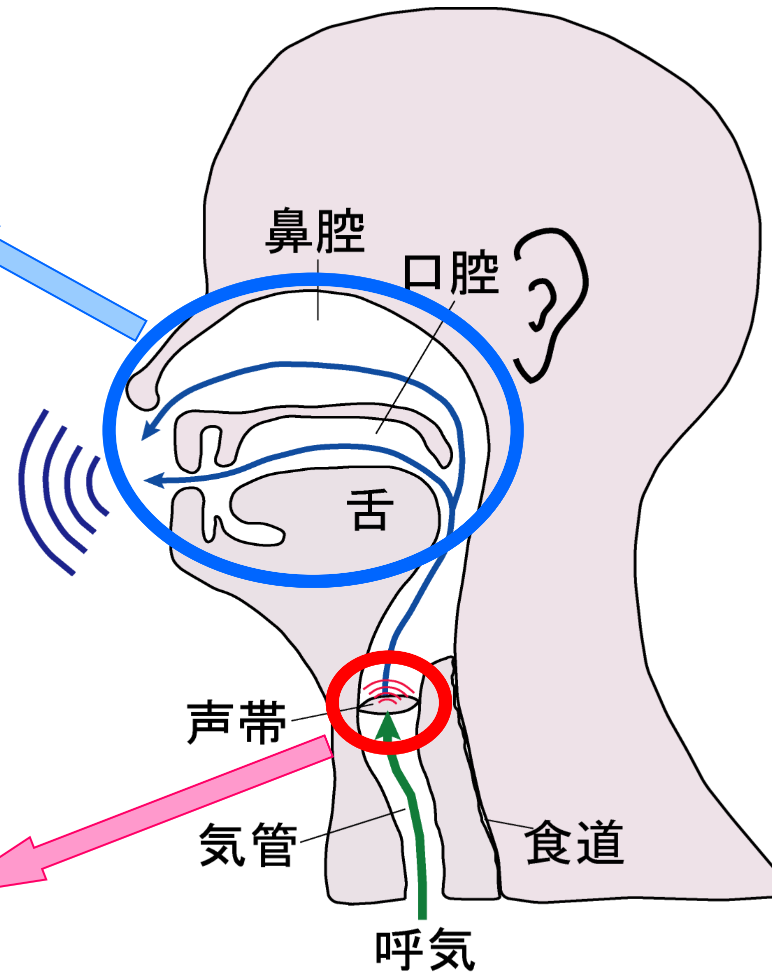
声になる！



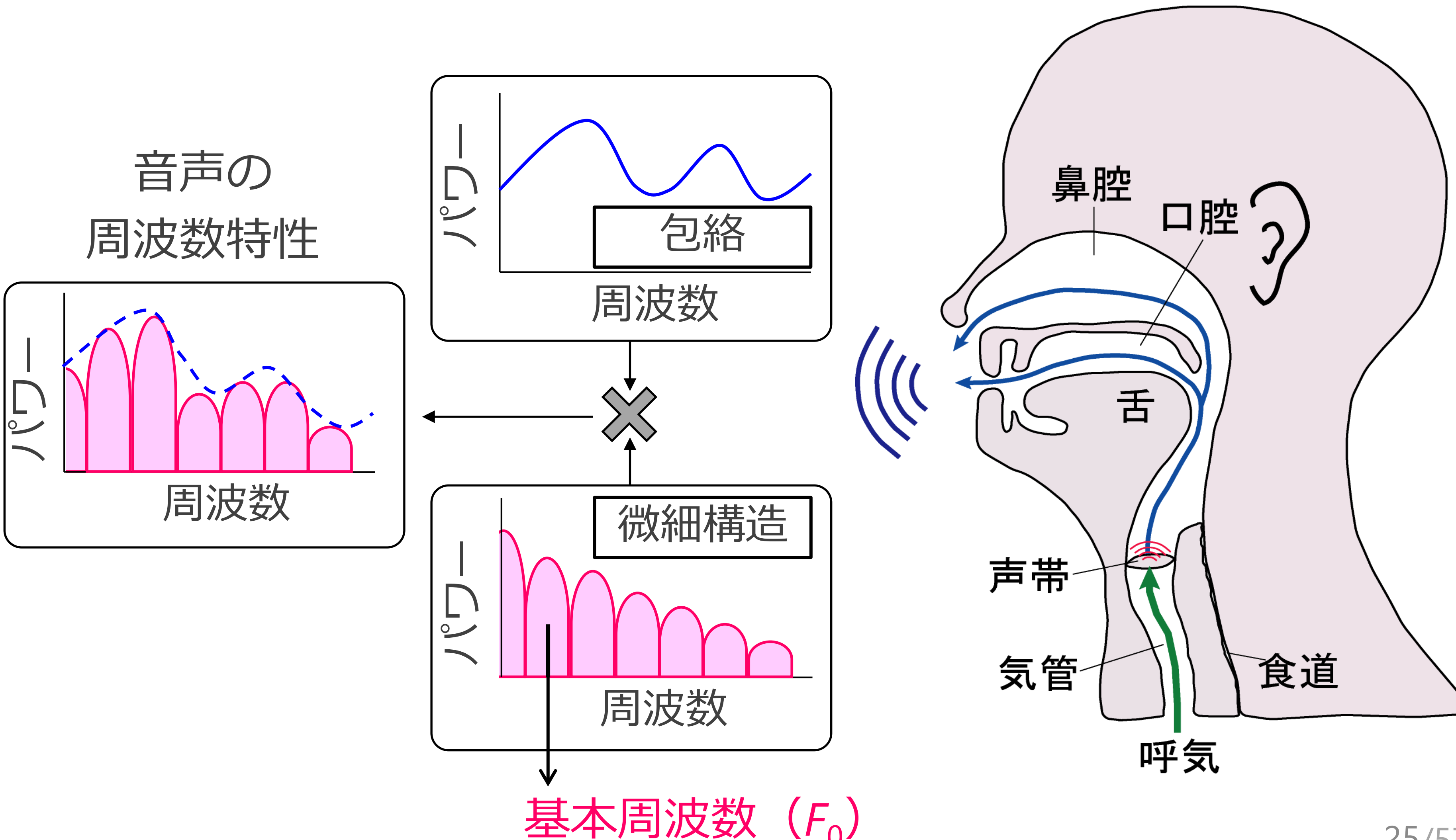
畳み込むと…

音高の生成

声帯を開閉させて，
空気を振動させる！

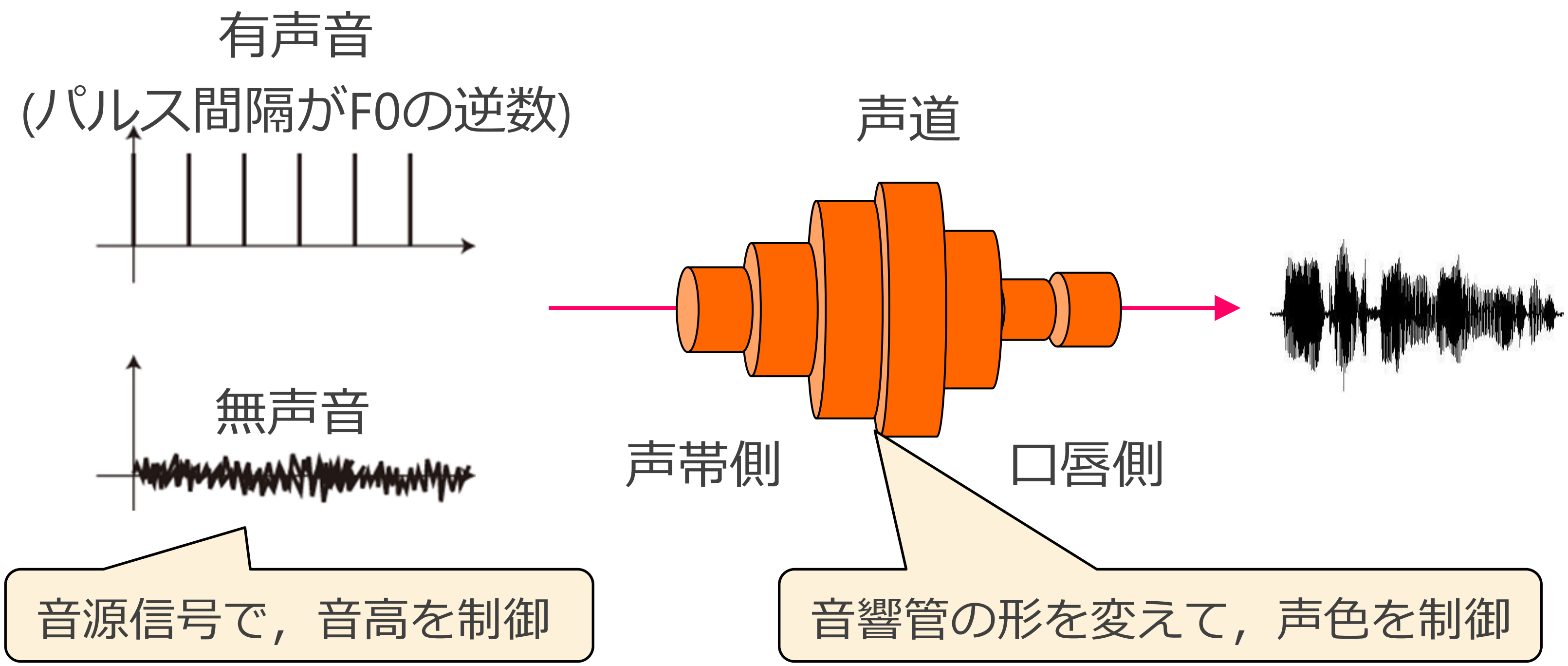


音声のスペクトル構造
（音声のスペクトル構造の2要素）



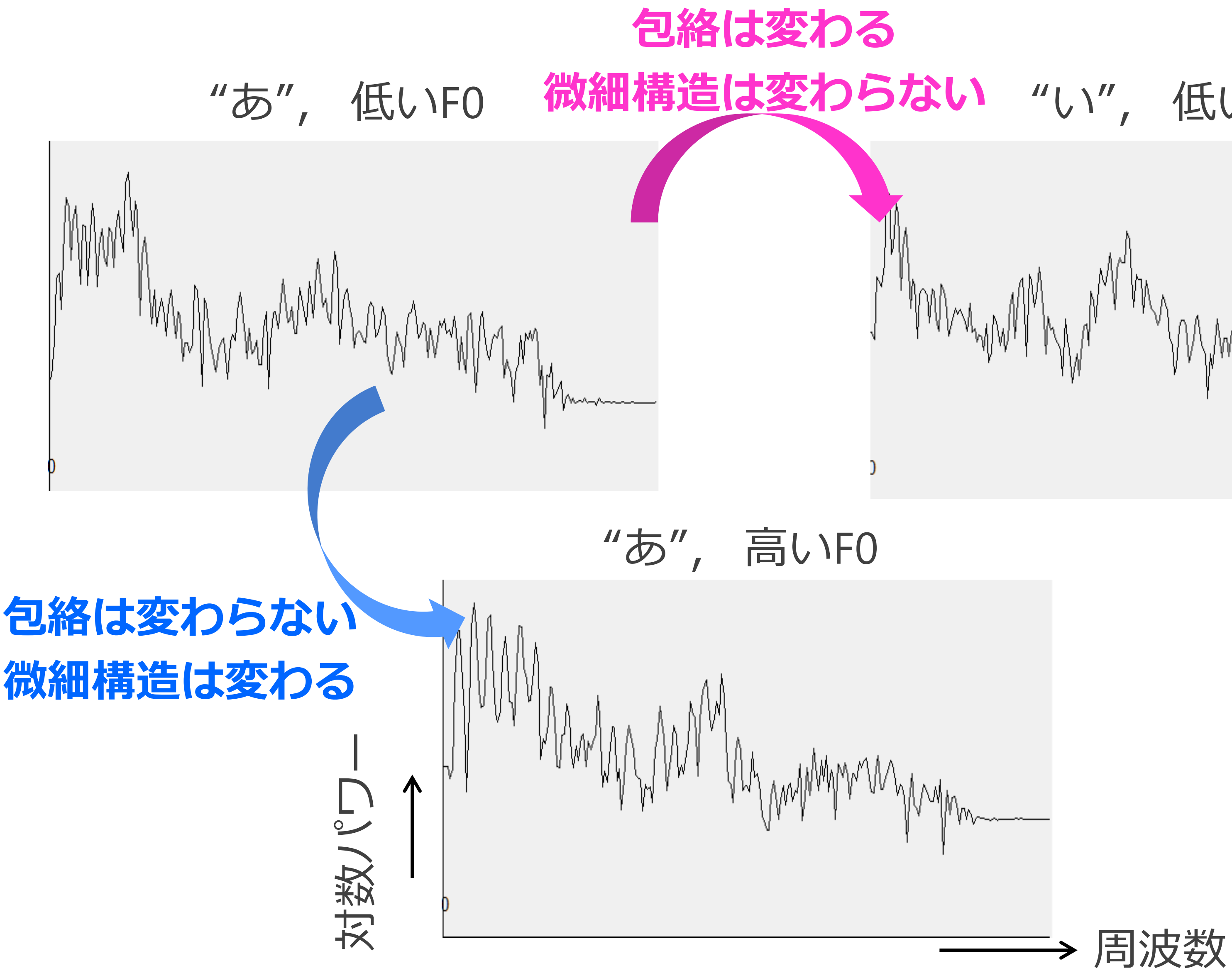
音源生成と，音響管としての声道

音源信号はインパルス列 or 白色雑音，声道は音響管連接



* http://ml.cs.yamanashi.ac.jp/media/20151114/1114slide.pptx から一部引用

スペクトル構造の例


包絡は変わる
微細構造は変わらない
“あ”，低いF0
“い”，低いF0
“あ”，高いF0
包絡は変わらない
微細構造は変わる
対数パワー
周波数
0

スペクトログラム

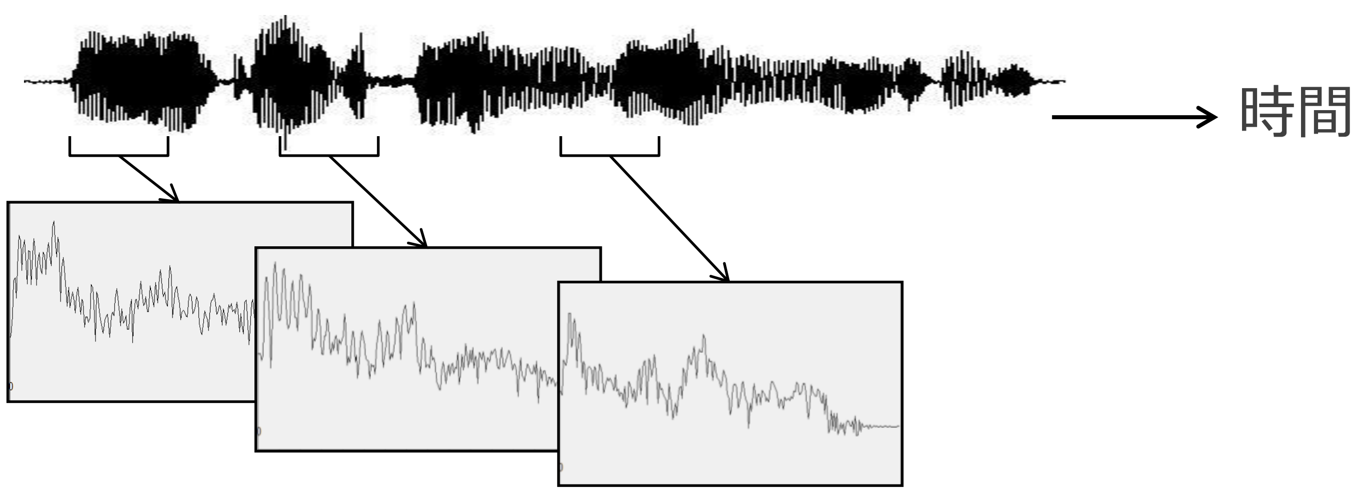
短時間の波形に対するDFT

利点： 比較的定常な部分の静的特徴を見られる

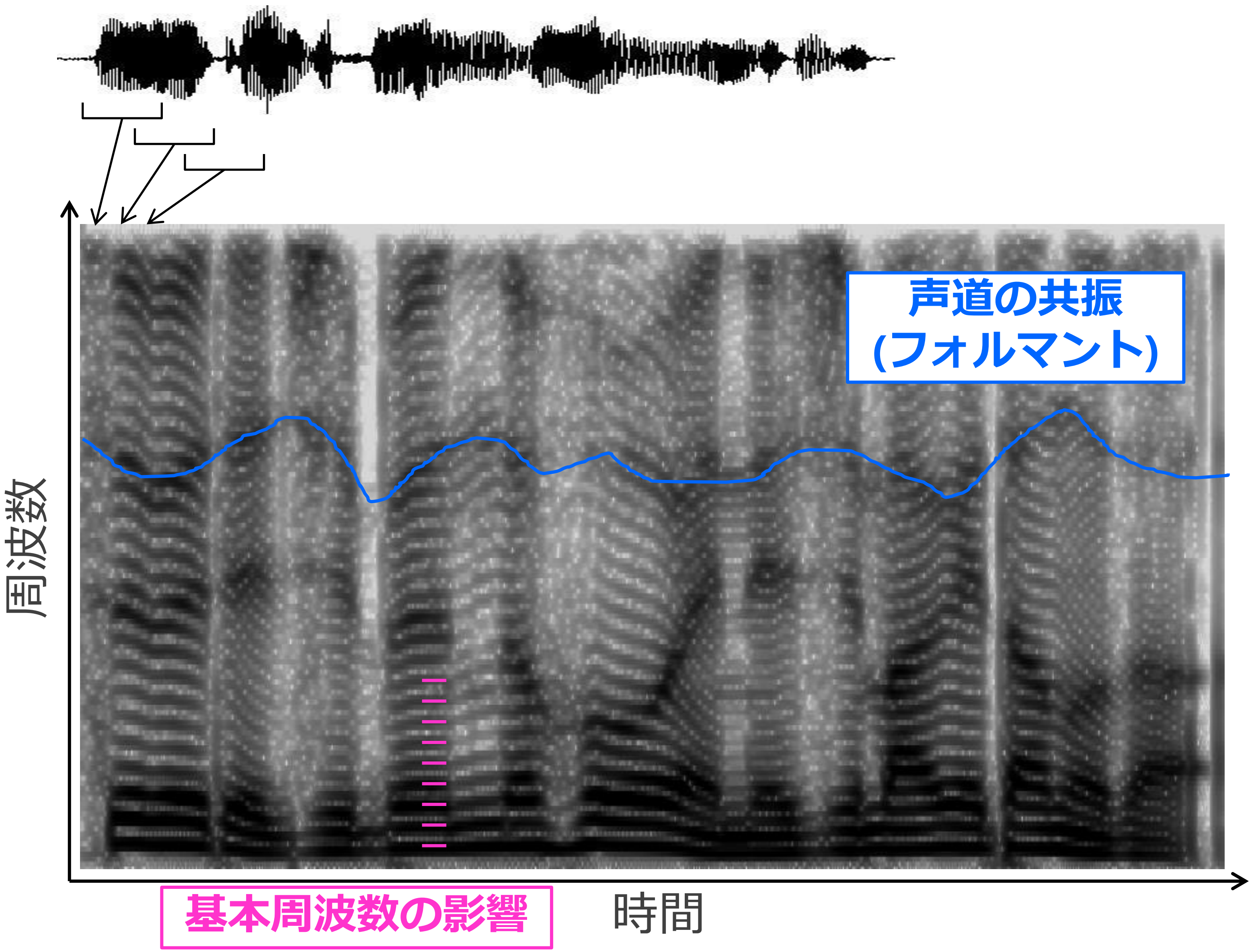
欠点： 音声が定常とみなせるのは数十msec程度なので
音声波形全体がどう変化しているかを見られない

スペクトログラム

離散フーリエ変換による分析を時間軸方向に連続して実行し，
時間ー周波数領域における2次元表示



スペクトログラムの例
(濃いほど，対数パワーが大きい)



音声の特徴抽出

2つの音声分析法：ケプストラムとLPC

ケプストラム分析

- ノンパラメトリックな分析法
- 周波数特性をフーリエ基底で波と捉える
- 時間波形のパワースペクトルの対数のフーリエ変換

LPC (Linear Predictive Coding)分析

- パラメトリックな分析法
- 声道を音響管連接と考え，自己回帰モデルと捉える

2つの音声分析法：ケプストラムとLPC

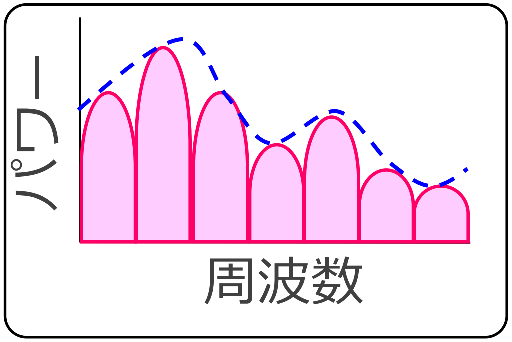
- **ケプストラム分析**
  - ノンパラメトリックな分析法
  - 周波数特性をフーリエ基底で波と捉える
  - 時間波形のパワースペクトルの対数のフーリエ変換

LPC (Linear Predictive Coding)分析

パラメトリックな分析法

声道を音響管連接と考え，自己回帰モデルと捉える

ケプストラムのモチベーション



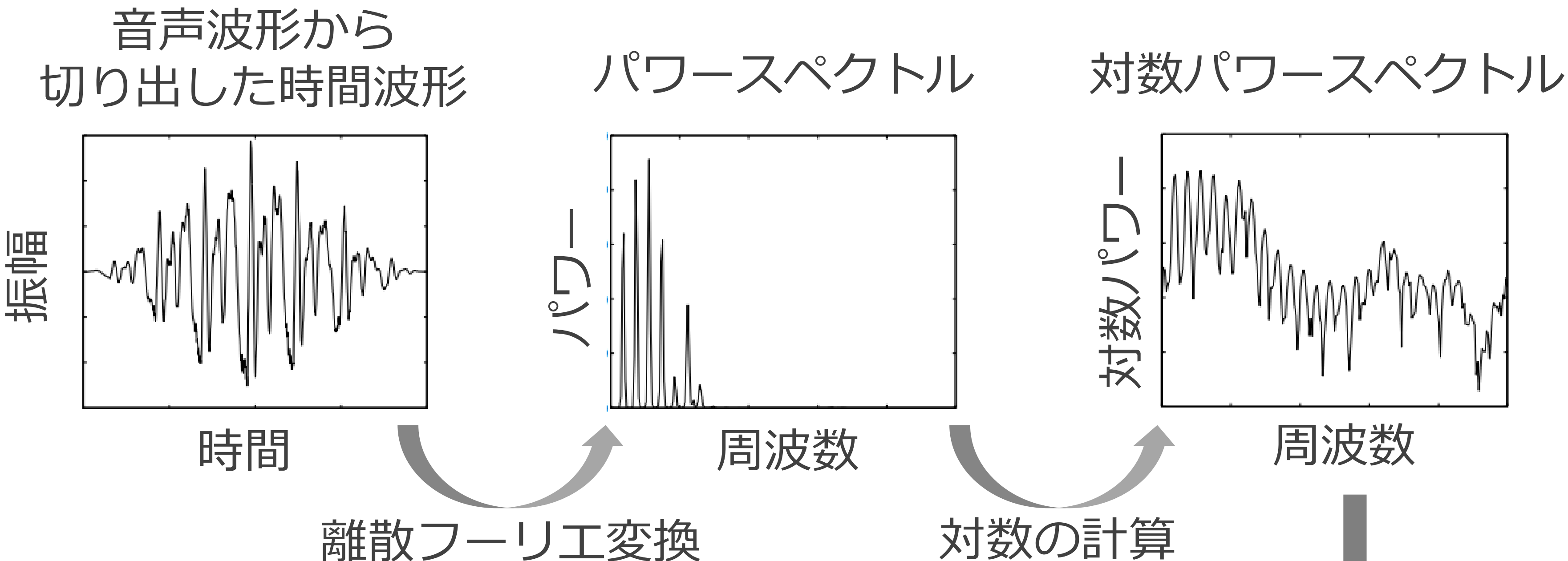
音声から**声道の特性**と**音源の特性**を
抽出（分離）できないかな？
（でも，混ざってるんだよな・・・）

声道の特性と**音源の特性**の形に違いはないかな・・・？

よく見ると，**声道の特性**は緩やかに変動して，
逆に，**音源の特性**は激しく変動しているな．

じゃあ，上図の信号を，**緩やかに振動する低周波数成分**と
激しく振動する高周波数成分に分ければいいんだ！

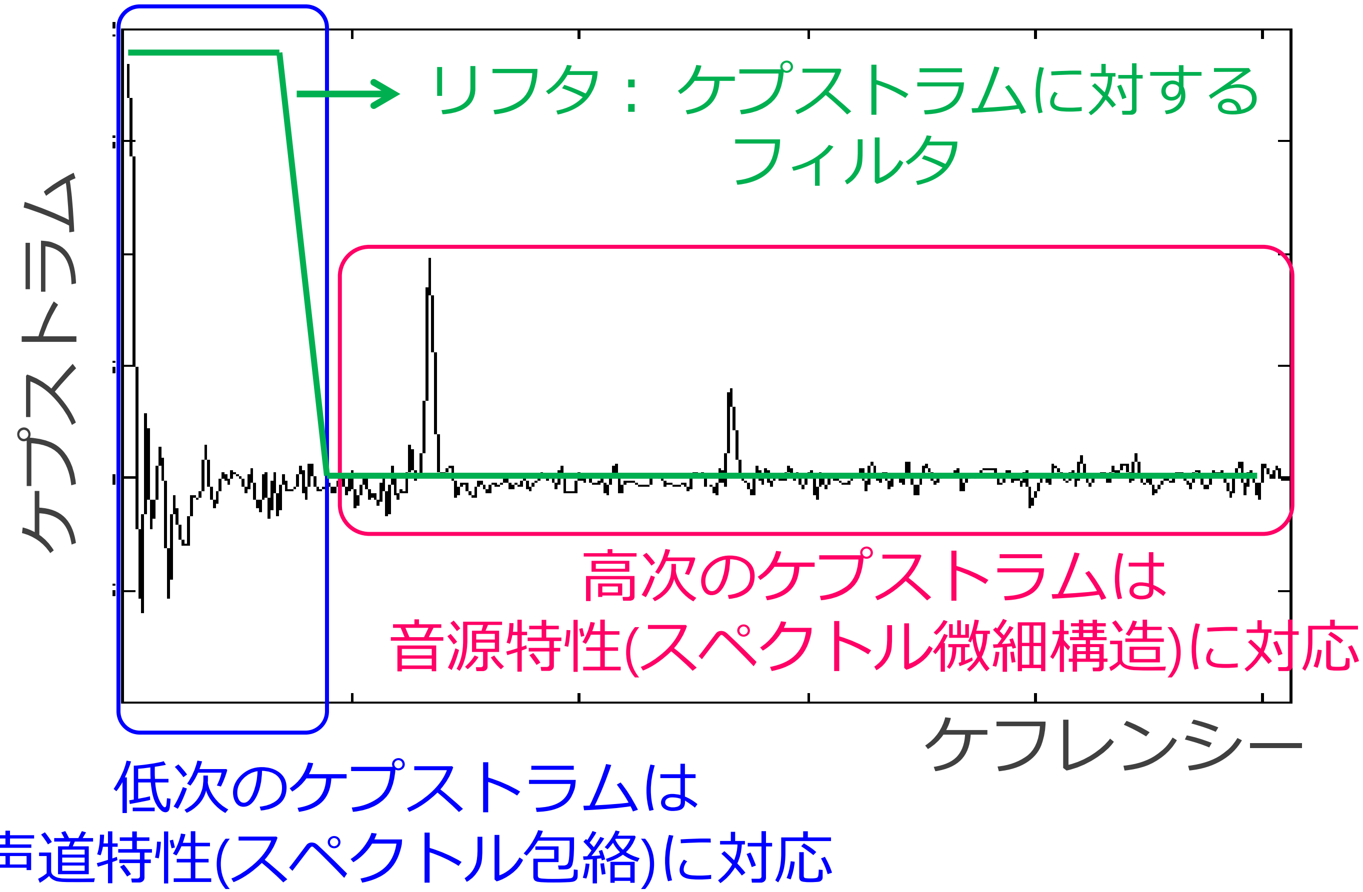
計算手順



対数パワースペクトルを時間波形だと思ってDFT
=> ケプストラムが計算される！

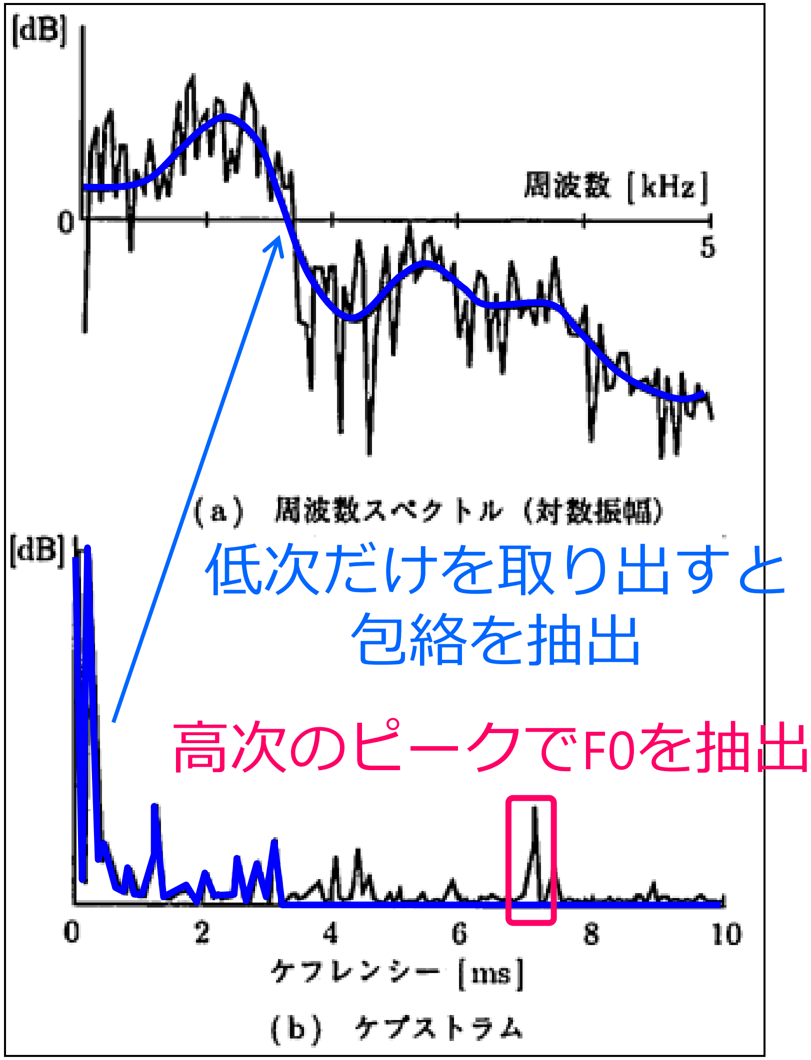
声道特性(包絡)と音源特性(微細構造)が
分離されて現れる(はず)！

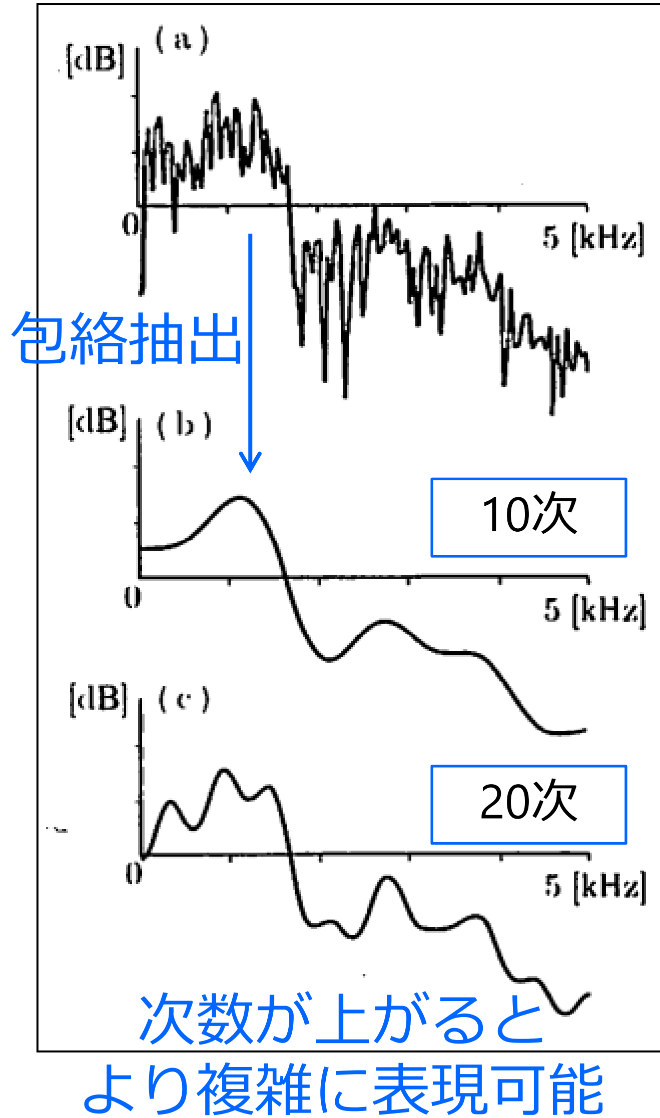
ケプストラムの例



リフタ：ケプストラムに対する
フィルタ
ケプストラム
ケフレンシー
高次のケプストラムは
音源特性(スペクトル微細構造)に対応
低次のケプストラムは
声道特性(スペクトル包絡)に対応

ケプストラムの次数による変化





板橋 他， 音声工学，図4.5から引用

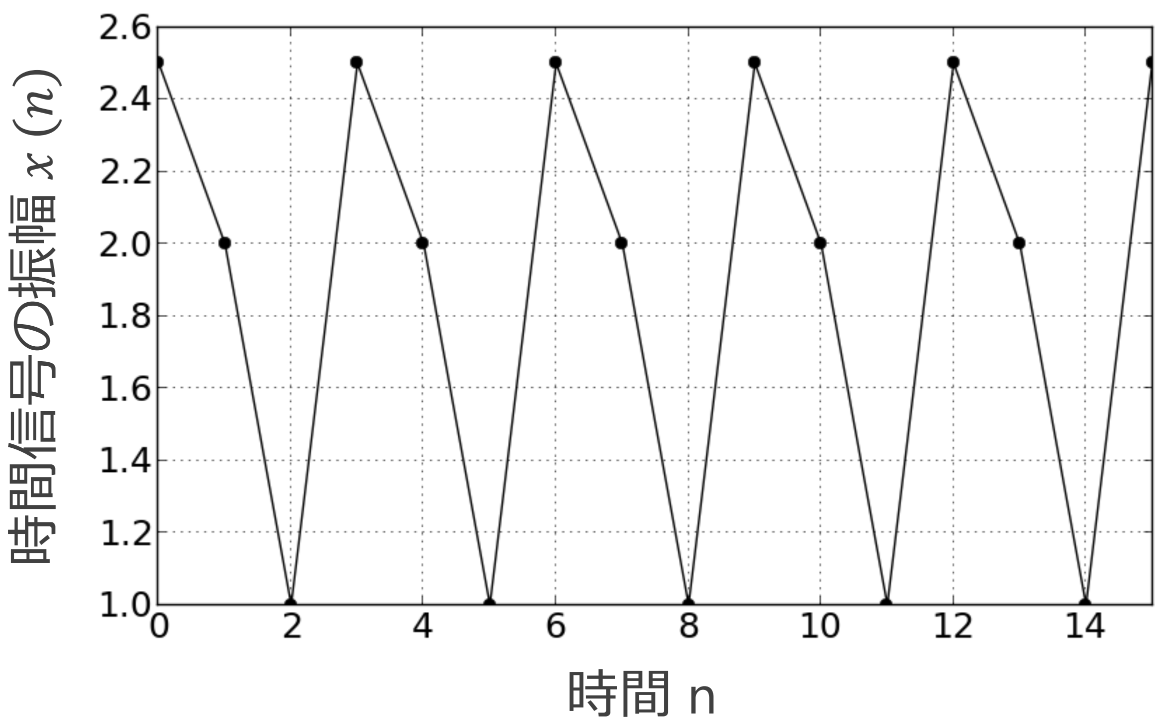
計算してみよう！

Q. 時間信号のスペクトル包絡を抽出せよ．条件は以下の通り

$\boldsymbol{x}$ = (2.5, 2.0, 1.0, 2.5, 2.0, 1.0, 2.5, 2.0, 1.0, 2.5, 2.0, 1.0, 2.5, 2.0, 1.0, 2.5)

信号長さN: 16

ケプストラムの次数: 4 (0~3次のケプストラムを残す)

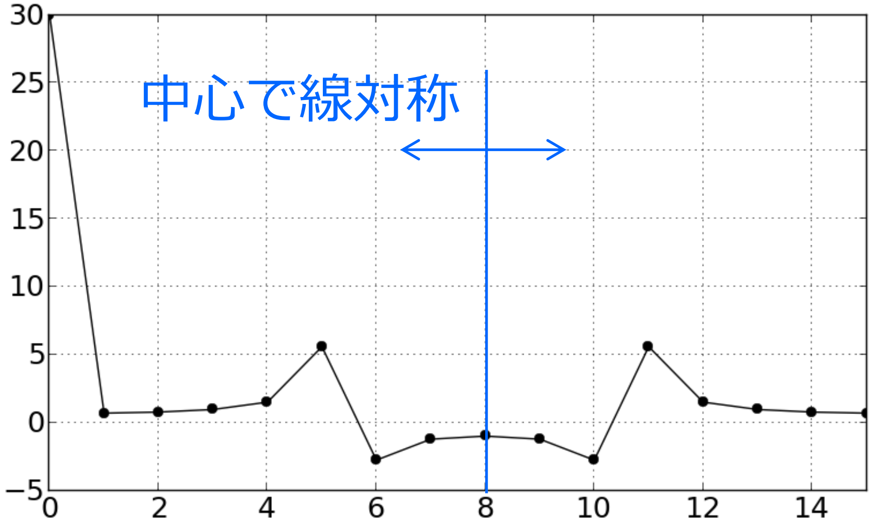


周波数特性 X(k) を計算

$$\boldsymbol{X} = \mathrm{DFT}(\boldsymbol{x})$$

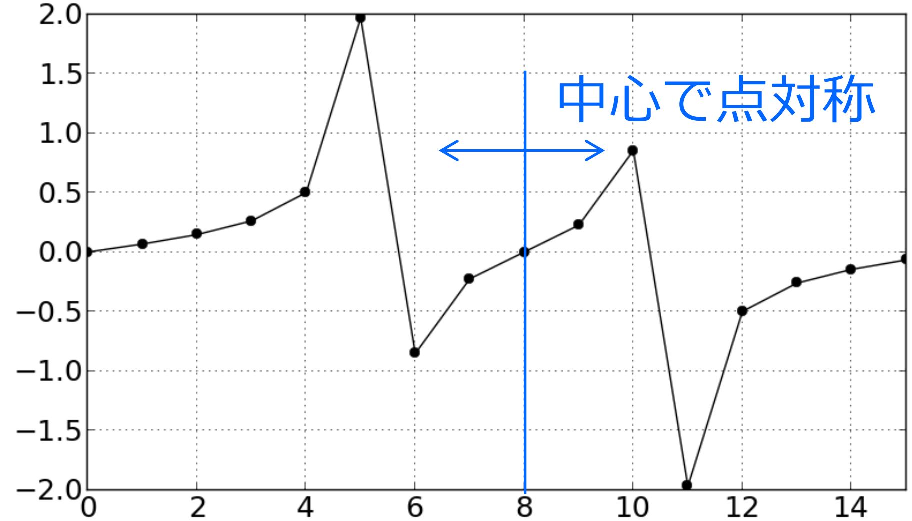
$\boldsymbol{X} = [X(0), \cdots, X(k), \cdots, X(N-1)]$ (各要素は複素数)

Re{$X(k)$}



周波数 k

Im{$X(k)$}

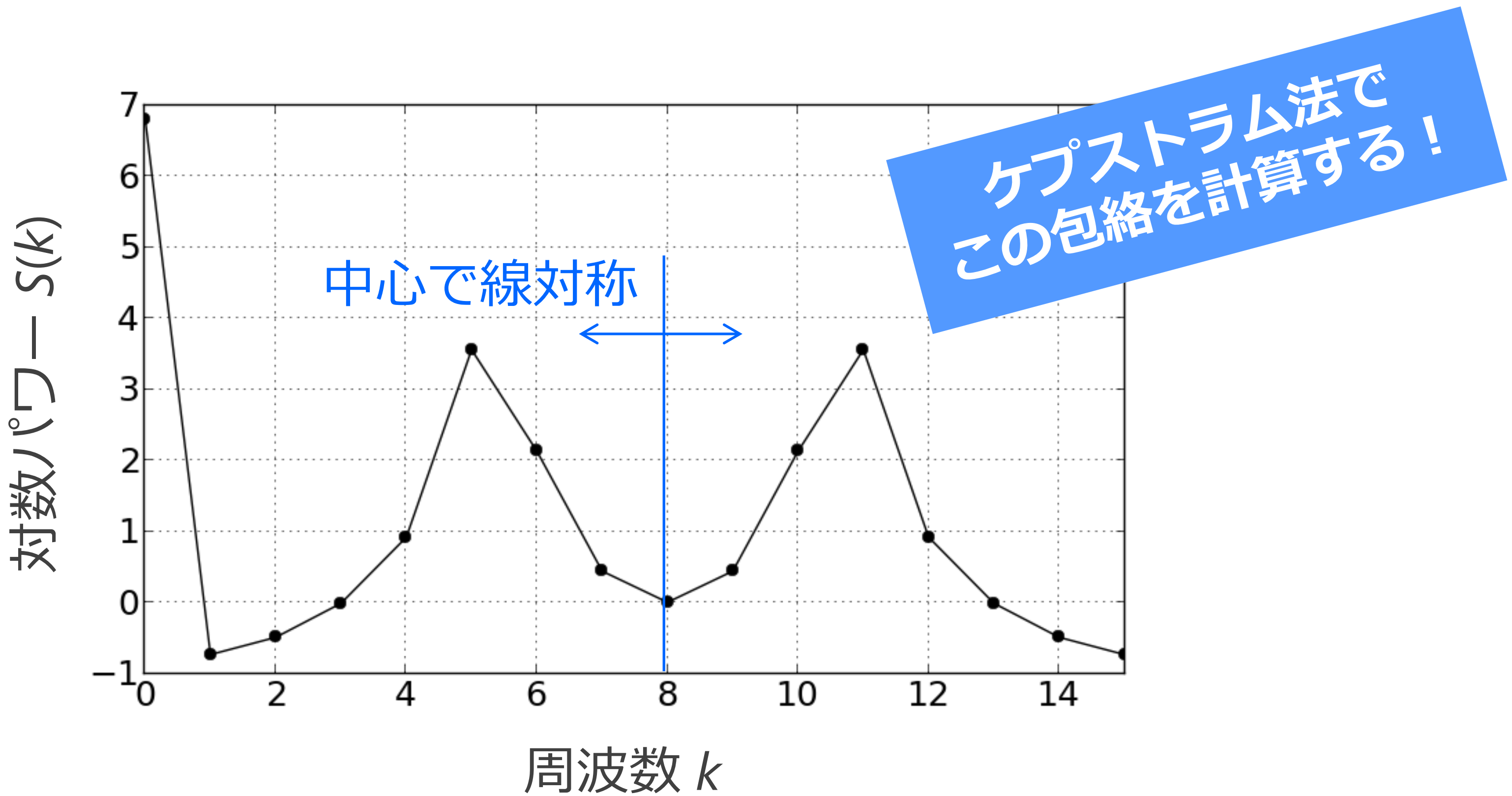


周波数 k

対数パワーを計算

$$S[k] = \log_{10}(|X(k)|^2)$$

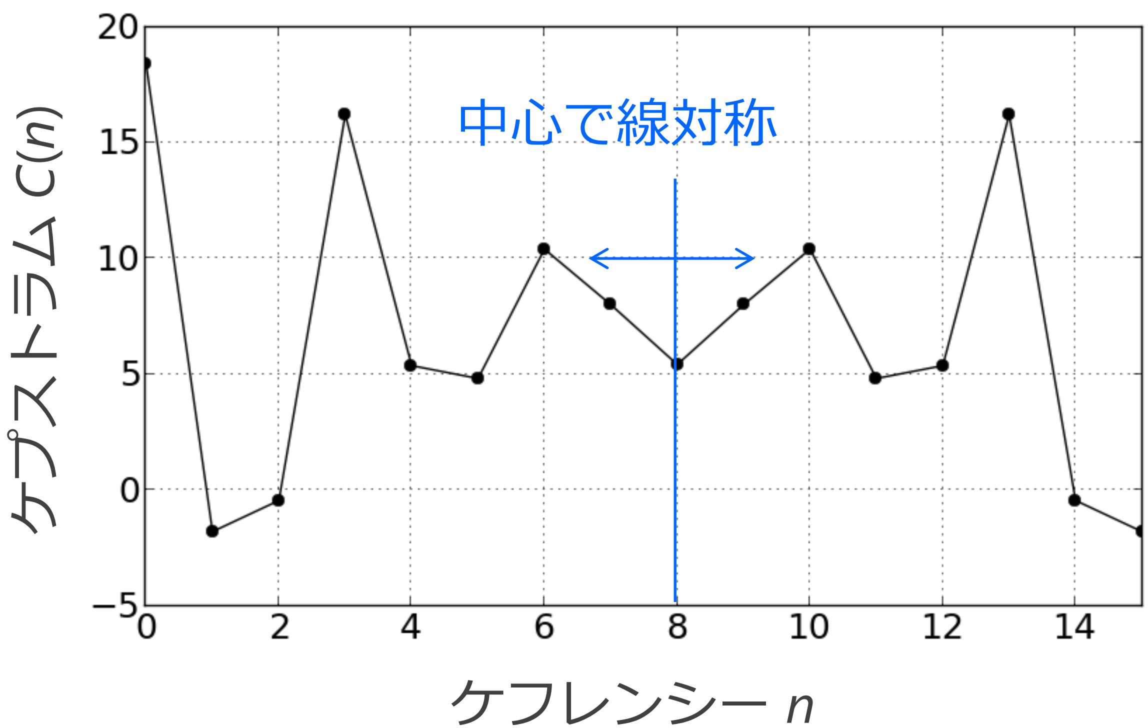
$S(k)$: 周波数kの対数パワー(実数)



ケプストラムを計算 (対数パワーをフーリエ変換)

$$\boldsymbol{C} = \mathrm{DFT}(\boldsymbol{S})$$

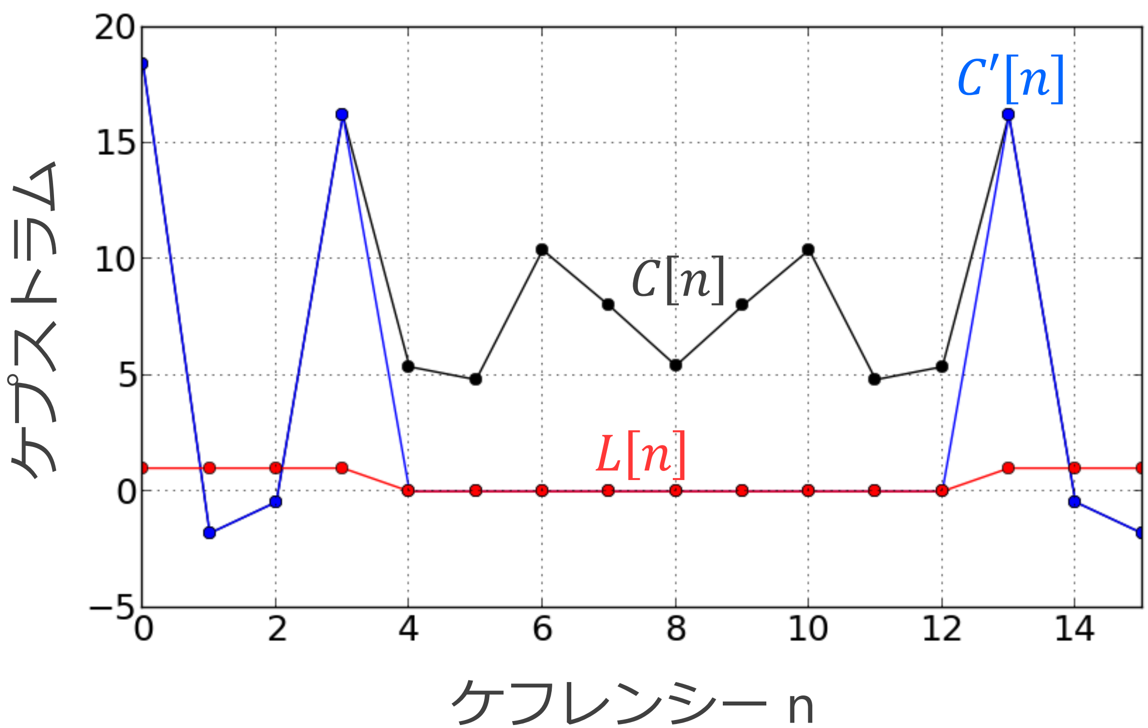
$\boldsymbol{S} = \big(S(0), \cdots, S(k), \cdots, S(N-1)\big)$: 対数パワー (実数)
$\boldsymbol{C} = \big(C(0), \cdots, C(k), \cdots, C(N-1)\big)$: ケプストラム (実数)



リフタをかける

$$C[n]' = L[n]C[n]$$

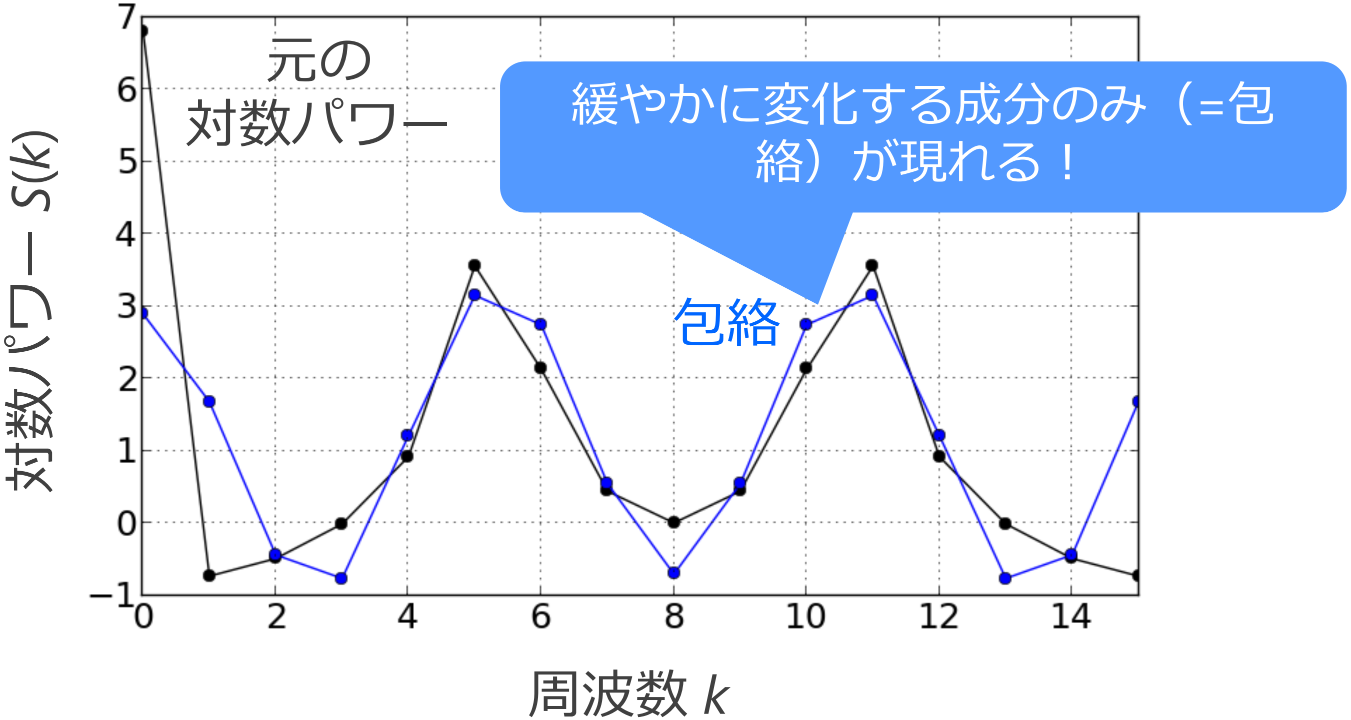
$$L[n] = \begin{cases} 1\ (n \leq 3\ or\ n \geq 13) \\ 0\ (\text{otherwise}) \end{cases}$$: リフタ(中心で線対称)



ケプストラムを逆フーリエ変換

$$\boldsymbol{s}' = \mathrm{IDFT}(\boldsymbol{c}')$$

$\boldsymbol{s}'$: スペクトル包絡， $\boldsymbol{c}'$: リフタリングされたケプストラム



ケプストラム分析の特徴

長所

単純な操作，少ない演算量でスペクトル包絡を抽出可能
高次ケプストラムの考慮により，F0も抽出可能

問題点

リフタリングのカットオフとデータ量のトレードオフ
スペクトル包絡に，フォルマント共振があまり反映されない*
→**共振点に敏感な聴覚系を踏まえると，非効率なモデリング**

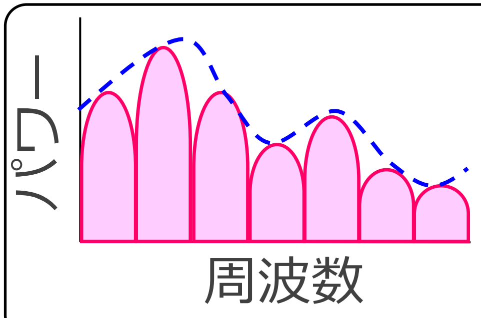
*フォルマントを考慮したケプストラム分析もあるが，本講義では説明しない

2つの音声分析法：ケプストラムとLPC

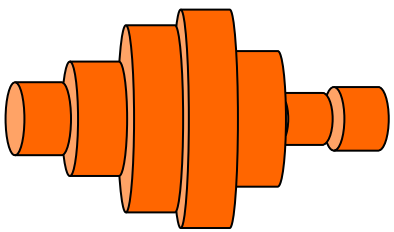
ケプストラム分析
- ノンパラメトリックな分析法
- 周波数特性をフーリエ基底で波と捉える
- 時間波形のパワースペクトルの対数のフーリエ変換

LPC (Linear Predictive Coding)分析
- パラメトリックな分析法
- 声道を音響管連接と考え，自己回帰モデルと捉える

LPCのモチベーション



声道のスペクトル包絡を
効率よくモデル化できないかな？



人間の**声道**って確か，音響管の
連接でモデル化できるよな・・・

そして，音響管の共振で音色が付くんだよね・・・

じゃあ，**声道を音響管だと思って，**
その特性を抽出できればいいんじゃない？

線形予測の原理

音声信号 $x(n)$ について，次式のAR過程が成り立つと仮定

$$\boxed{x(n) + \alpha_1 x(n-1) + \cdots + \alpha_p x(n-p) = e(n)}$$

$e(n)$: $N(\cdot, 0, \sigma^2)$に従う**線形予測誤差**

α_i: **線形予測係数**

$e(n)$ **を最小にするように** α_i **を決める**

上式のz変換は以下の通り与えられる

$$X(z) + \alpha_1 X(z) z^{-1} + \cdots + \alpha_p X(z) z^{-p} = E(z)$$

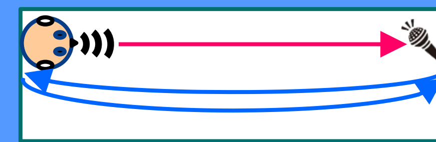
$$X(z) = \frac{1}{1 + \alpha_1 z^{-1} + \cdots + \alpha_p z^{-p}} E(z)$$

線形予測係数は何を表している？

この式は何を表す？

$$X(z) = \frac{1}{1 + \alpha_1 z^{-1} + \cdots + \alpha_p z^{-p}} E(z)$$

部屋での共振の式！



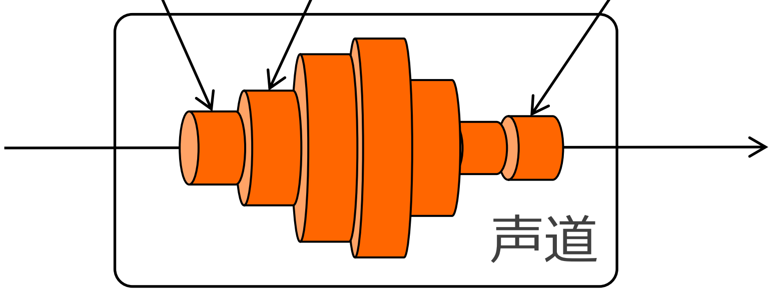
因数分解してみる

$$X(z) = \left(\frac{1}{1 + \beta_1 z^{-1}}\right)\left(\frac{1}{1 + \beta_2 z^{-1}}\right) \cdots \left(\frac{1}{1 + \beta_p z^{-1}}\right) E(z)$$

つまり…

$E(z)$

声帯の音源信号



$X(z)$

音声信号

声道を音響管の連接と捉え，その特性を推定している！

線形予測の推定(1)

LPC分析で推定される線形予測係数は，AR過程を仮定

つまり「**声帯信号のパワーを最小化するようにARモデルを推定**」しており，「**声道特性を共振のみで表現**」する分析法

どうやって，線形予測係数を推定する？

当該時間区間内の声帯信号のパワーを最小化する (次のページへ)

$$\rightarrow \underbrace{\frac{\partial}{\partial \alpha_i}}_{\text{i番目の予測係数}} \left(\underbrace{\sum_{n=n_0}^{n=n_1}}_{\text{時間区間}} \underbrace{e(n)^2}_{\text{残差(声帯信号)}} \right) = 0$$

線形予測の推定(2)

予測残差を展開

$$\sum_{n=n_0}^{n_1} e(n)^2 = \sum_{n=n_0}^{n_1}\left(\sum_{i=0}^{p}\alpha_i x(n-i)\right)^2$$

和の二乗を展開

$$= \sum_{n=n_0}^{n_1}\sum_{i=0}^{p}\sum_{j=0}^{p}\alpha_i\alpha_j x(n-i)x(n-j)$$

nに関する総和を
自己相関関数へ

$$= \sum_{i=0}^{p}\sum_{j=0}^{p}\alpha_i\alpha_j v_{ij}$$

$$\sum_{n=n_0}^{n_1} x(n-i)x(n-j)$$ **自己相関関数**

上式は，α_i に関する2次式であるため， α_iによる微分=0と置くと解けるが，安定して解が求まる保証はない

→ 条件を導入

線形予測の推定(3)

条件

当該時間区間外では $x(n) = 0$

無限長の信号を考える ($n_0 = -\infty$, $n_1 = \infty$)

この条件下で自己相関関数は次式のように変形できる

$$v_{ij} = \sum_{n=n_0}^{n_1} x(n-i)x(n-j) = \sum_{n=n_0}^{n_1} x(n)x(n-|i-j|) = r_{|i-j|}$$

iとjの2変数に依存していた自己相関関数が
$|i-j|$の1変数のみに依存

この変形により安定して解を推定できる（次ページ）

線形予測の推定(4)

微分値を0とおいてα_iを推定

$$\frac{\partial}{\partial \alpha_i}\left(\sum_{n=-\infty}^{\infty} e(n)^2\right) = \frac{\partial}{\partial \alpha_i}\left(\sum_{i=0}^{p}\sum_{j=0}^{p} \alpha_i \alpha_j v_{ij}\right) = 2\sum_{j=0}^{p} \alpha_j v_{ij} = 0$$

$\alpha_0 = 1$として変形

$$\sum_{j=1}^{p} \alpha_j v_{i,j} = v_{i0}$$

行列で表現すると…

$$\begin{bmatrix} v_{1,1} & \cdots & v_{1,j} & \cdots & v_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{i,1} & \cdots & v_{i,j} & \cdots & v_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{p,1} & \cdots & v_{p,j} & \cdots & v_{p,p} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_j \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} v_{1,0} \\ \vdots \\ v_{i,0} \\ \vdots \\ v_{p,0} \end{bmatrix}$$

線形予測の推定(5)

安定化条件による導出 $v_{i,j} = r_{|i-j|}$ **を代入すると…**

$$
\begin{bmatrix}
r_0 & r_1 & r_2 & \cdots & r_{p-1} \\
r_1 & r_0 & r_1 & \ddots & \vdots \\
r_2 & r_1 & r_0 & \cdots & \vdots \\
\vdots & \ddots & \vdots & \ddots & r_1 \\
r_{p-1} & \cdots & \cdots & r_1 & r_0
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \alpha_p
\end{bmatrix}
=
\begin{bmatrix}
r_1 \\ r_2 \\ \vdots \\ \vdots \\ r_p
\end{bmatrix}
$$

テプリッツ型行列 → 正定値行列 → 逆行列が必ず存在

利点

- 線形予測係数が必ず求まる
- 高速解法(Durbinの再帰的解法)が利用可能
- 推定されたARモデルは絶対安定

線形予測分析とケプストラム分析の比較

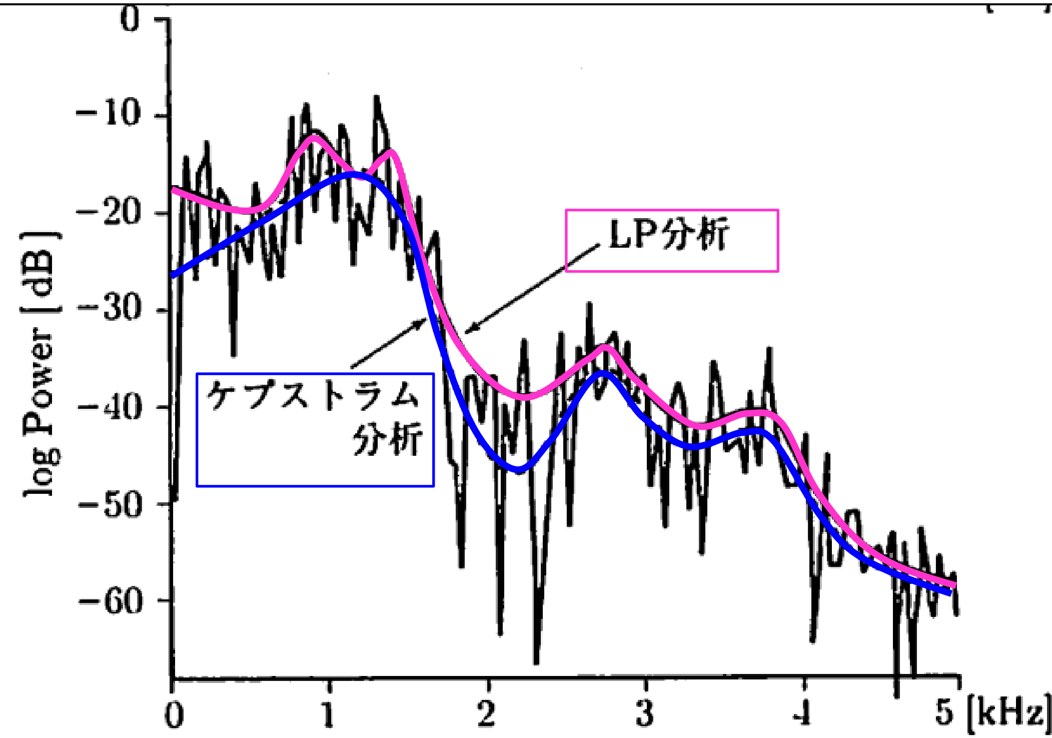
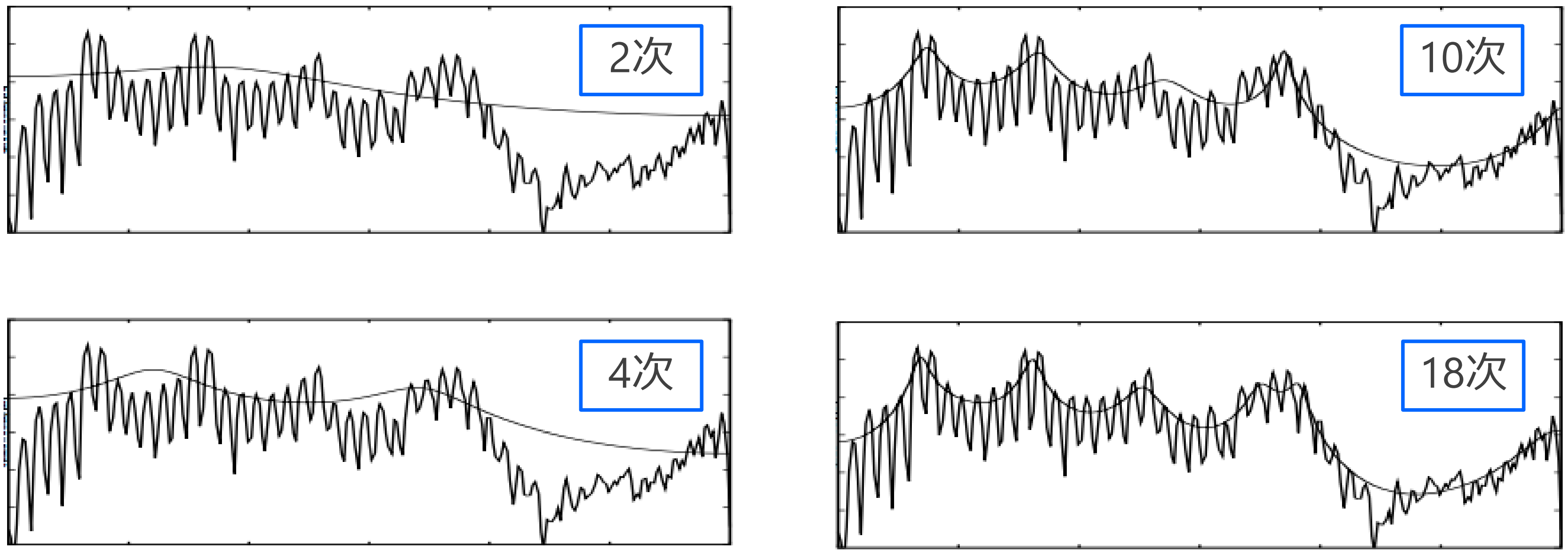


図 4.13　自己相関法で推定された母音のスペクトル包絡 (実線，次数は 13). 点線はケプストラム分析により得られたスペクトル包絡を示す.

ケプストラム分析より，フォルマント(ピーク)を重視
→ 少ない次数によりスペクトルを表現可能

* 板橋 他， 音声工学，図4.13より引用

線形予測分析の次数による違い



ケプストラムと同じように，
次数が増えるほど細かくモデル化できる

* 嵯峨山茂樹, “応用音響学 講義資料 2009”より引用

線形予測分析の特徴

長所

高速解法により，単純な操作でスペクトル包絡を抽出可能

フォルマントを強調した包絡を抽出

少量のパラメータ数で効率的に包絡を表現

問題点

線形予測係数を量子化・伝送する場合，伝送誤差等により
不安定なフィルタになりやすい

(一つの線形予測係数の誤差だけで，安定性が崩れる)

→ **PARCORやLSPによる改善**

PARCORやLSP

PARCOR分析

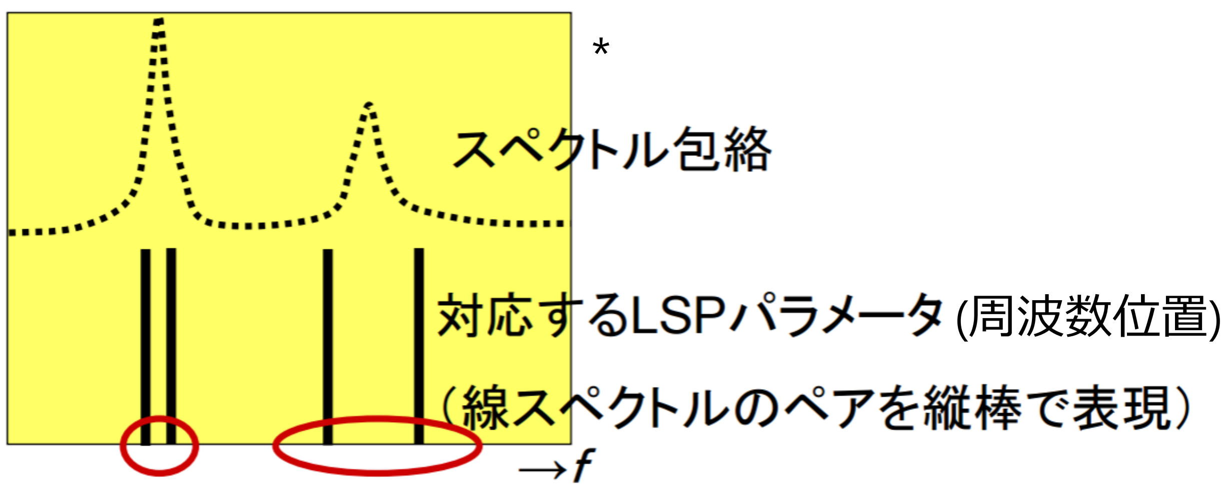
線形予測係数を，音響管の各管の反射係数に変換

反射係数は1を超えない → 伝送誤差で1を超えても後処理で補償

→ **絶対安定な伝達関数を受信可能**

LSP (線スペクトル対) 分析

PARCOR係数を周波数領域へ → **安定性を保持しつつ時間方向でのスペクトル補間が可能** → 時間方向での情報削減が可能



*猿渡 洋, "音情報処理論音声処理における信号処理2"より引用

本講義のまとめ

音声の特徴とは何か，それをどう定量化するか

デジタル信号処理の基礎

離散フーリエ変換 ... 振動する波で音声を表現

z変換 ... 増減・振動する波で音声を表現．安定性を図れる．

音声とは

音声の生成過程 ... スペクトル包絡・基本周波数

音声の特徴抽出

ケプストラム分析 ... 対数パワースペクトルを時間波形と捉える

LPC分析 ... 声道を音響管連接と捉える