言語情報処理論
2007-11-28

峯松 信明
東京大学大学院新領域創成科学研究科

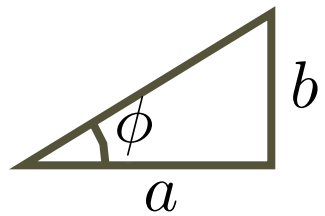
前回のおさらい

- **フーリエ級数展開について＋α**
  - サイン君＋コサイン君→サイン君
  - フーリエ級数展開を使った波形変形　～フィルタリング～
- **分節的特徴と韻律的特徴，及びその抽出方法**
  - 音声生成とソース・フィルターモデル
  - スペクトル包絡特性の推定方法　～ケプストラム法とLPC～
  - 基本周波数の推定方法　～ケプストラム法と自己相関～
- **音声の分析合成系とその応用　～音声変形～**
  - 高性能分析合成系 STRAIGHT
  - 感情・年齢・性別・音韻，色んなものを変形しちゃいます
- **音声合成の原理と波形編集型音声合成**
- **まとめ**

サイン君＋コサイン君→サイン君

合成の公式

$$a\sin\theta + b\cos\theta = \sqrt{a^2+b^2}\sin(\theta+\phi), \quad \phi = \tan^{-1}\frac{b}{a}$$



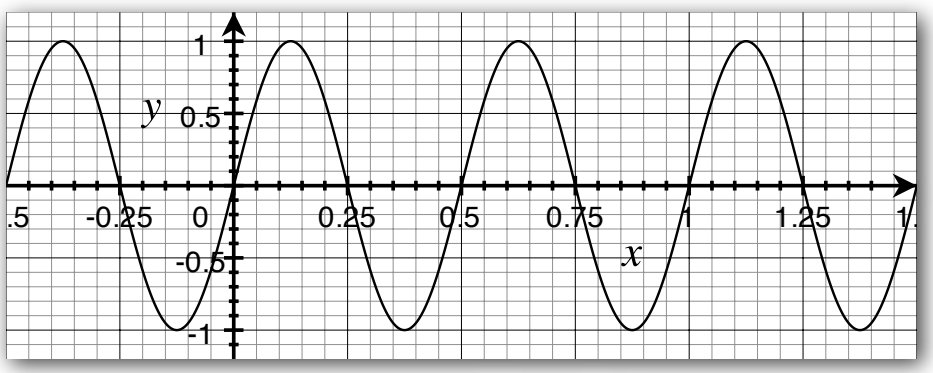
コサイン君，さようなら・・・

$$y_s = s_1\sin(2\pi\, t) + s_2\sin(2\pi\, 2t) + s_3\sin(2\pi\, 3t) + s_4\sin(2\pi\, 4t) + s_5\sin(2\pi\, 5t)$$

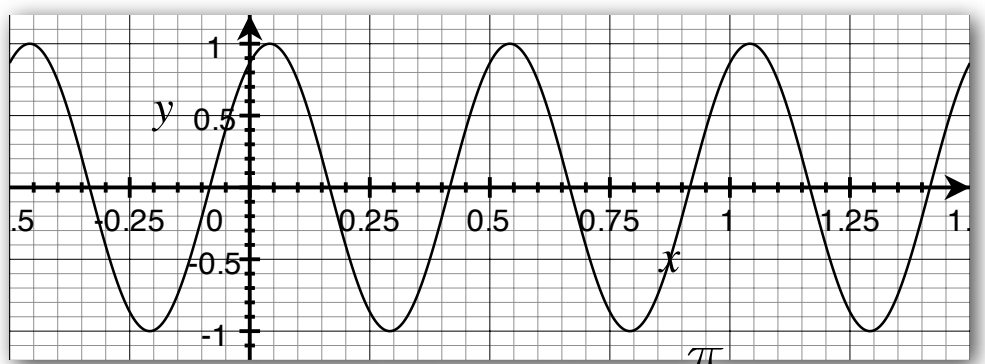
$$y_c = c_1\cos(2\pi\, t) + c_2\cos(2\pi\, 2t) + c_3\cos(2\pi\, 3t) + c_4\cos(2\pi\, 4t) + c_5\cos(2\pi\, 5t)$$

1Hz　2Hz　3Hz　4Hz　5Hz

$$\begin{aligned} y &= y_s + y_c \\ &= S_1\sin(2\pi\, t + \phi_1) + S_2\sin(2\pi\, 2t + \phi_2) + S_3\sin(2\pi\, 3t + \phi_3) + \cdots \end{aligned}$$



$y = \sin(2\pi\, 2t)$



$y = \sin(2\pi\, 2t + \frac{\pi}{3})$

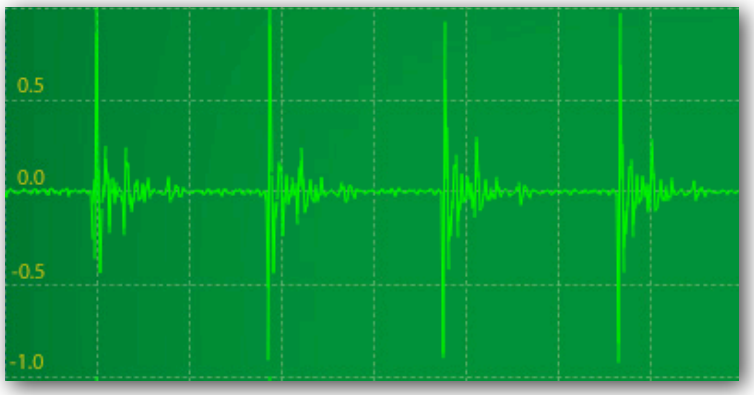
合成波⇄要素波の足し合わせ

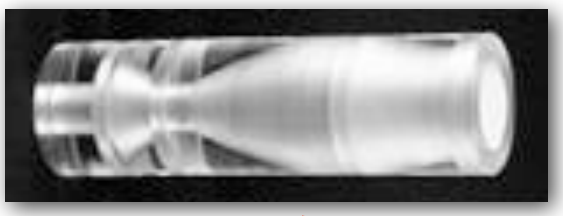
$y \leftrightarrow (S_1, S_2, S_3, \ldots, S_N)$

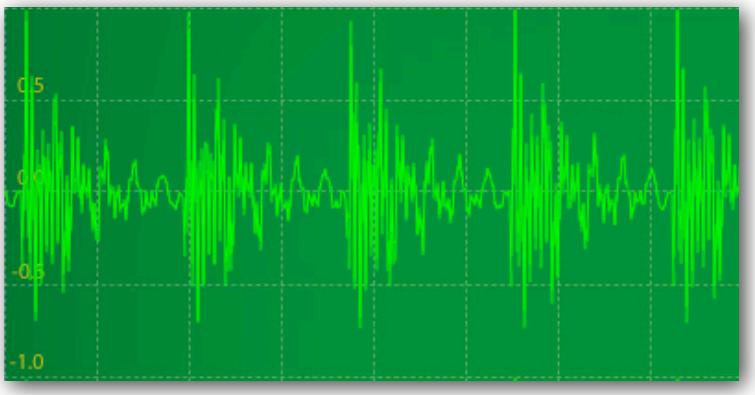
- 合成波を要素波群に分解し，S_i の値を操作して元に戻す。
  - フィルタリング（細かな振動だけ，大まかな振動だけ，を取り出す）
  - イコライザー

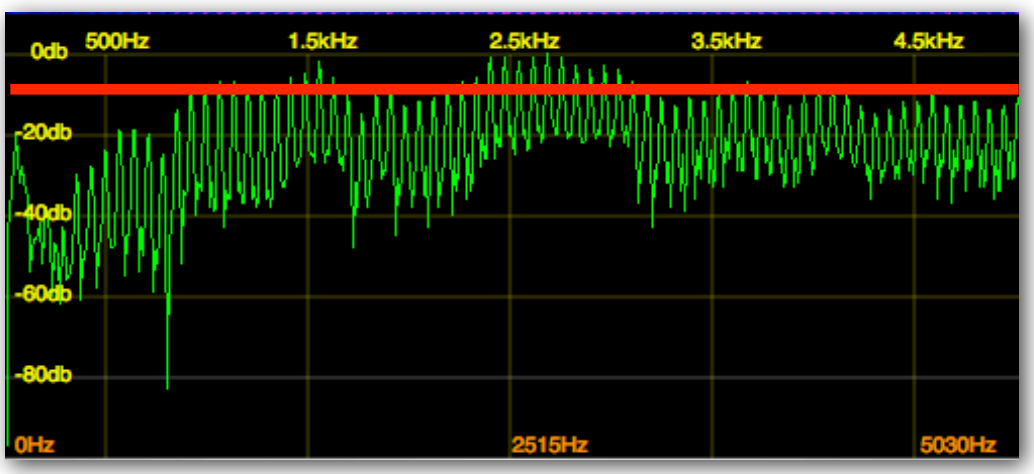
この図がちょっと違って見える？

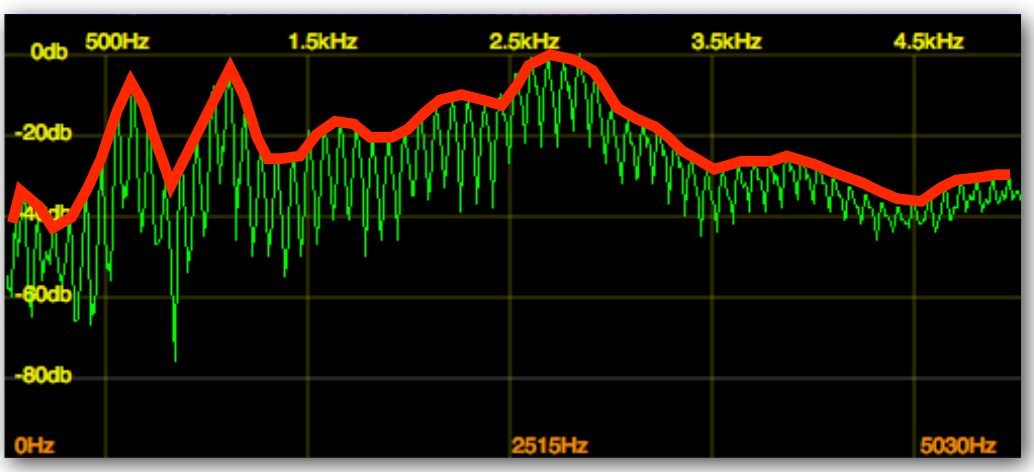
定常波の生成⇄フーリエ解析＆重み操作＆重ね合わせ











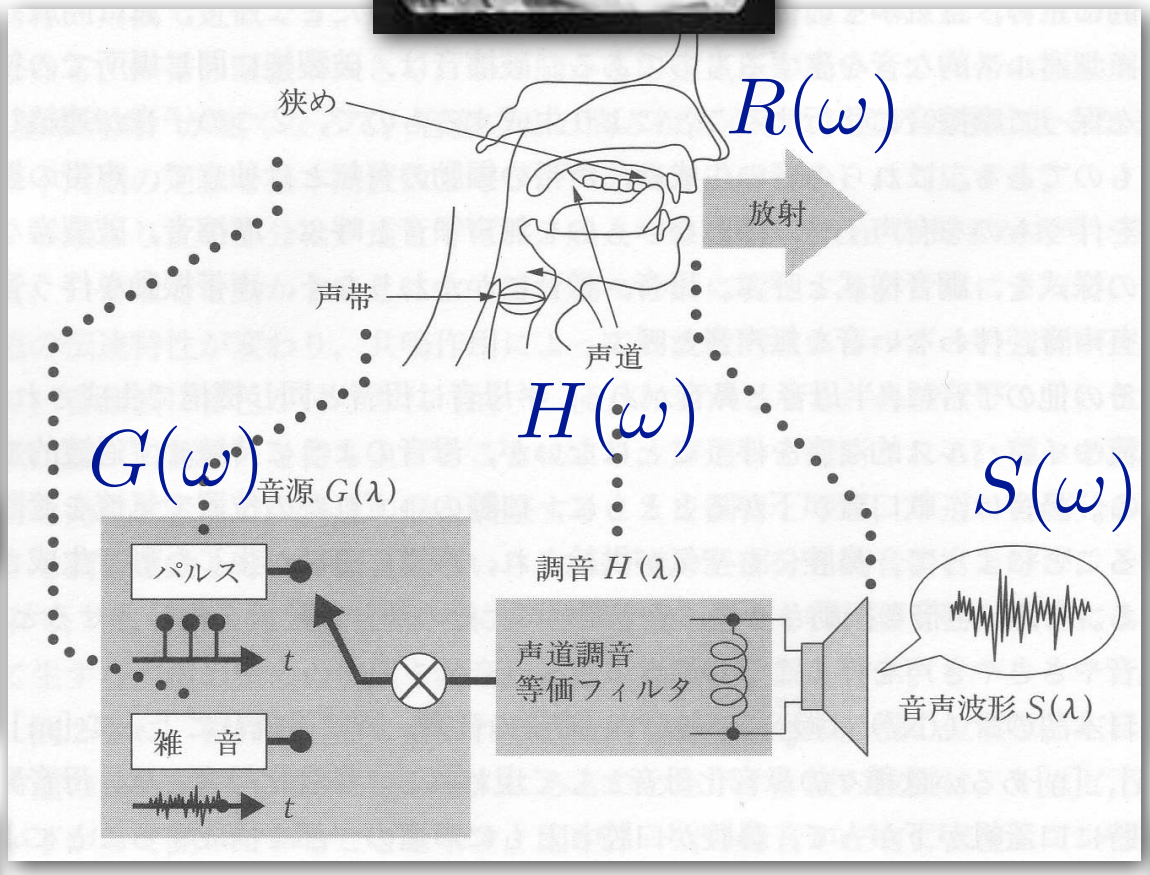
- 音響管の形に基づいて生き残る or 死に絶えるサイン波がある
- スペクトル包絡の特性さえ与えられれば，フーリエ解析で可能
  - 調音器官の形を指定するのか／所望の音響特性を指定するのか

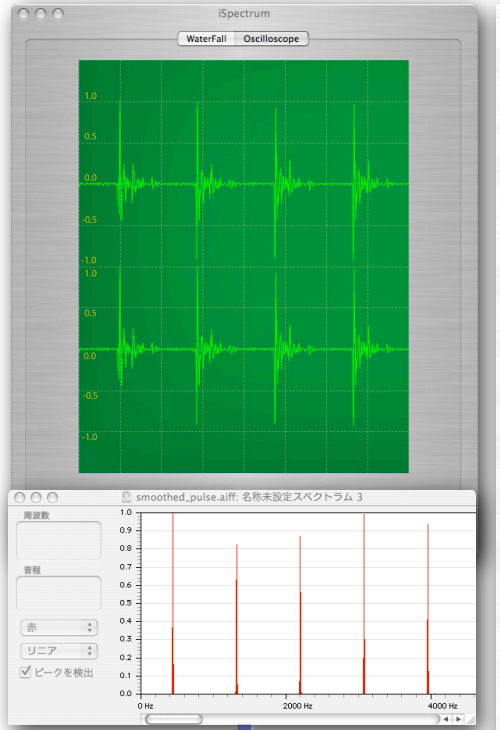
音声の音響分析

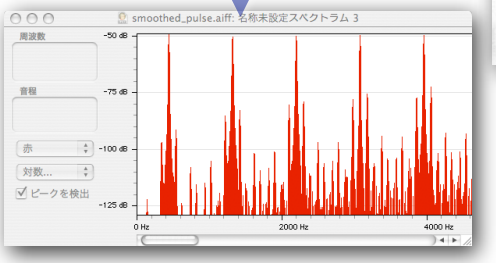
音声（母音）生成のソース・フィルタモデル

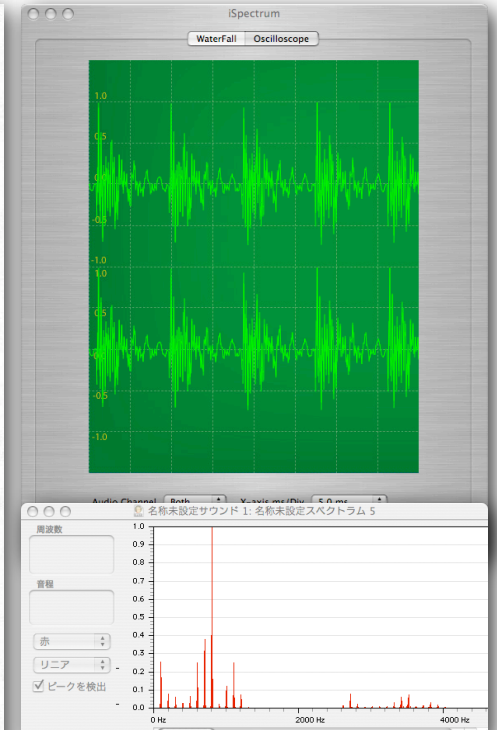
- 音声→音源の特性＋声道フィルタの特性

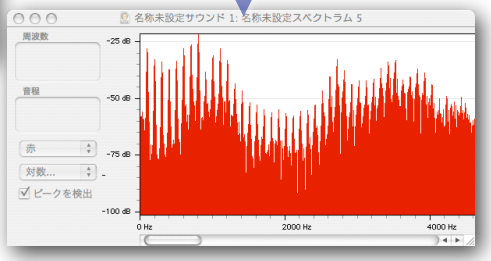
$$y \leftrightarrow (S_1, S_2, S_3, \ldots, S_N)$$











$$S(\omega) = G(\omega)H(\omega)R(\omega)$$

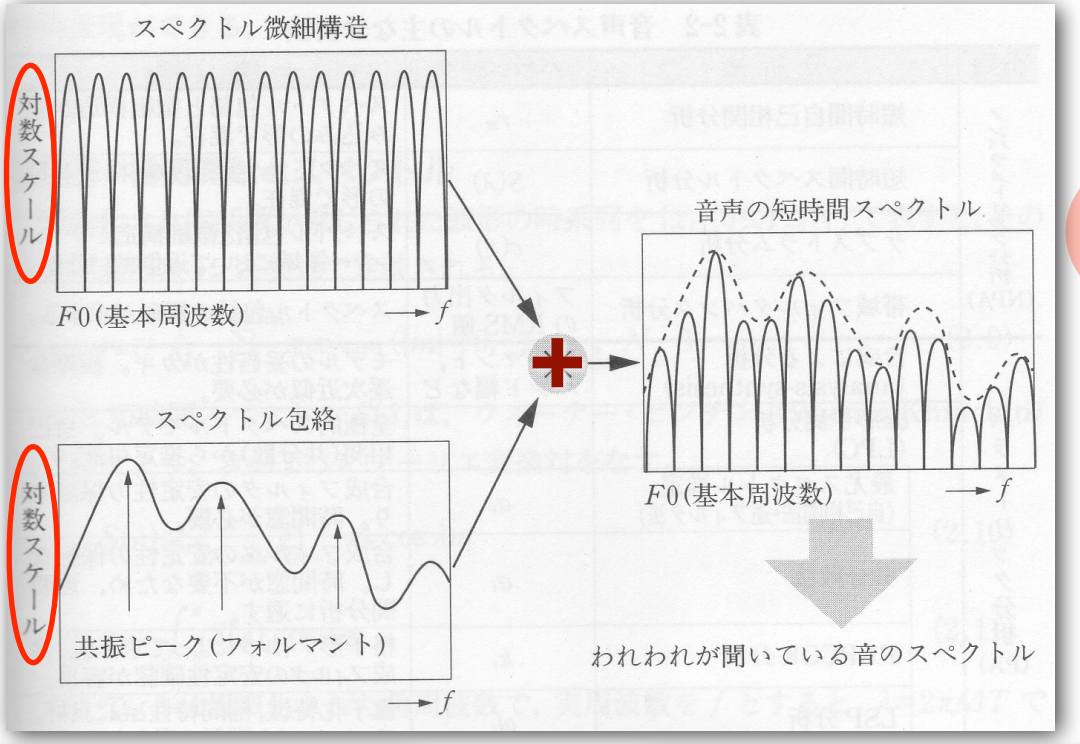
音声の音響分析

音声（母音）生成のソース・フィルタモデル

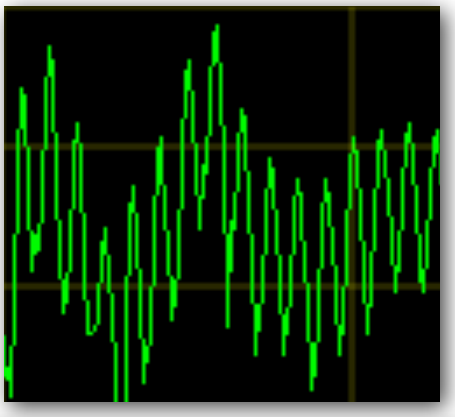
- 対数スペクトル特性においては，音源特性＋声道特性

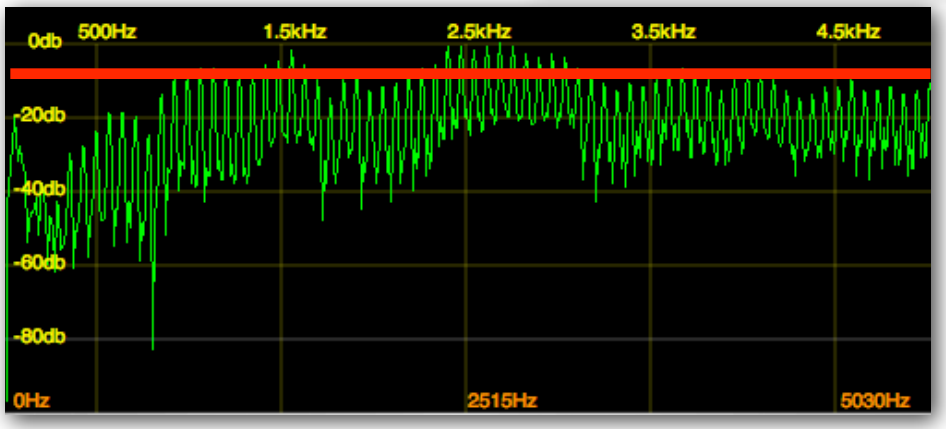
音

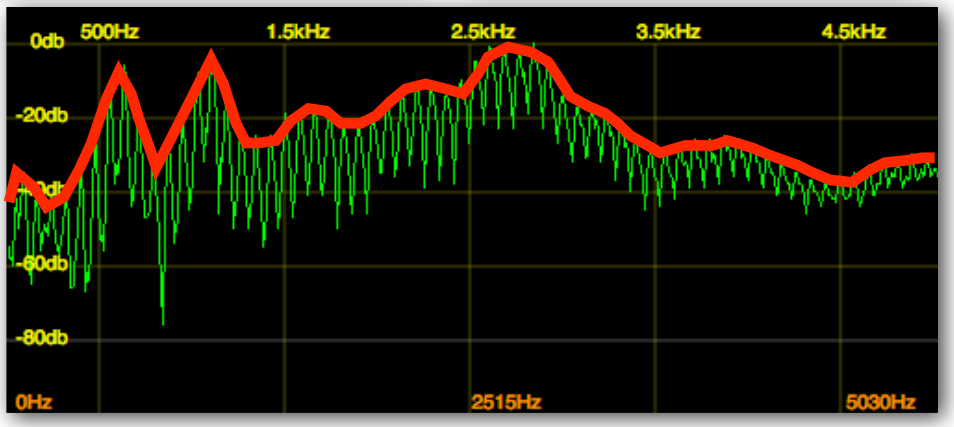
音変形



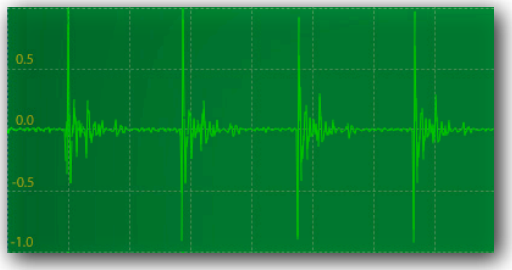
音

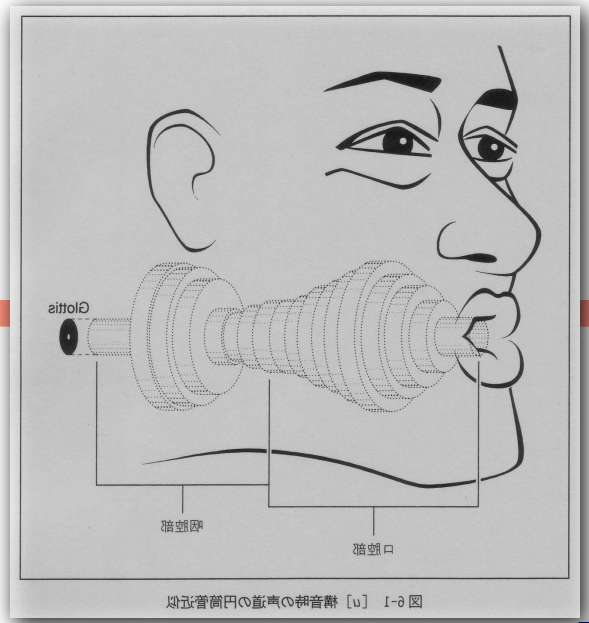


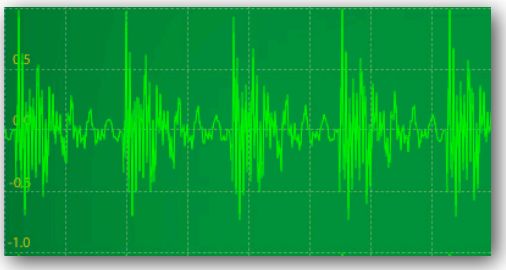


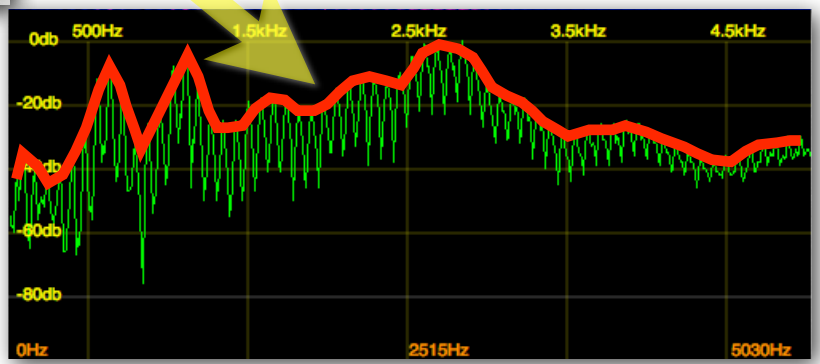


音声波形を使って推定できるもの

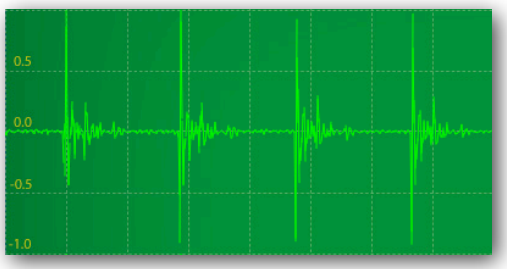


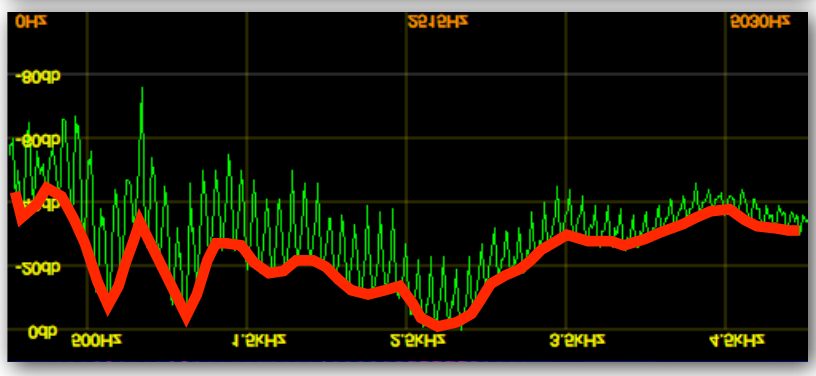




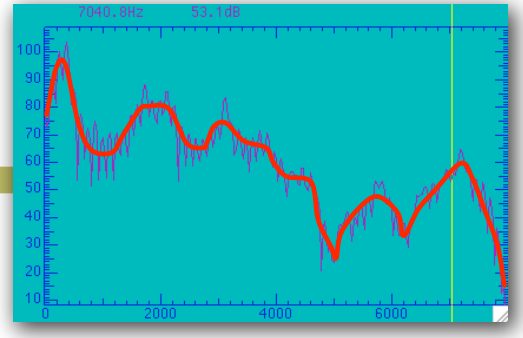


スペクトル包絡特性
は大変重要！





声道特性の抽出技術

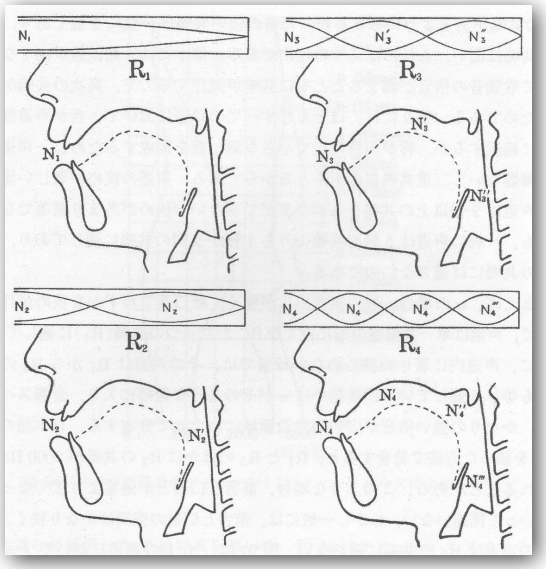


パラメトリックな手法

- モデルを仮定し，観測量がそのモデルから生成されたと解釈
- 既存の観測量が最も確率高く得られるモデルパラメータを推定
  - 最尤推定法
- 長所と短所
  - モデルによる制約がある分，推定にかかる演算量が少ない。
  - モデルが不適切であれば，全てが台無し。
- 線形予測分析（LPC, Linear Predictive Coding）

ノン・パラメトリックな手法

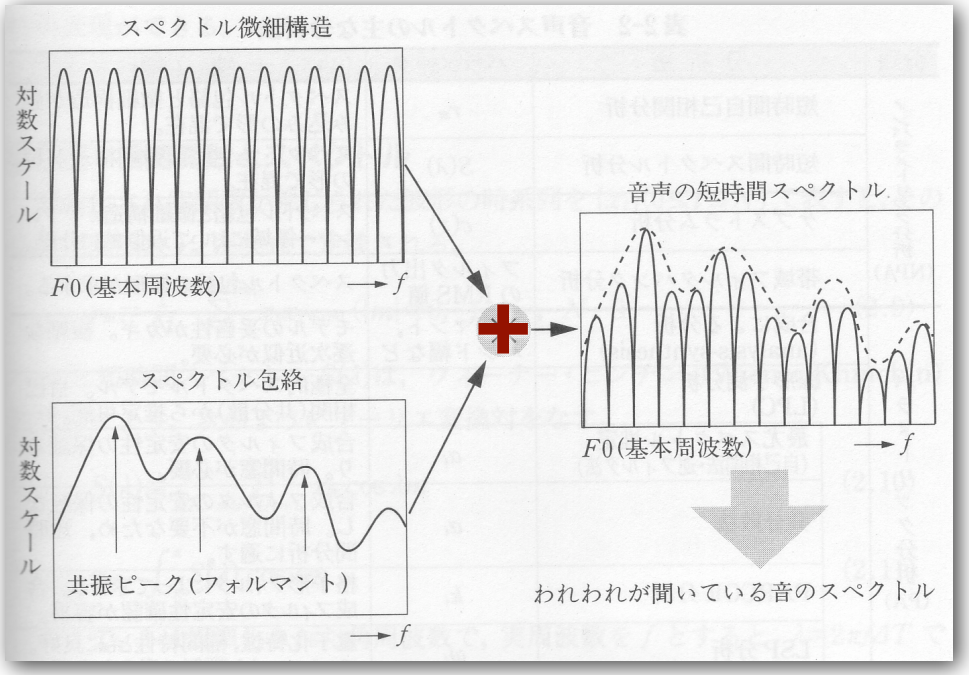
- モデルの仮定無し。あらゆる信号を解析可能
- 長所と短所
  - 自由度高，演算量高
- ケプストラム分析

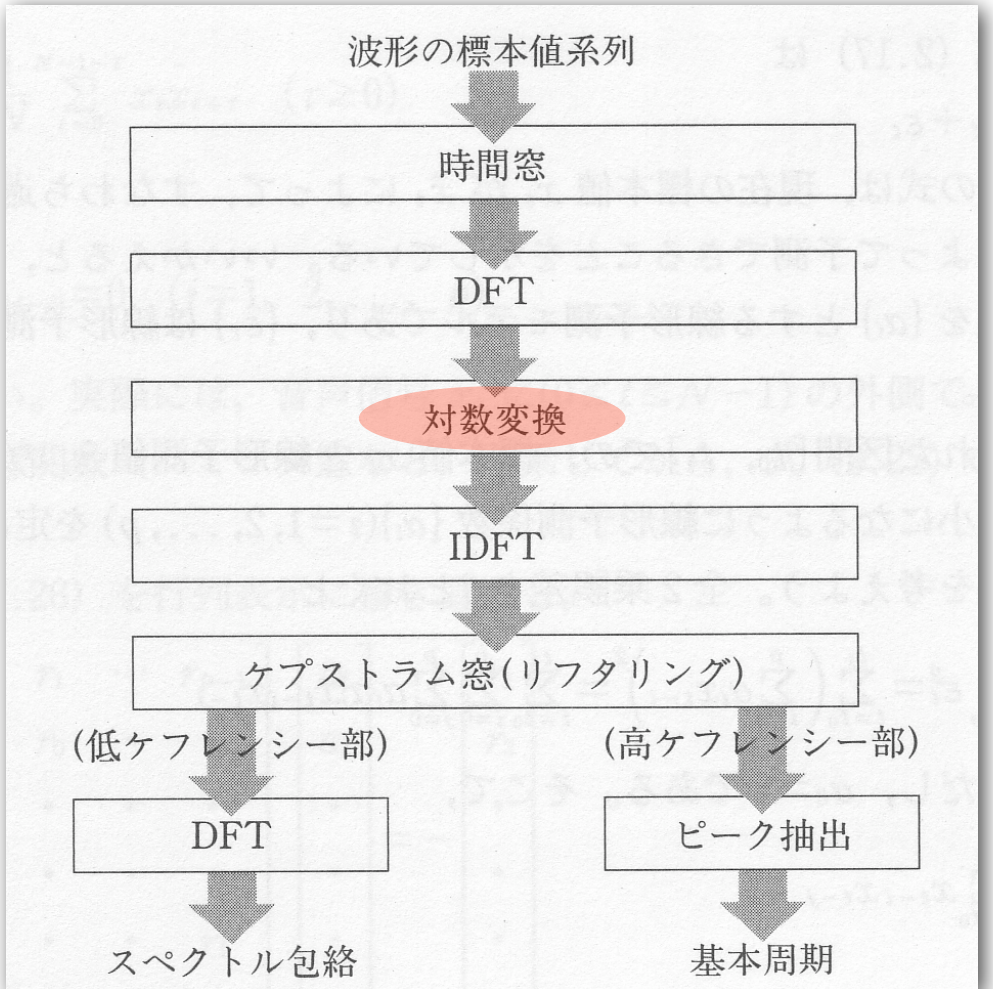


声道特性の抽出技術

ケプストラム法

- フーリエ変換による LPF を用いたスペクトルスムージング
  - スペクトルの大局的な成分と局所的な成分との分割
  - スペクトル包絡と基本周波数の同時推定



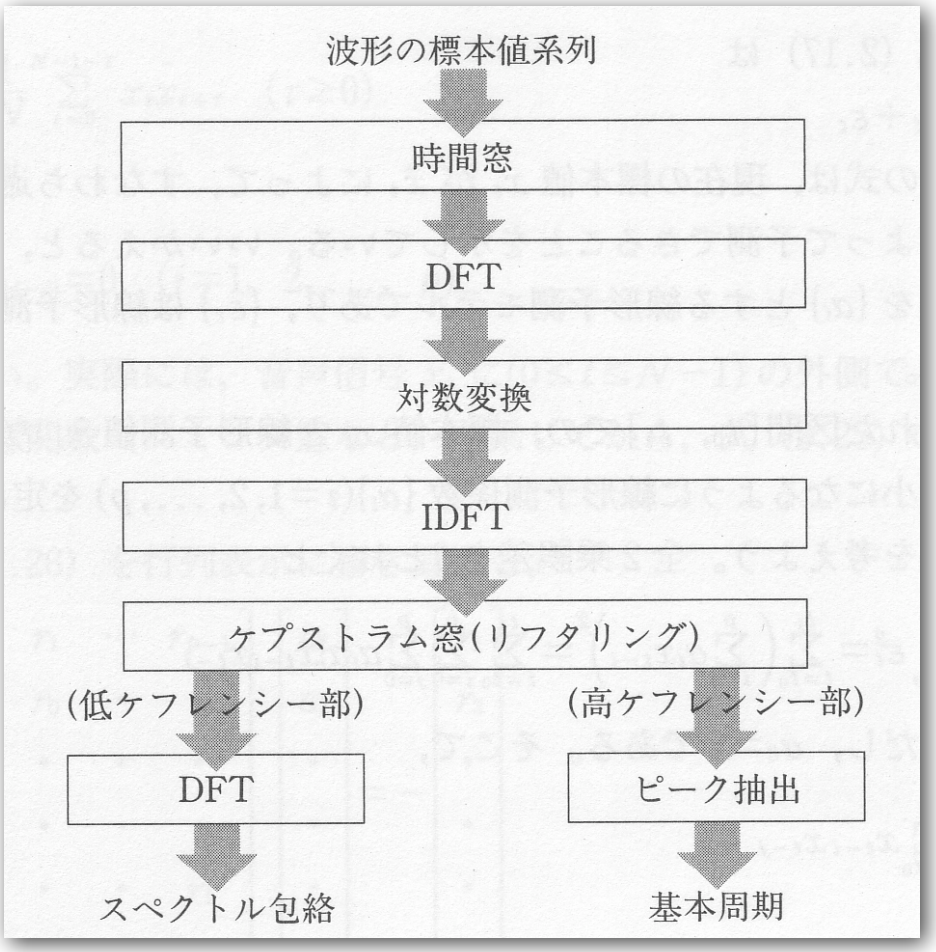


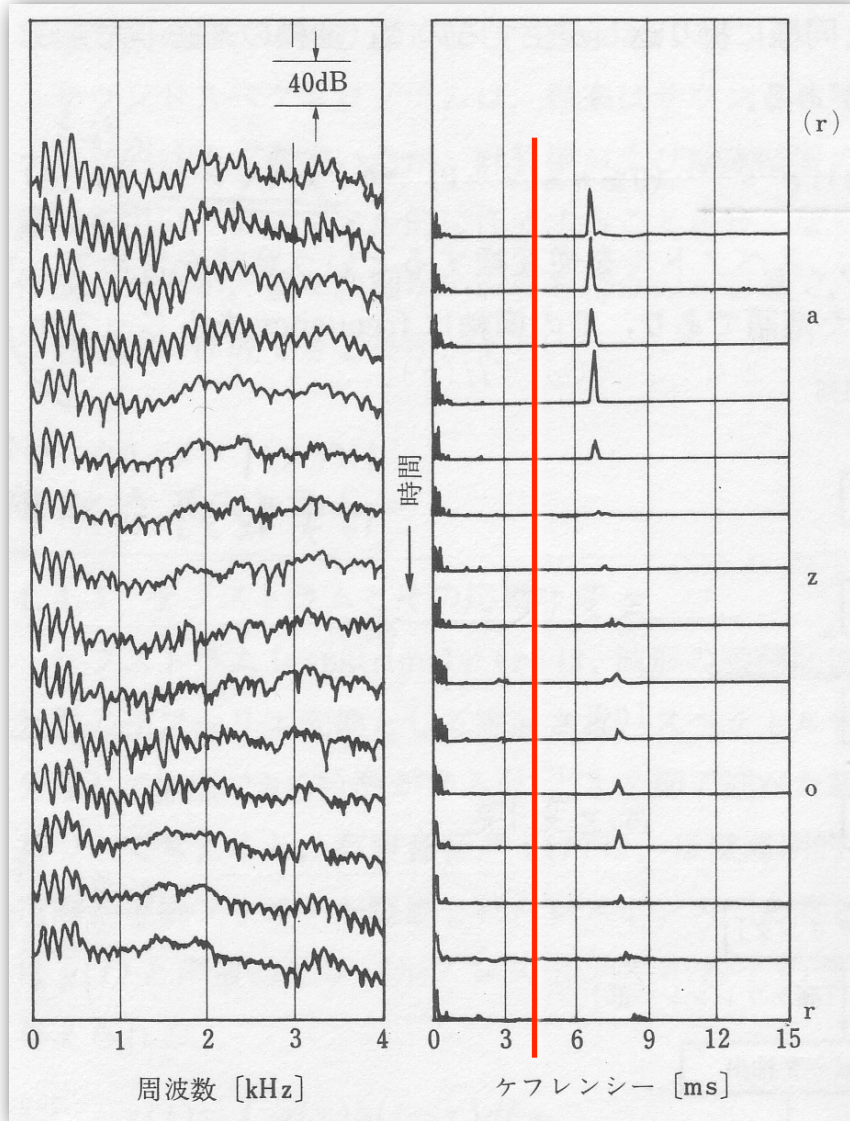
DFT = Discrete Fourier Transform
IDFT = Inverse DFT

声道特性の抽出技術

ケプストラム法

- スペクトル包絡情報と基本周波数情報の分離

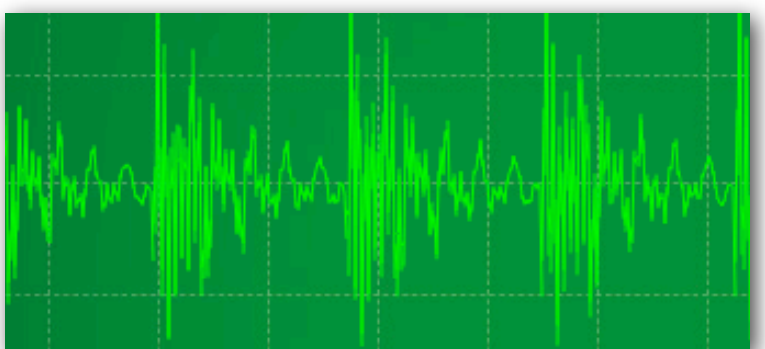


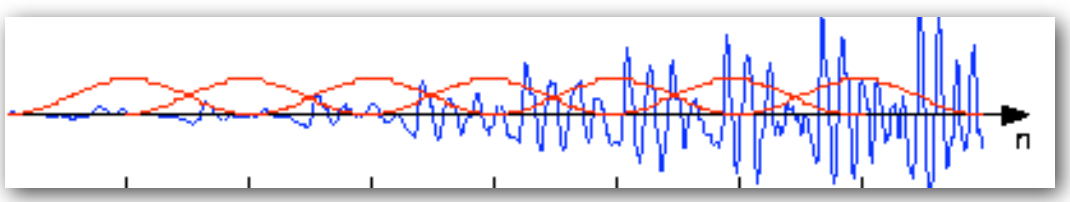


声道特性の抽出技術

線形予測分析（時間軸でのLPC）

- 時刻 t の信号 s_t を過去のサンプルより線形に予測（モデル化）
  - $s_t \leftarrow s'_t = \alpha_1 s_{t-1} + \alpha_2 s_{t-2} + \ldots + \alpha_p s_{t-p}$　（一種のFIRフィルタ）
- 予測された信号と観測された信号間の誤差を最小化
  - $\sum_t |s'_t - s_t|^2$ の最小化
  - ある時間区間（例えば250msec）の間は α_n は不変として，その区間において誤差が最小となるように α_n を求める。（誤差＝残差）
  - 下記の方程式へ帰着（$r_\tau = \frac{1}{N}\sum_{t=0}^{N-1-\tau} s_t s_{t+\tau}$，自己相関関数）



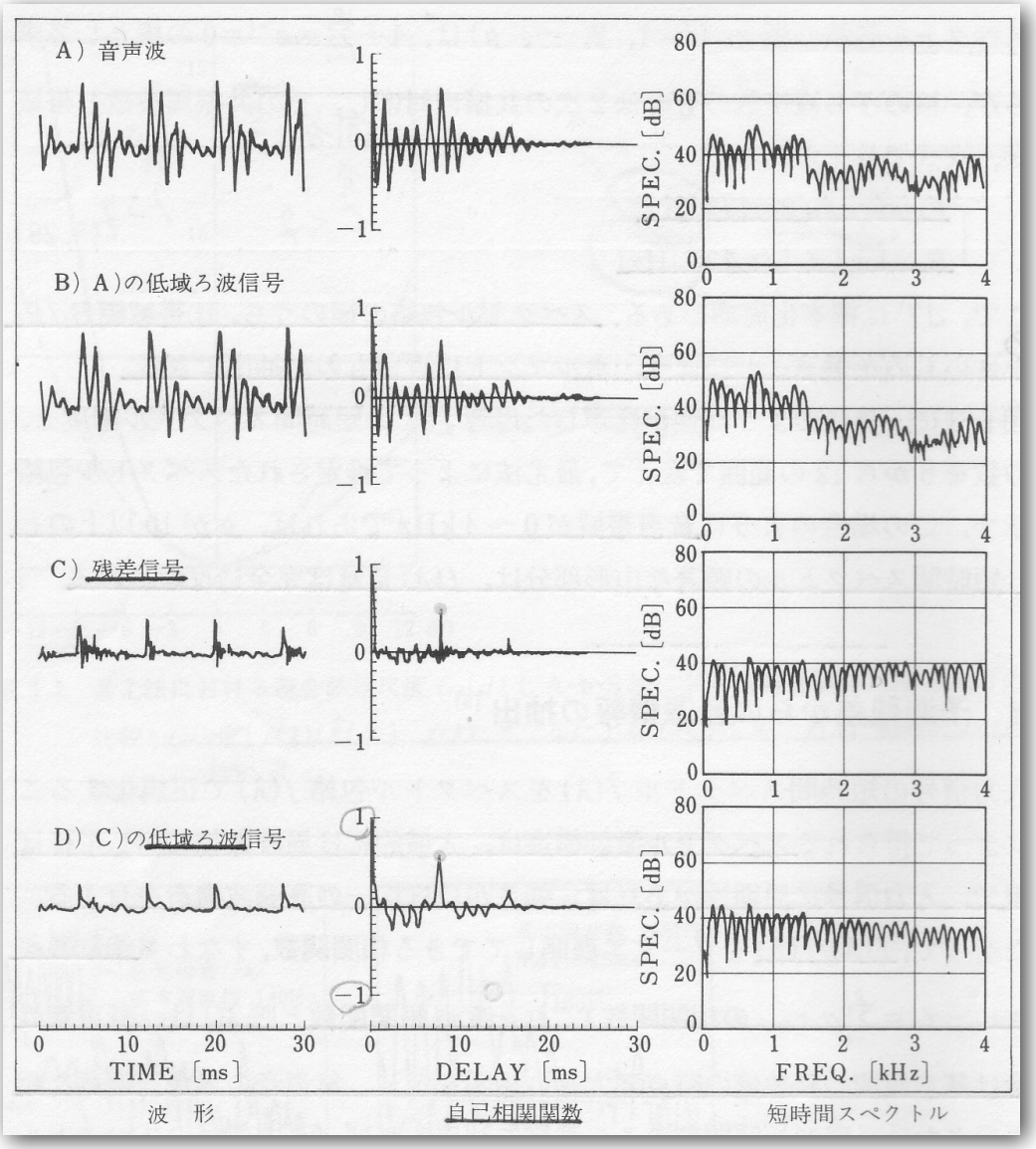


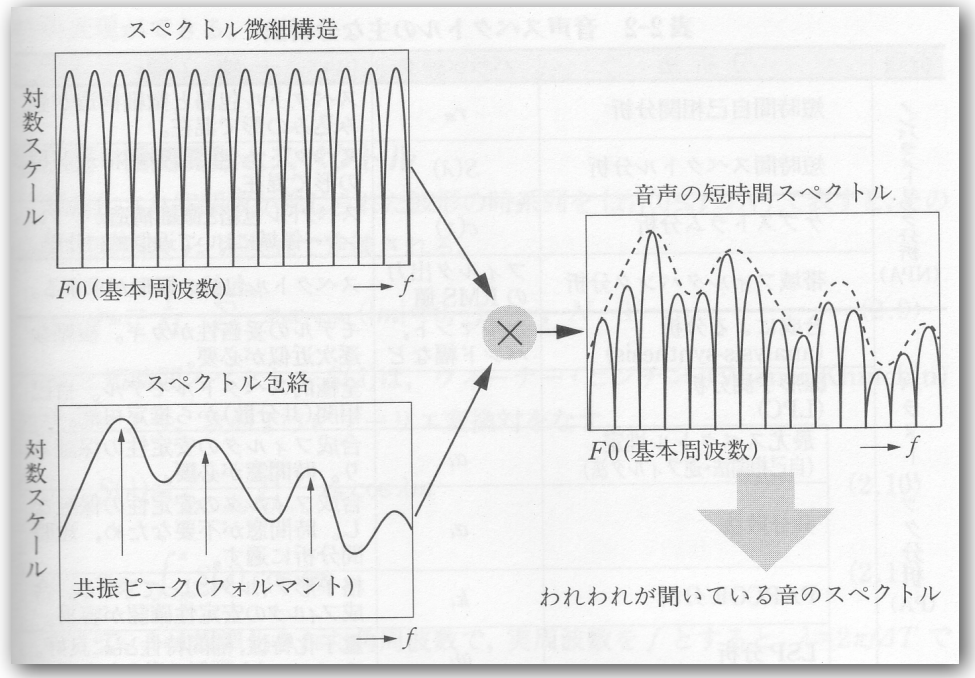
$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & r_1 \\ r_{p-1} & \cdots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix}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

音源特性の抽出技術

時系列信号に対する周期構造の抽出

- ＝自己相関のピークの検出（$r_\tau = \frac{1}{N}\sum_{t=0}^{N-1-\tau} s_t s_{t+\tau}$ ）



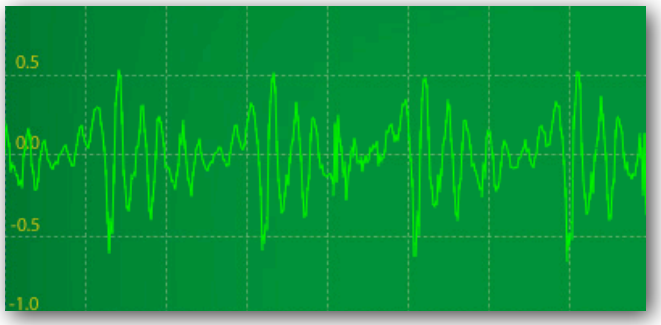


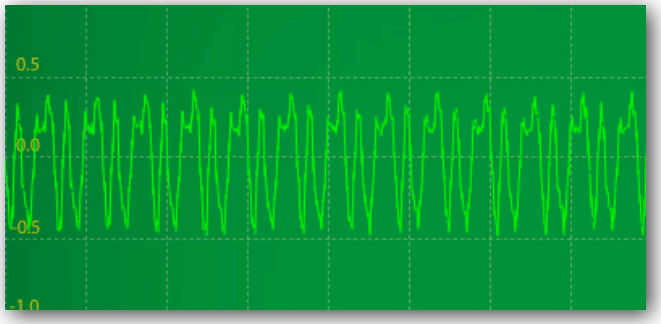
- 音声に対して反転フィルターを通して出力される音＝残差信号

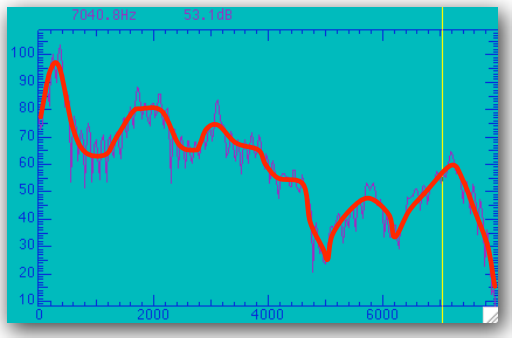
音の心理量とその物理量

音の心理的四要素と対象となる物理量

- 心理量と物理量のマッピング
- 音の高さ
  - 波形における周期の長短＝高さの低高
- 音の長さ
  - 波形を見れば一目瞭然
- 音の強さ
  - 波形振幅に反映される。
- 音色
  - 音楽的には高・長・強以外の要素を音色と定義したりする。
  - 耳が位相に鈍感であることを考慮し，対数スペクトル包絡
  - 各倍音のエネルギーの強さ

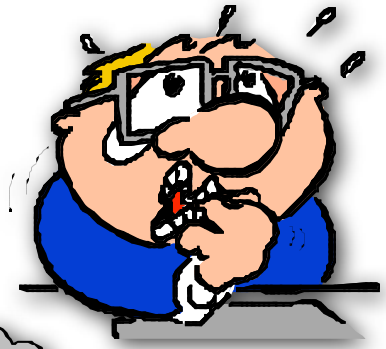






音声によって伝搬される情報

言語・パラ言語・非言語情報



- 語彙（意味）・テキストの情報
  - 制御可能／動的
- 話者の感情・意図・態度
  - 制御可能／動的
- 話者の個人性・性別・年齢・健康状態
  - 制御不可能／静的

分節的特徴と超分節的特徴

- Segmental features & Supra-segmental features
  - segment＝分節音＝所謂音素のこと
  - 音声を細かい言語単位（音素）に分割するために着眼する音響特性
    - スペクトル包絡特性
  - それよりも大きな単位で観測して初めて意味を持つ音響特性
    - 基本周波数パターン，パワーパターン，継続長→韻律的特徴

STRAIGHT

分析（再）合成技術

- 音声→分析→包絡＋音源→包絡’＋音源’→合成（分析$^{-1}$）→音声
  - 音の音色＝スペクトル包絡
  - 音の高さ＝基本周波数
  - 音の強さ＝波形振幅
  - 音の長さ＝（ある音素の）継続時間長
  - これらの特徴量を色々変形して，再合成（分析と逆の処理）する

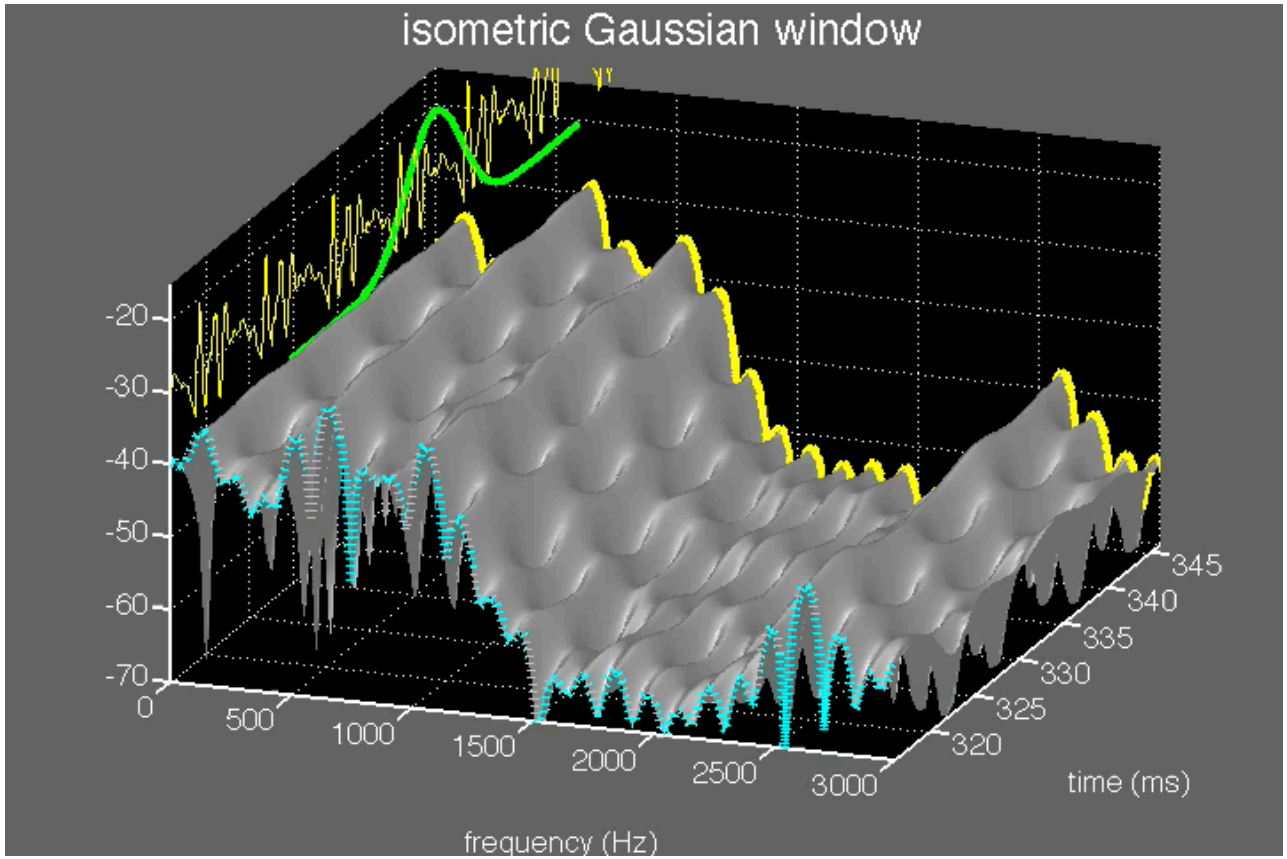
STRAIGHT

- **S**peech **T**ransformation and **R**epresentation based on **A**daptive **I**nterpolation of wei**GHT**ed spectrogram
  - 適応的に内挿されたスペクトル表現
  - 河原英紀@和歌山大学システム工学部による開発
    - http://www.wakayama-u.ac.jp/~kawahara/

STRAIGHT

つまり，こうなります（SFT = 短時間フーリエ変換）

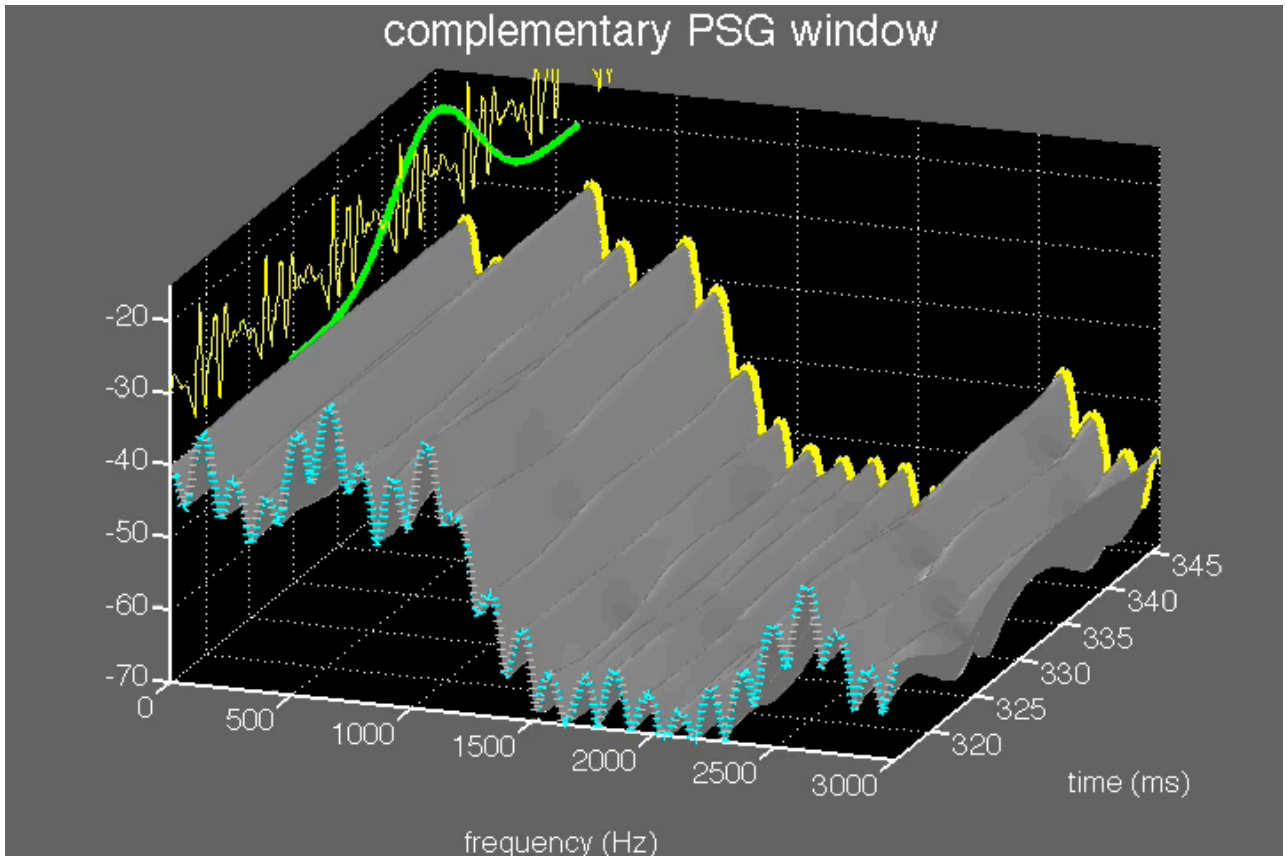
SFT-based spectrogram



STRAIGHT

ピッチに同期して窓を動かしてあげると・・

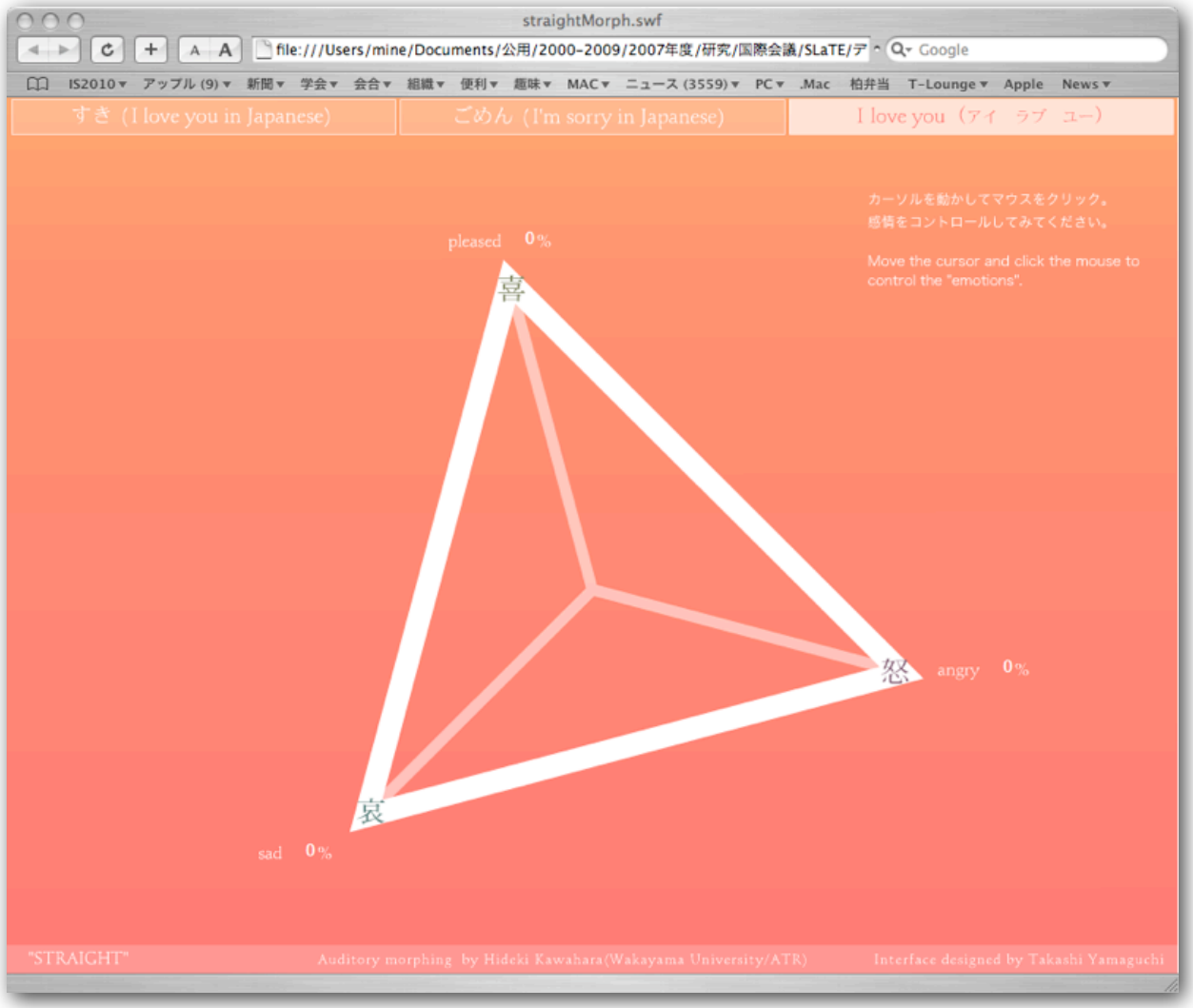
Complementary pitch-synchronous Gaussian window removes the repetitive structure in the time domain



音声モーフィング（変形）

感情を変えてみる。平静／怒／悲／喜

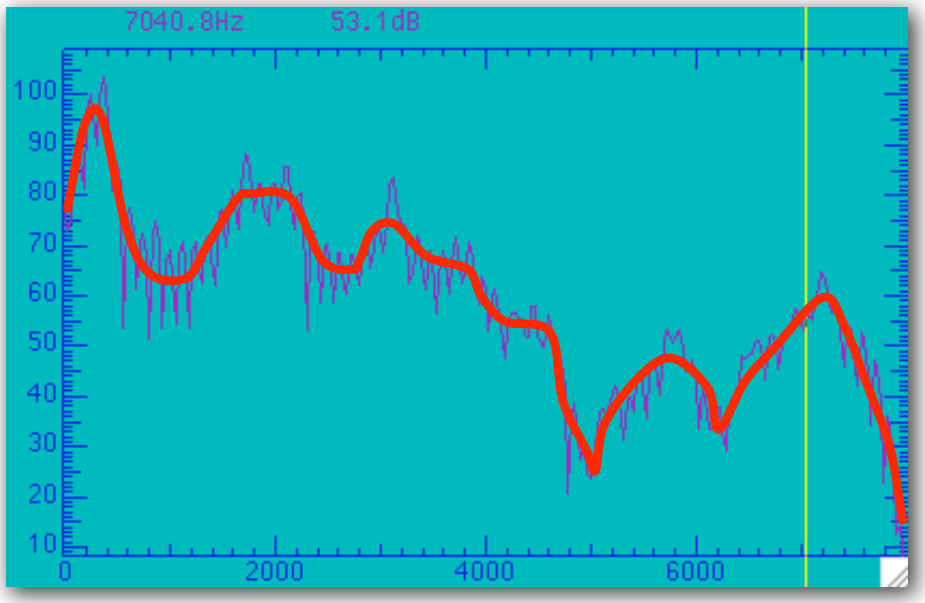
- 音色・高さ・長さ・時間長のどこがどう変わっているのか？
- 操作対象の物理量に思いを馳せながら聞きましょう。



音声モーフィング（変形）

性別・年齢を変えてみる２。

- 読み上げ音声に対する分析再合成
  - 周波数軸：x 倍　　x > 1＝フォルマント上昇
  - 原音声
  - 再合成音声（変形無し）
  - 再合成音声（Fo：2倍，周波数軸：1.25倍）
  - 再合成音声（Fo：3倍，周波数軸：1.44倍）
  - 再合成音声（Fo：0.5倍，周波数軸：0.8倍）



音声モーフィング（変形）

普通の声を歌声に変えてみる。

- 歌声音声に対する分析再合成
  - 女声の持続母音/a/（440Hz，原音声）
    - 再合成音声（変換無し）
    - 再合成音声（ソプラノ化）
    - 再合成音声（アルト化）
  - 男声による旋律/a/（原音声）
    - 再合成音声（変換無し，バリトンとして使用）
    - 再合成音声（テノール化）
    - 再合成音声（バス化）
  - 出来上がった５声による演奏（元は２音声）

音声モーフィング（変形）

何から何に変わっているのか，当てて下さい。

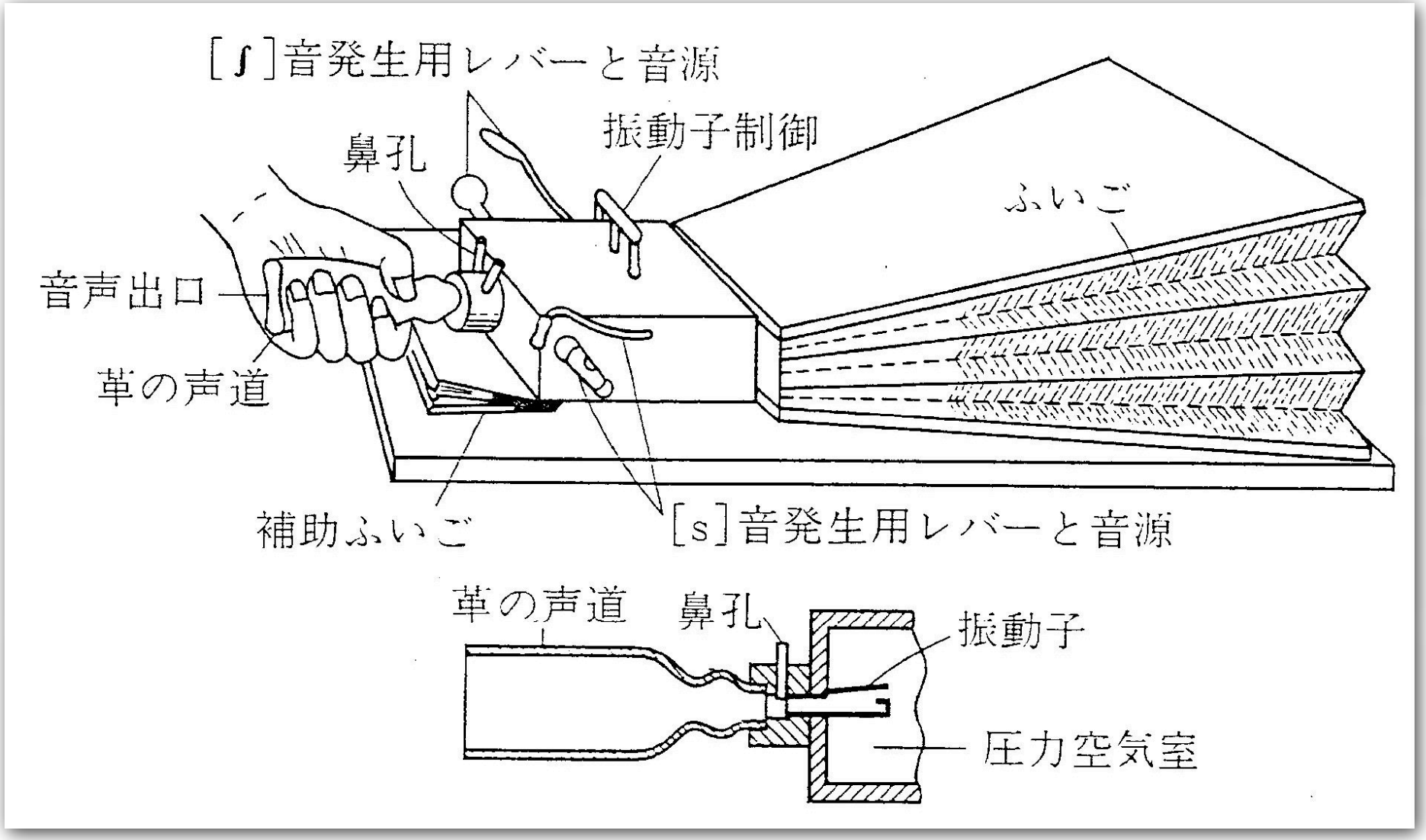
A　B　C

本日のメニュー

- **音声合成の原理と波形編集型音声合成**
  - 言語処理／音韻処理／波形処理
  - 音韻処理といっても結構面倒・・・
- **音声認識セミナーより**
  - 音声特徴の分析と照合
  - 隠れマルコフモデル（HMM）
  - サブワードモデルを用いた音声認識
  - 認識文法を用いた連続音声認識
  - 日本語ディクテーションにおける基礎技術
  - ビデオ鑑賞
- **宿題**
- **まとめ**

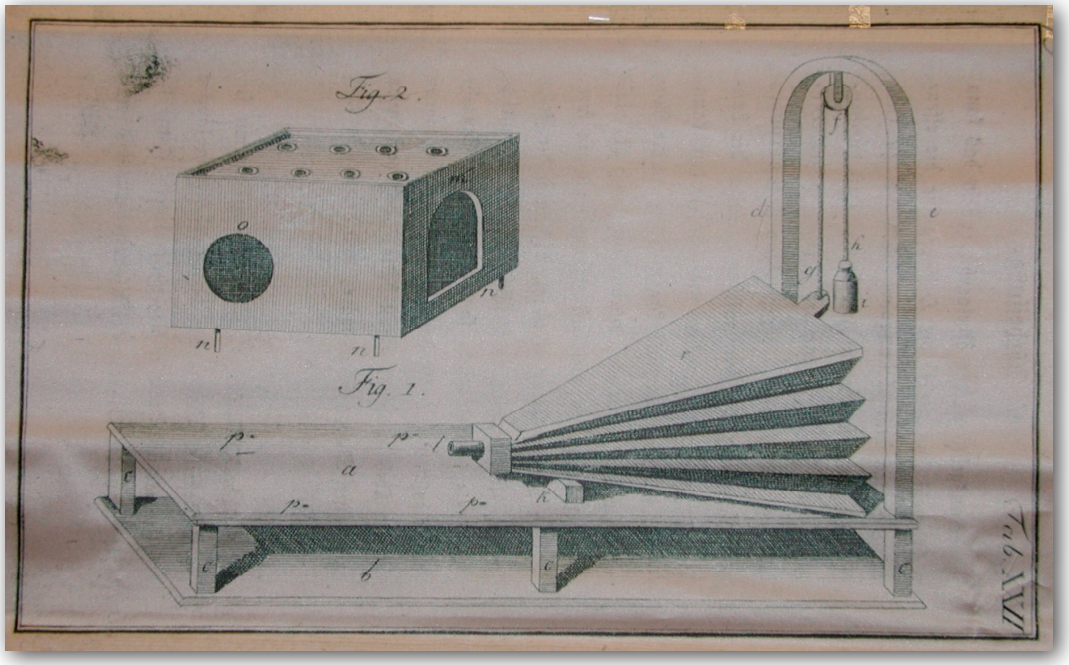
世界最古の音声合成器

von Kempelenの機械式音声合成器



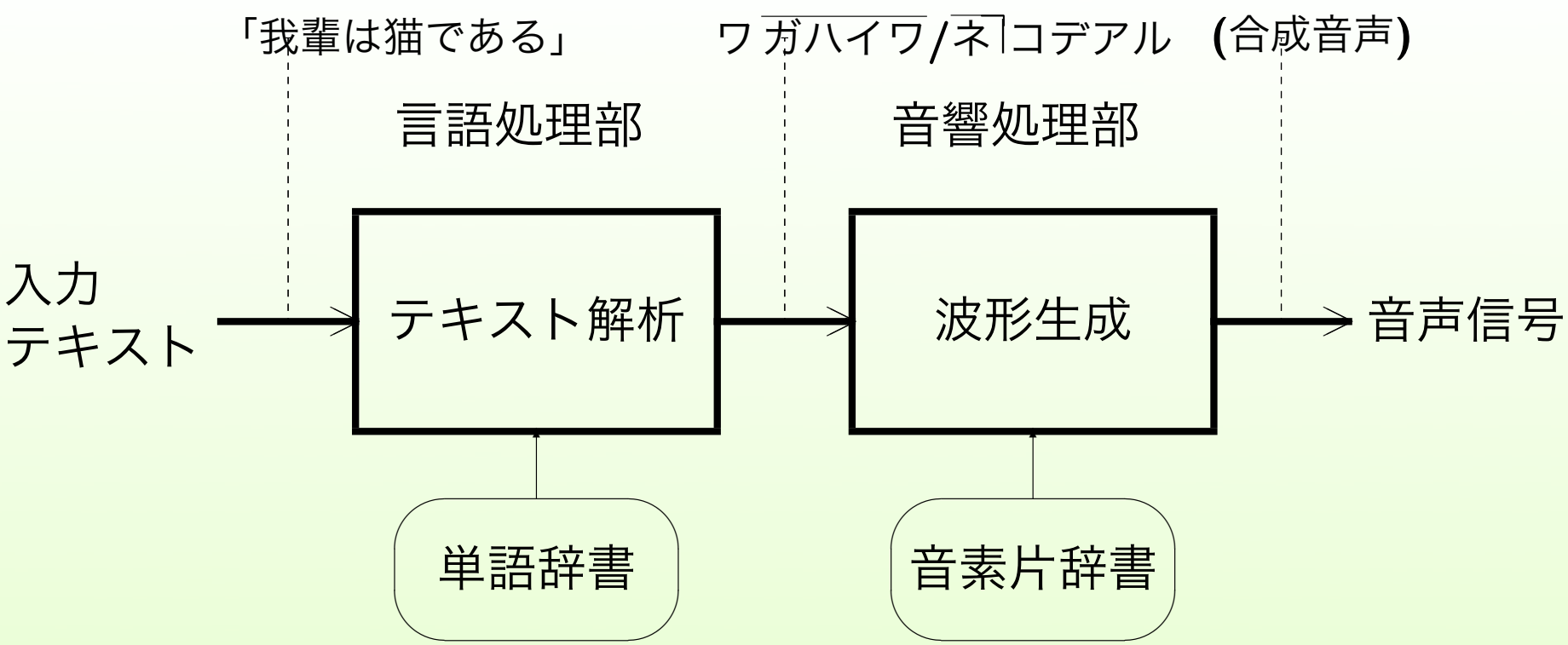
世界最古の音声合成器

- von Kempelenの機械式音声合成器



テキスト音声合成

まずは概要から



テキスト音声合成

テキスト音声合成とは

- 任意の漢字仮名混じり文を入力として音声波形を生成する技術

何が必要？

- 言語処理（文解析）
  - 形態素解析器によるテキストの単語への区分
    - 品詞，統語情報，係り受けの推定（意味，談話情報は難）
- 音韻処理
  - 文解析結果を実際の発音の表記へと変換
    - 最終的に「読み」の情報と「アクセント」の情報を推定
- 音響処理（波形生成）
  - 発音表記に従って，音声波形を生成する。
    - モデルベースの方法，波形素片をペタペタ貼付ける方法

以下，言語処理が終わった後の過程を考えます

音韻処理

- **音韻性に関する処理（分節的特徴）**
  - 音素表記の生成
    - 音訓，長音化，促音化，連濁
  - 単音表記の生成（異音化処理）
    - 無声化，鼻音化，撥音
- **韻律性に関する処理（超分節的・韻律的特徴）**
  - 単語レベルの処理
    - アクセント型，複合語（接頭語，接尾語）
  - 句レベルの処理
    - アクセント結合
  - 文レベルの処理
    - 強調，フレージング
- **単語分割語にどれだけの処理をしているのか確認せよ**

音素表記の生成

音読み・訓読み

- 文法情報：工夫する（サ変動詞），強力な（形容動詞）
- 意味素性：十分以内（数詞＋以内），関東平野（地名＋平野）
- 係り受け：講演を行なった　／　方程式の根

助詞

- は→ワ，へ→エ

長音化

- 消耗（オに続くウ），映画（エに続くイ），大阪（同一母音）

促音化（特に数詞＋助数詞）

- 一巻，一本

連濁（カ，サ，タ，ハ行の語頭清音の濁音化）

- 目覚まし時計，遅咲き，　　ぺん先

数詞＋助数詞

数詞の読み

- 03-5841-6662，　123,456円

促音化

- 一本→イッポン，　1cm→イッセンチ

濁音化（連濁）

- 三本→サンボン，　三階→サンガイ

特例

- 一日（ツイタチ），　二日（フツカ）

単音表記の生成（異音化処理）

- **母音無声化（無声子音に挟まれたイとウ）**
  - アシカ，エンピツ，スキヤキ，シチ
- **鼻音化**
  - カブシキガイシャ
- **撥音**
  - ネンバンガン
  - p, b, m の前（カンパン），t, d, n の前（カンタン）
  - k, g, ng の前（カンカク），語尾（カンザン）

記号・省略文字・未登録語

記号が入力されても某の音にしなければならない

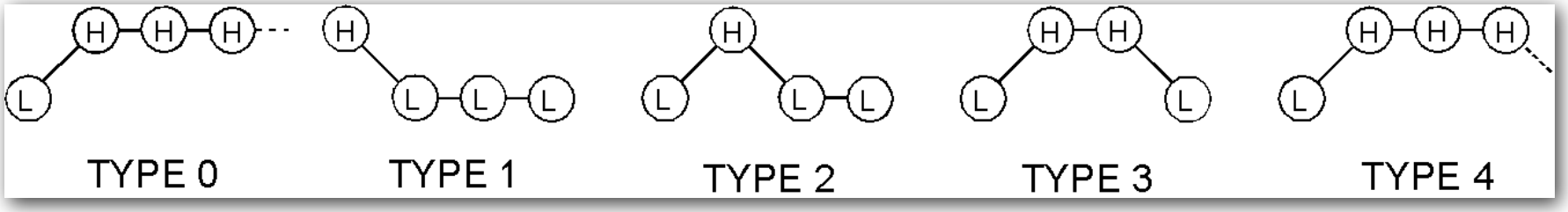
- %, m, 1/s, などなど
- HMM, ICSLP, ICASSP, IT などなど
- 未知語
  - 氏名，地名など

- 最終的には漢字毎のデフォルトの読み，デフォルトの単語アクセントでごまかすことになる。

単語アクセントとアクセント結合

nモーラ単語でn+1個のアクセント型が可能

- 基本周波数の落ちの直前のモーラ＝アクセント核



トモダチ　フジサン　イロガミ　カガリビ　オトート

アクセント結合

- アクセント核の移動

  アル┐ク (歩く) ＋ マス → アルキマ┐ス

- アクセント核の生起

  ケイタイ (携帯) ＋ デンワ (電話) → ケイタイデ┐ンワ

- アクセント核の消失

  ケ┐イザイ (経済) ＋ テキ (的) → ケイザイテキ

単語アクセントとアクセント結合

アクセント結合の規則化

- N_1 モーラ M_1 型アクセント単語と
- N_2 モーラ $\tilde{M}_2$ 型結合アクセント価をもつ後続語が接続
- N_c モーラ M_c 型アクセントとなる
  - アクセント価：連接後に伴い現れる核の位置を連接境界からのモーラ数で表現したもの。連接位置に核（落ち）があれば，０価となる。

付属語アクセント結合規則

- あそぶ　＋　ます　→　あそびます

複合名詞アクセント結合規則

- とうきょう　＋　だいがく　→　とうきょうだいがく

接頭辞アクセント結合規則

- ぜん　＋　せかい　→　ぜんせかい

単語アクセントとアクセント結合

(N_1 モーラ M_1 型+N_2 モーラ $\widetilde{M}_2$ 価→ N_c モーラ M_c 型)

(a) 付属語アクセント結合規則

結合様式	文節のアクセント型 M_c	
	$M_1 = 0$	$M_1 \neq 0$
(F1) 従属型	M_1	
(F2) 不完全支配型	$N_1 + \widetilde{M}_2$	M_1
(F3) 融合型	M_1	$N_1 + \widetilde{M}_2$
(F4) 支配型	$N_1 + \widetilde{M}_2$	
(F5) 平板化型	0	

(b) 複合名詞アクセント結合規則

結合様式	後続名詞の性質	複合名詞 M_c
(C1) 保存型	$N_2 \geqq 2, M_2 \neq 0, {N_2}^\dagger$	$N_1 + M_2$
(C2) 生起型	$N_2 \geqq 2, M_2 = 0, {N_2}^\dagger$	$N_1 + 1$
(C3) 標準型	$N_2 \leqq 2$	N_1
(C4) 平板型	$N_2 \leqq 2$	0

(c) 接頭辞アクセント結合規則

結合様式	文節のアクセント型 M_c	
	$M_2 = 0, {N_2}^\dagger$	$M_2 \neq 0, {N_2}^\dagger$
(P1) 一体化型	0	$N_1 + M_2$
(P2) 自立語結合型	$N_1 + 1$	$N_1 + M_2$
(P3) 分離型	M_1	M_1 (and $N_1 + M_2$)
(P4) 混合型	$N_1 + 1$ (or)M_1	M_1 (and/or) $N_1 + M_2$

単語アクセントとアクセント結合

巡回的な規則適用

- おきる＋られる＋そーな＋ので
- おきられる＋そーな＋ので
- おきられそーな＋ので
- おきられそーなので

その他の考慮

- 音節内核移動規則
  - 撥音，促音，長母音，重母音などのモーラにアクセント核がくると，アクセント核は原則として1モーラ前にずれる。
- 無声化に伴う移動規則
  - 無声化した母音にアクセント核が来ると，アクセント核は原則として1モーラ前にずれる。

テキスト音声合成

テキスト音声合成とは

- 任意の漢字仮名混じり文を入力として音声波形を生成する技術

何が必要？

- 言語処理（文解析）
  - 形態素解析器によるテキストの単語への区分
    - 品詞，統語情報，係り受けの推定（意味，談話情報は難）
- 音韻処理
  - 文解析結果を実際の発音の表記へと変換
    - 最終的に「読み」の情報と「アクセント」の情報を推定
- 音響処理（波形生成）
  - 発音表記に従って，音声波形を生成する。
    - モデルベースの方法，波形素片をペタペタ貼付ける方法

以下，言語処理が終わった後の過程を考えます

音韻処理

- **音韻性に関する処理（分節的特徴）**
  - 音素表記の生成
    - 音訓，長音化，促音化，連濁
  - 単音表記の生成（異音化処理）
    - 無声化，鼻音化，撥音
- **韻律性に関する処理（超分節的・韻律的特徴）**
  - 単語レベルの処理
    - アクセント型，複合語（接頭語，接尾語）
  - 句レベルの処理
    - アクセント結合
  - 文レベルの処理
    - 強調，フレージング
- **単語分割語にどれだけの処理をしているのか確認せよ**

音響処理（波形生成）

波形編集方式

- 時間領域での音声素片加工（TD-PSOLAなど）
- 周波数領域での音声素片加工（FD-PSOLAなど）
  - ノン・パラメトリックな分析が有効

分析合成方式

- 線形予測分析法，ケプストラム分析法
- HMMに基づく合成法

ターミナルアナログ方式（フォルマント合成）

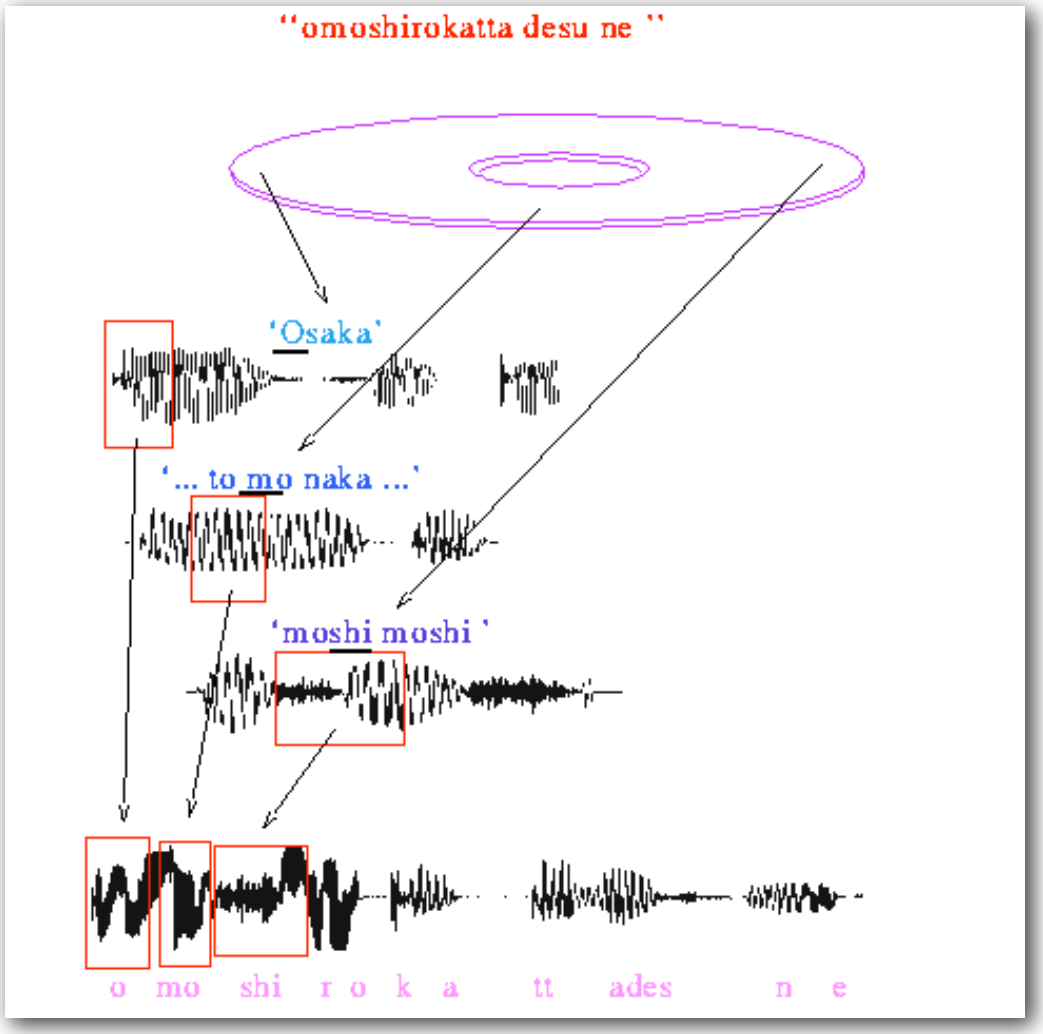
- 周波数領域での音声生成過程の模擬

声道アナログ方式

- 調音器官（メカニクス）での音声生成過程の模擬

波形編集方式

- 音声DBをある単位に基づいて素片DB化
- 与えられたテキストに対して最適な素片を求めて連結
- 必要な波形処理を施して滑らかに接続



波形編集方式

- **音声DBをある単位に基づいて素片DB化**
  - 音素，音節，半音節，どれを使う？
- **与えられたテキストに対して最適な素片を求めて連結**
  - 複数の素片候補のどれとどれを選び，連結するのか？
- **必要な波形処理を施して滑らかに接続**
  - 例えば欲しい高さの音がなかった場合どうする？

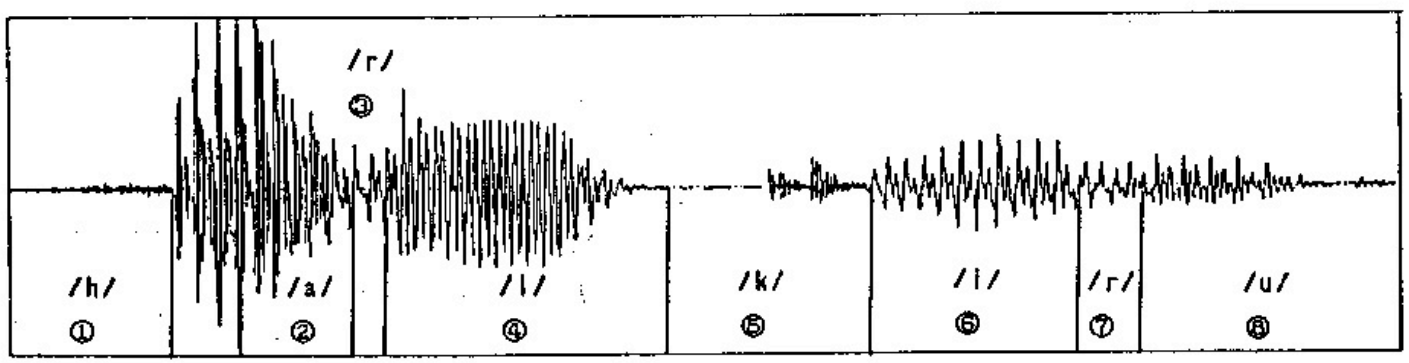
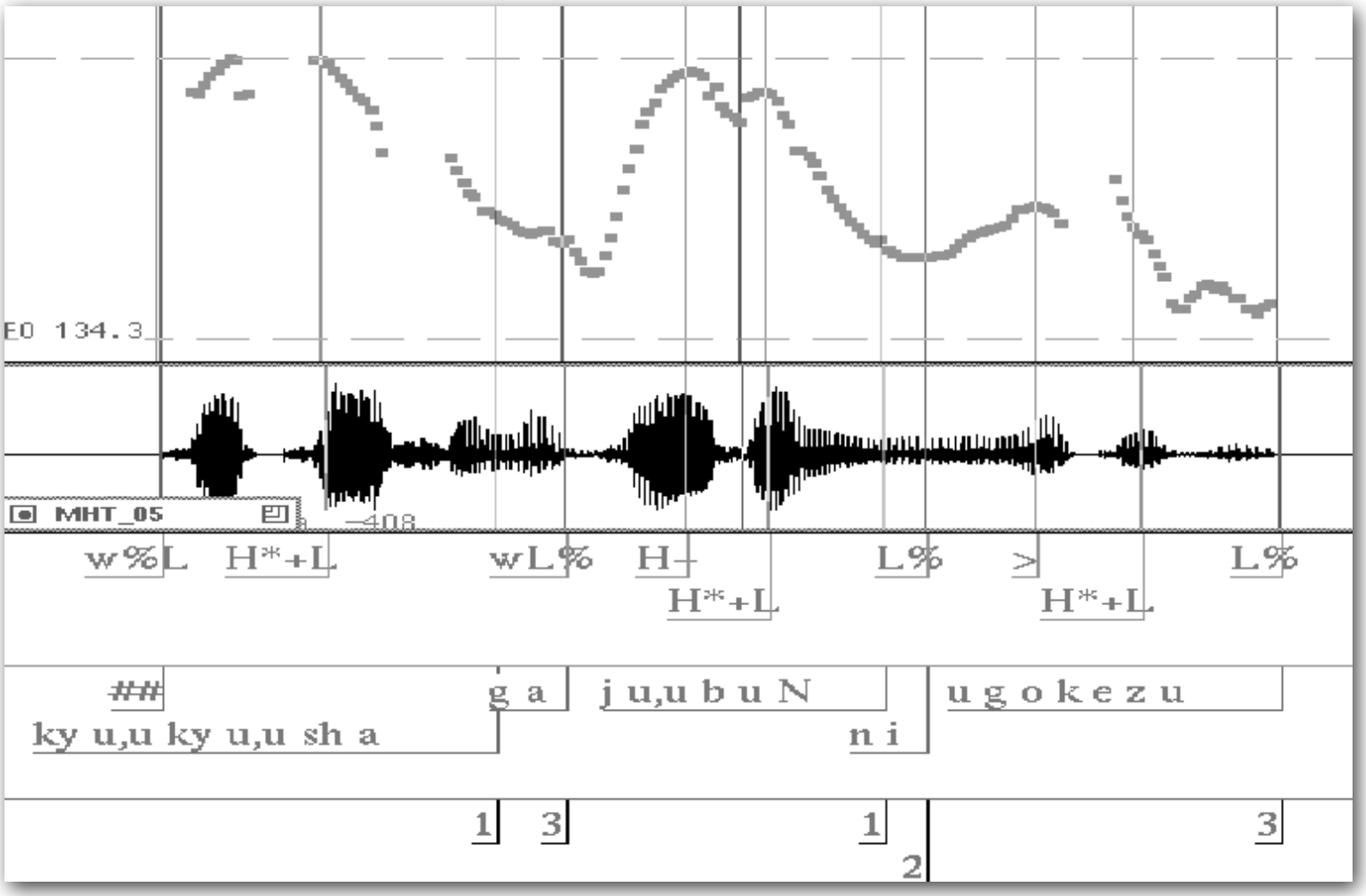


図７．合成音波形例　/harikiru/

波形編集方式

音声DBに対するラベリング・素片DB化

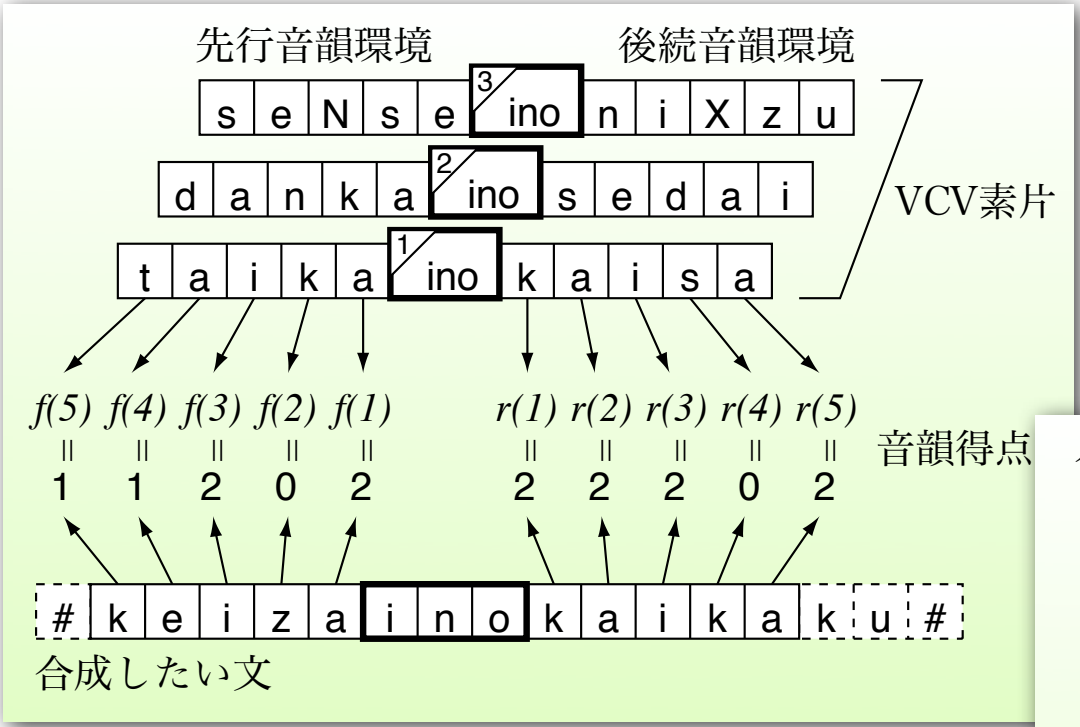
- 音韻ラベリングと韻律ラベリング
- 手動ラベリング／音声認識技術を使った自動ラベリング

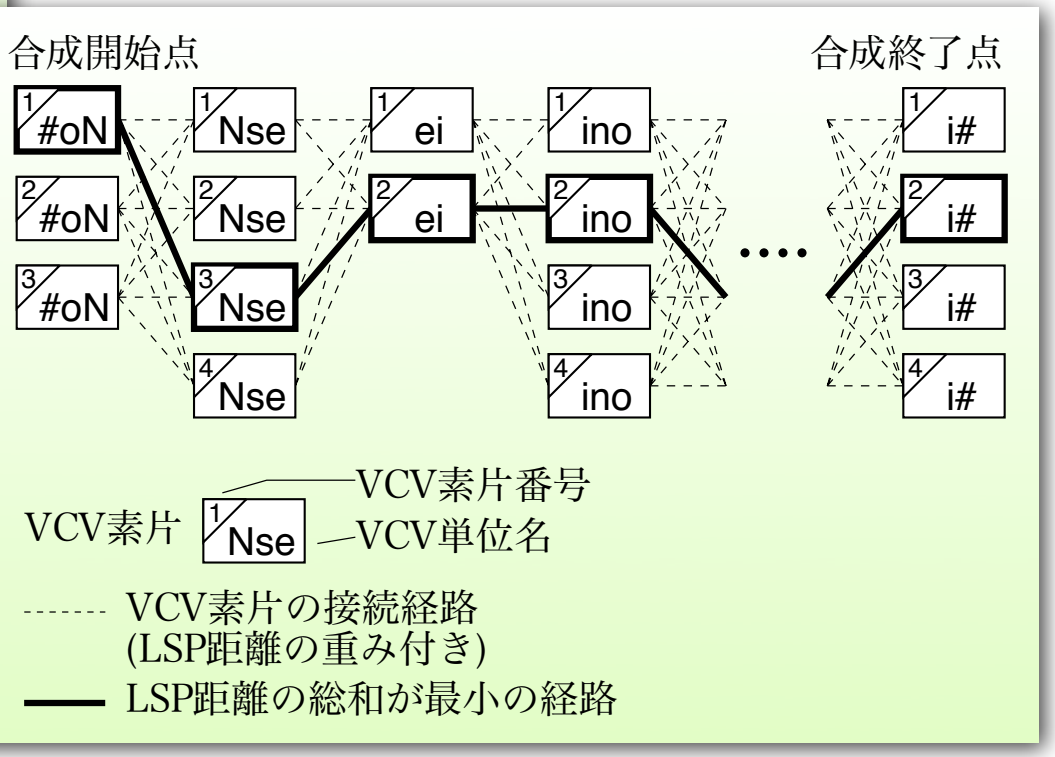


波形編集方式

接続コストを最小にする素片選択

- 素片を定義する／素片候補群から最適なものを選ぶ

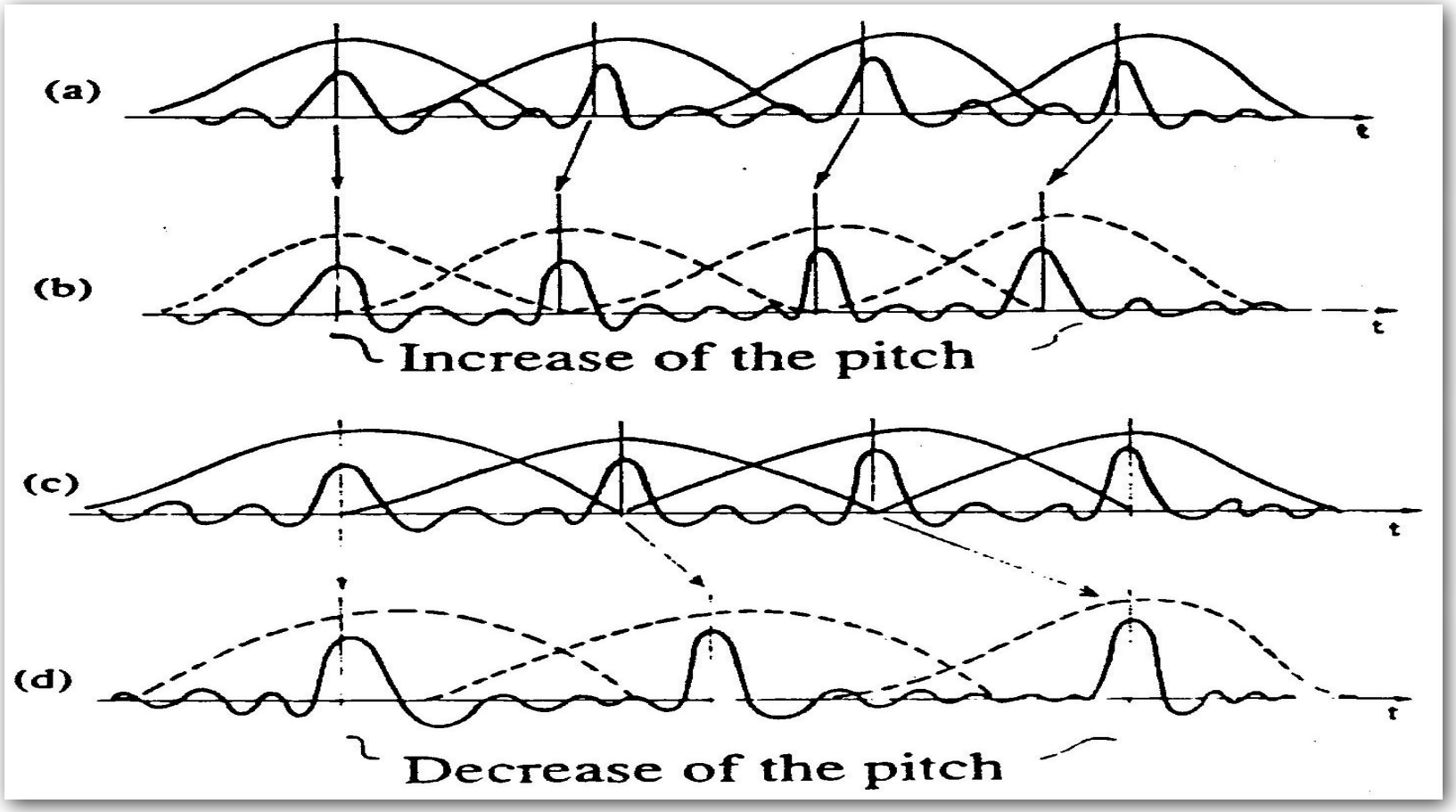




波形編集方式

所望の高さにしてから接続する

- TD-PSOLA
  - Time Domain Pitch Synchronous OverLap and Add
  - ピッチ波形を単位として，間隔を修正して切り貼り



波形編集方式

幾つかの合成サンプル

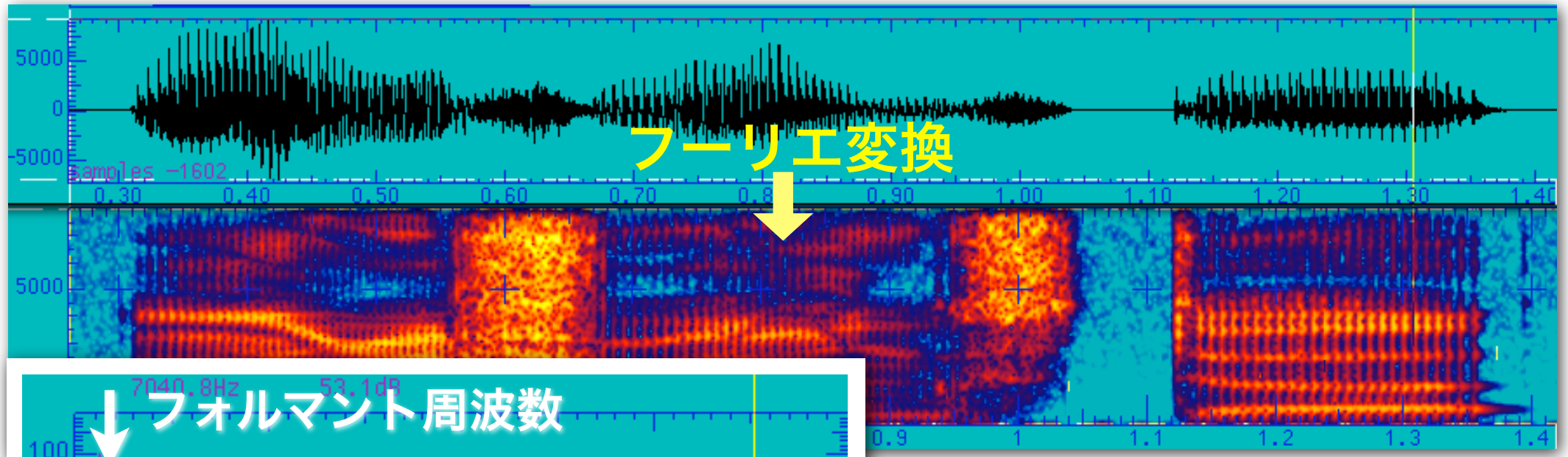
- 男声
- 女声
- その他
- http://www.ntt-it.co.jp/goods/vcj/voice/tts_demo.html

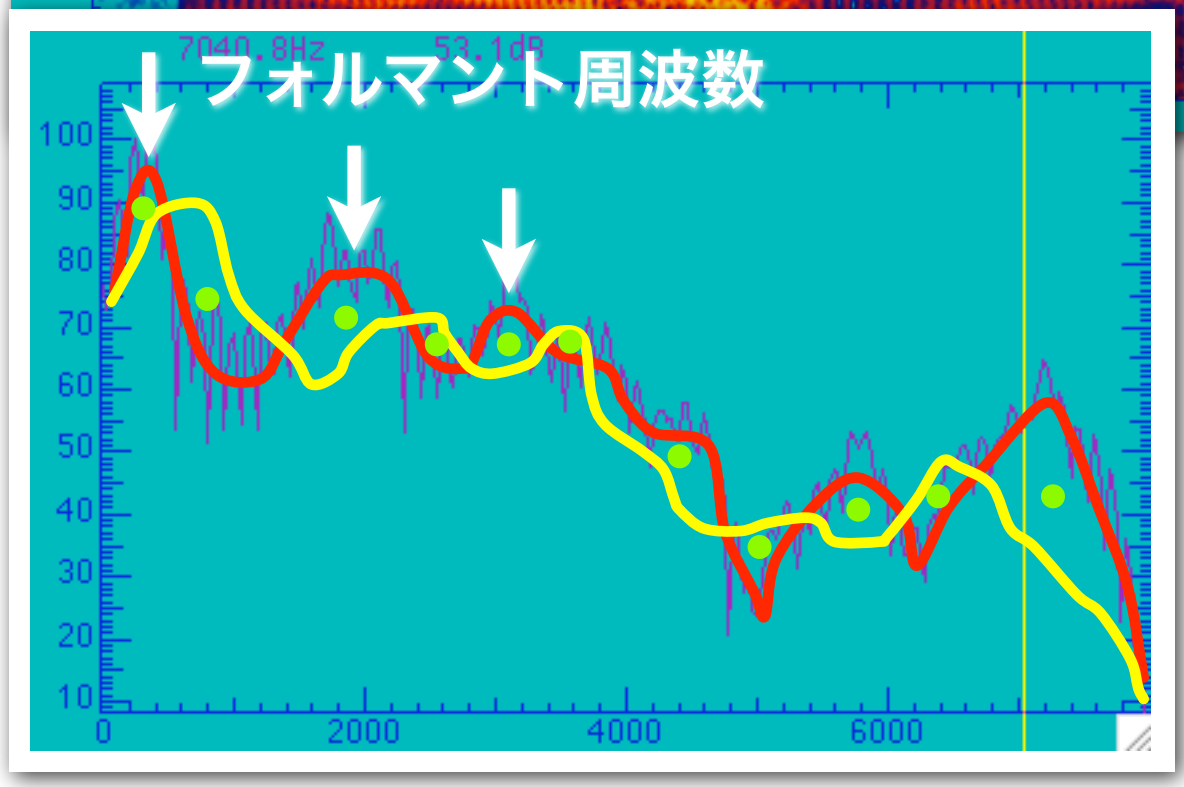
本日のメニュー

- **音声合成の原理と波形編集型音声合成**
  - 言語処理／音韻処理／波形処理
  - 音韻処理といっても結構面倒・・・
- **音声認識セミナーより**
  - 音声特徴の分析と照合
  - 隠れマルコフモデル（HMM）
  - サブワードモデルを用いた音声認識
  - 認識文法を用いた連続音声認識
  - 日本語ディクテーションにおける基礎技術
  - ビデオ鑑賞
- **宿題**
- **まとめ**

幾つか確認

音声波形→スペクトラム→ケプストラム





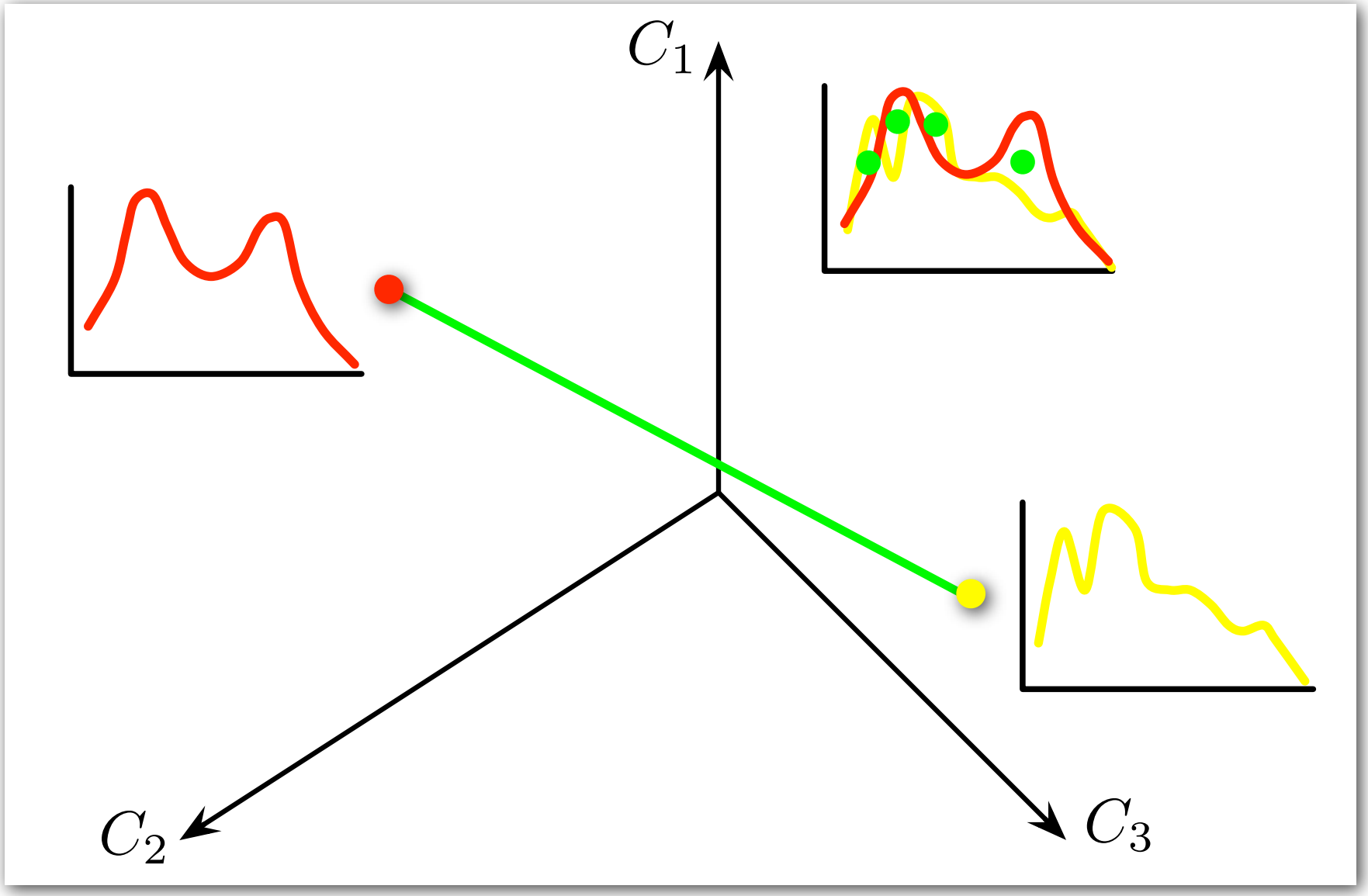
逆フーリエ変換

$$(c_0, c_1, \cdots, c_N)$$

$$\sum_{i=0}^{N} (c_i - c_i')^2$$

幾つか確認

- ケプストラム空間における「点」と「点間距離」



点＝スペクトル包絡，点間距離＝スペクトル間差異

宿題

課題１

- 第三回「音声の音響分析　波とフーリエ級数展開」
- 第四回「音声の音響分析と分析合成系」
- 第五回「音声合成と音声認識」
  - これらの講義に対して自主的に調査したことがあれば記しなさい。
  - 既に学んだ講義との関連性について記しなさい。
  - 理解できなかったことがあれば記しなさい。
  - 来年度以降に向けた改善点などあれば，それを記しなさい。

課題２

- 配布したスライドに対して，知るところを全て述べよ。
  - A, B, C... などとマーキングし，箇条書きにして記せ。
  - 文系の高校生に対して説明するように，自分の言葉で記述せよ。

提出日

- 12月5日の授業の開始時にマーク付きスライドと一緒に提出

本日のおさらい

- **音声合成の原理と波形編集型音声合成**
  - 言語処理／音韻処理／波形処理
  - 音韻処理といっても結構面倒・・・
- **音声認識セミナーより**
  - 音声特徴の分析と照合
  - 隠れマルコフモデル（HMM）
  - サブワードモデルを用いた音声認識
  - 認識文法を用いた連続音声認識
  - 日本語ディクテーションにおける基礎技術
  - ビデオ鑑賞
- **宿題**
- **まとめ**