

認知メディア論

2007-12-04

峯松 信明

新領域・基盤情報学専攻

前回のおさらい

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

まとめ

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

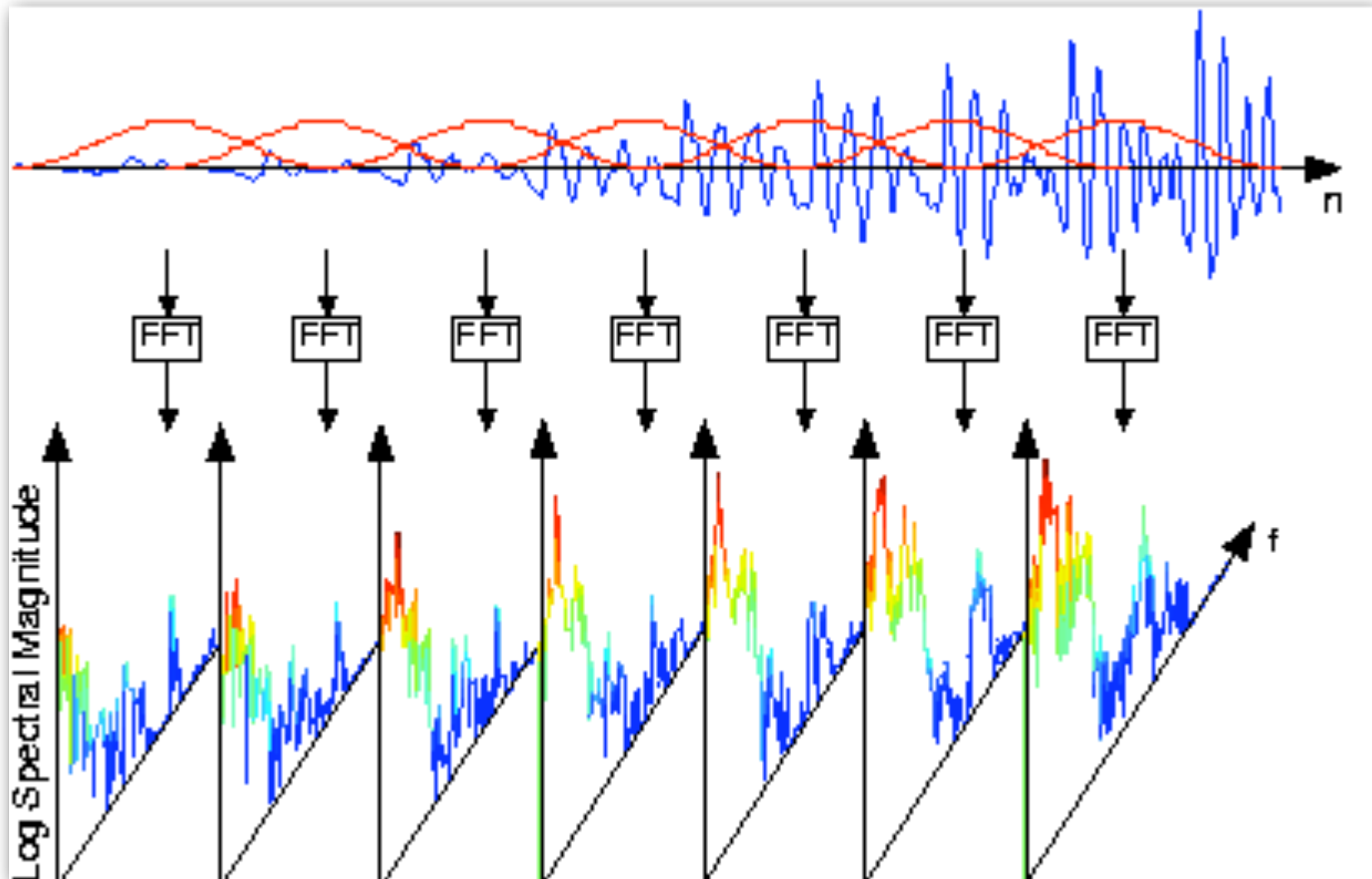
- Speech Transformation and Representation based on Adaptive Interpolation of weighted spectrogram
 - 適応的に内挿されたスペクトル表現
 - 河原英紀@和歌山大学システム工学部による開発
 - <http://www.wakayama-u.ac.jp/~kawahara/>



STRAIGHT

まずは復習。普通はこうやります。

● 32 msec の窓／シフト長は半分の 16 msec とか . . .

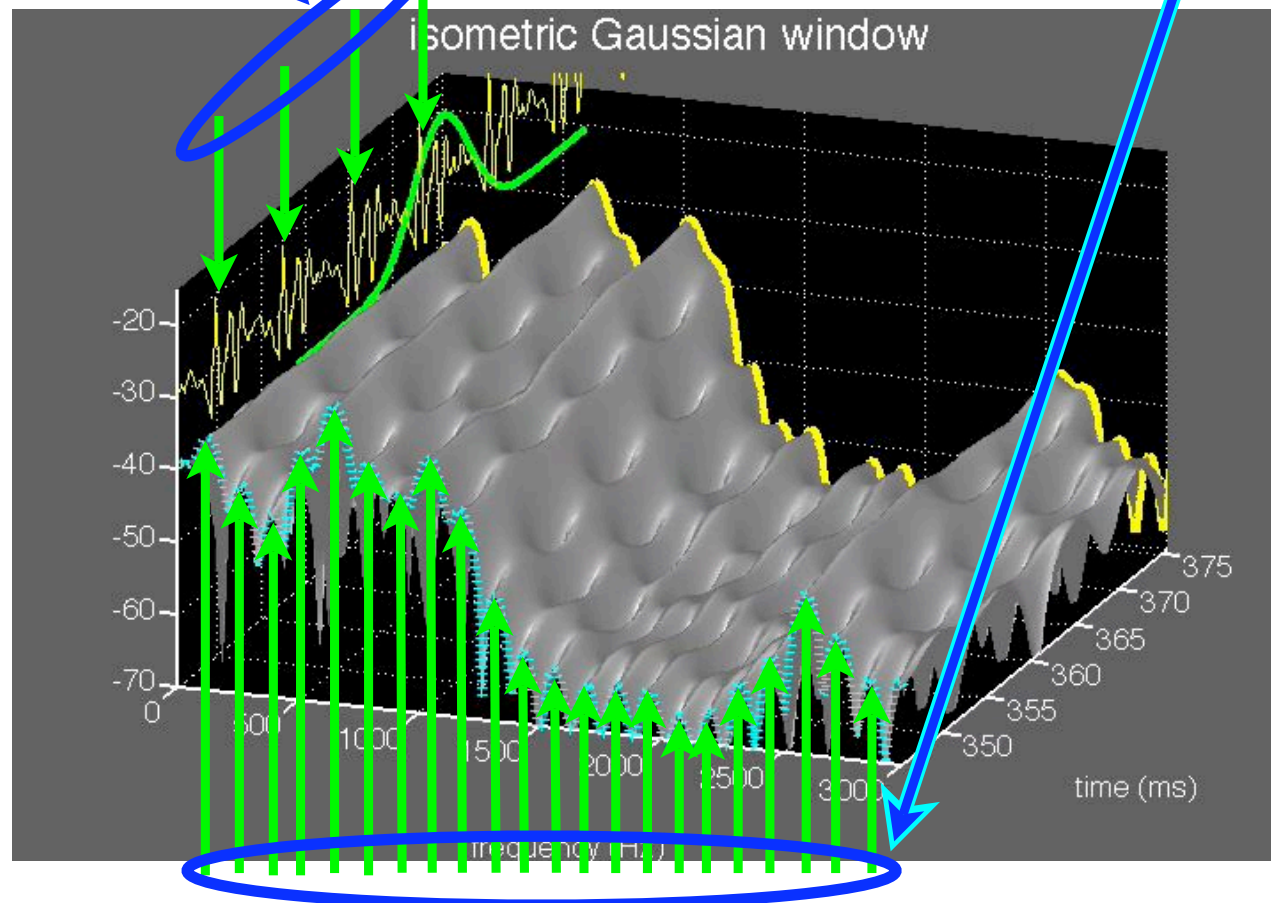


STRAIGHT

こういうところが嫌なのね・・・

periodic in the **time** domain

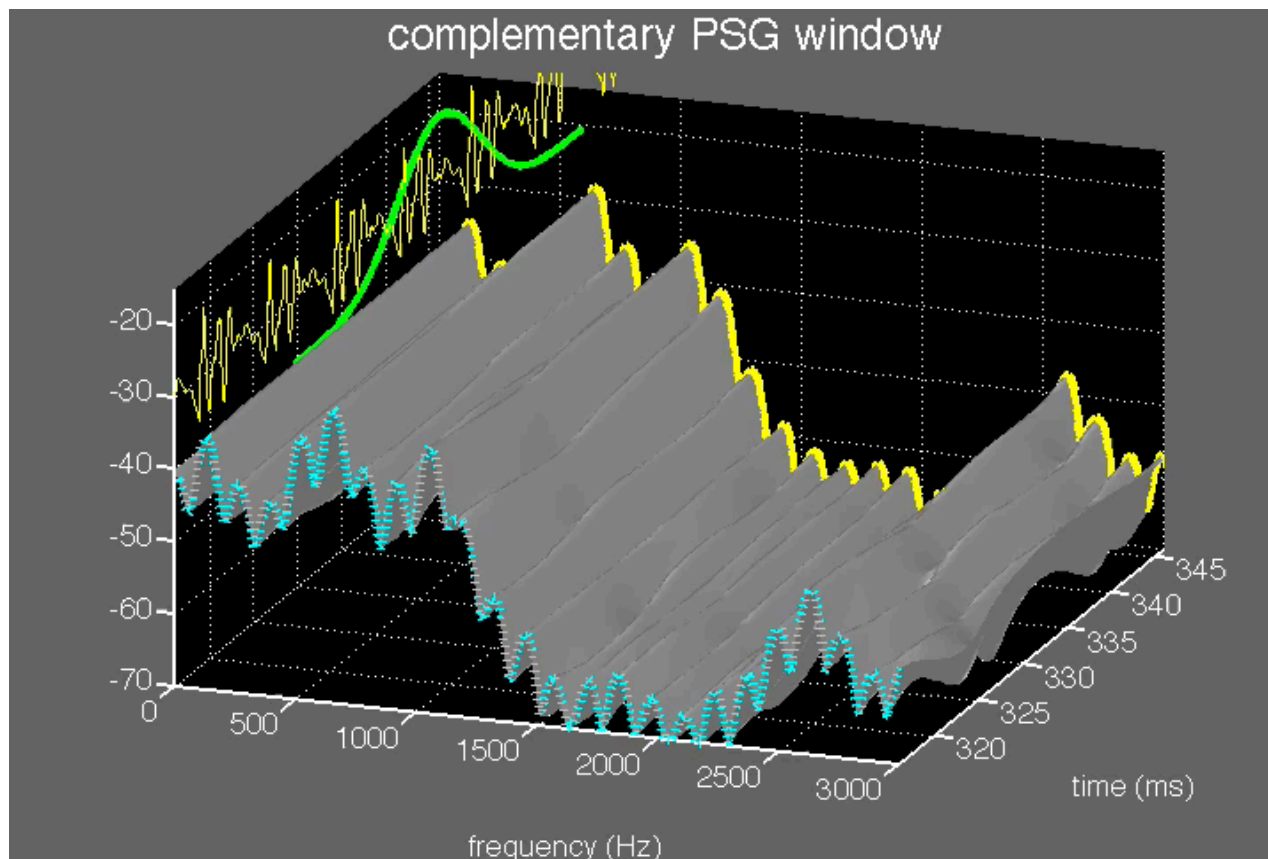
periodic in the **frequency** domain



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）

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

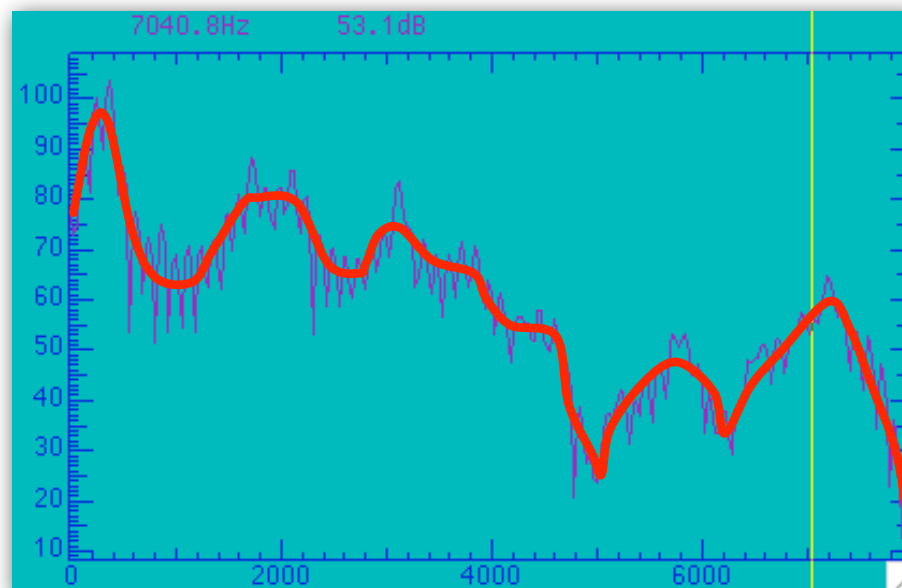
原音声

再合成音声（変形無し）

再合成音声（ F_0 ：2倍，周波数軸：1.25倍）

再合成音声（ F_0 ：3倍，周波数軸：1.44倍）

再合成音声（ F_0 ：0.5倍，周波数軸：0.8倍）



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

前回のおさらい

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

まとめ

音声によるコミュニケーション技術

～音声認識と音声合成～

峯松 信明

本日のメニュー

音声認識セミナーより

- 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

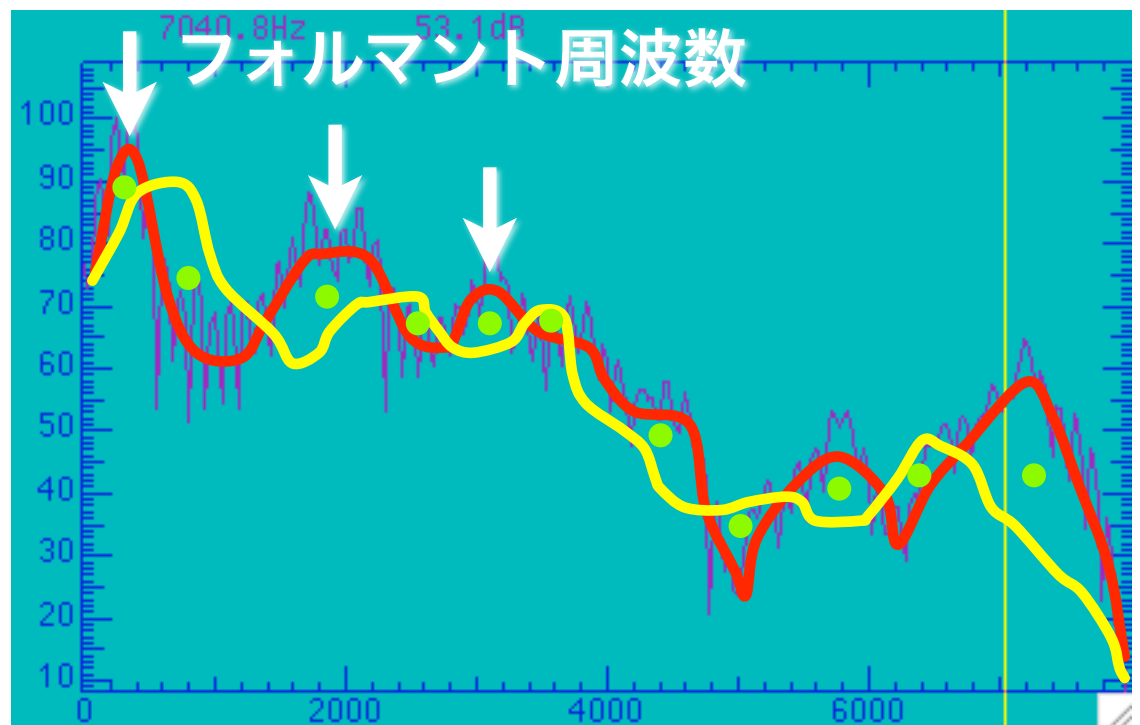
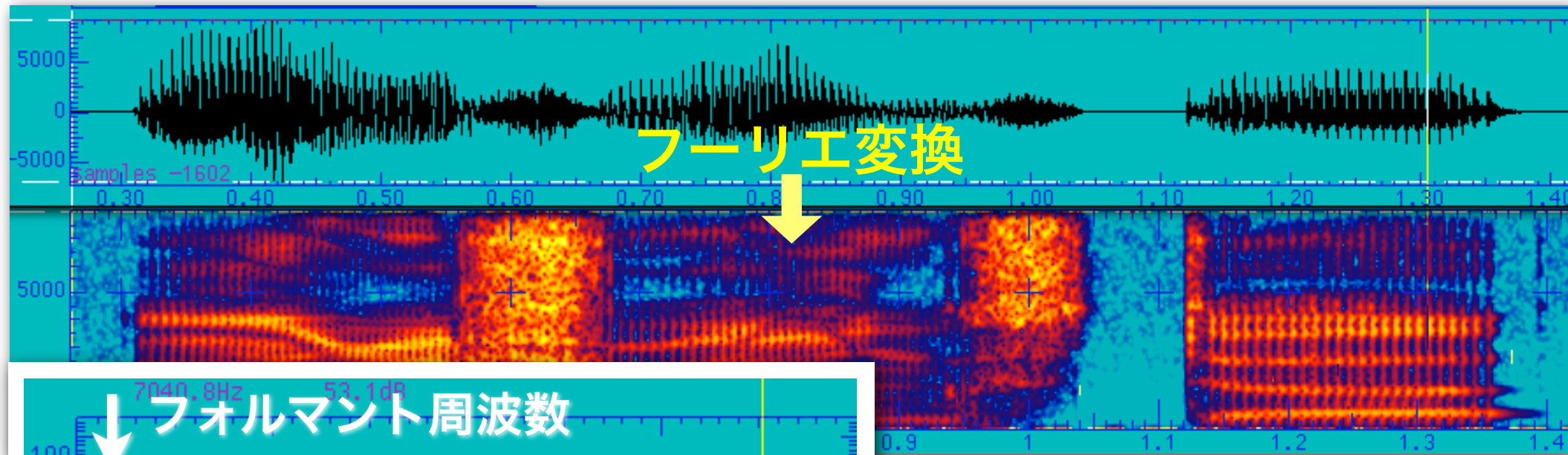
テキスト音声合成の原理

- 言語処理／音韻処理／波形生成
- 音韻処理：音素から単音へ／適切な韻律生成に向けて
- 波形生成：波形編集型，分析合成型（HMM音声合成）

まとめ

幾つか確認

音声波形→スペクトラム→ケプストラム



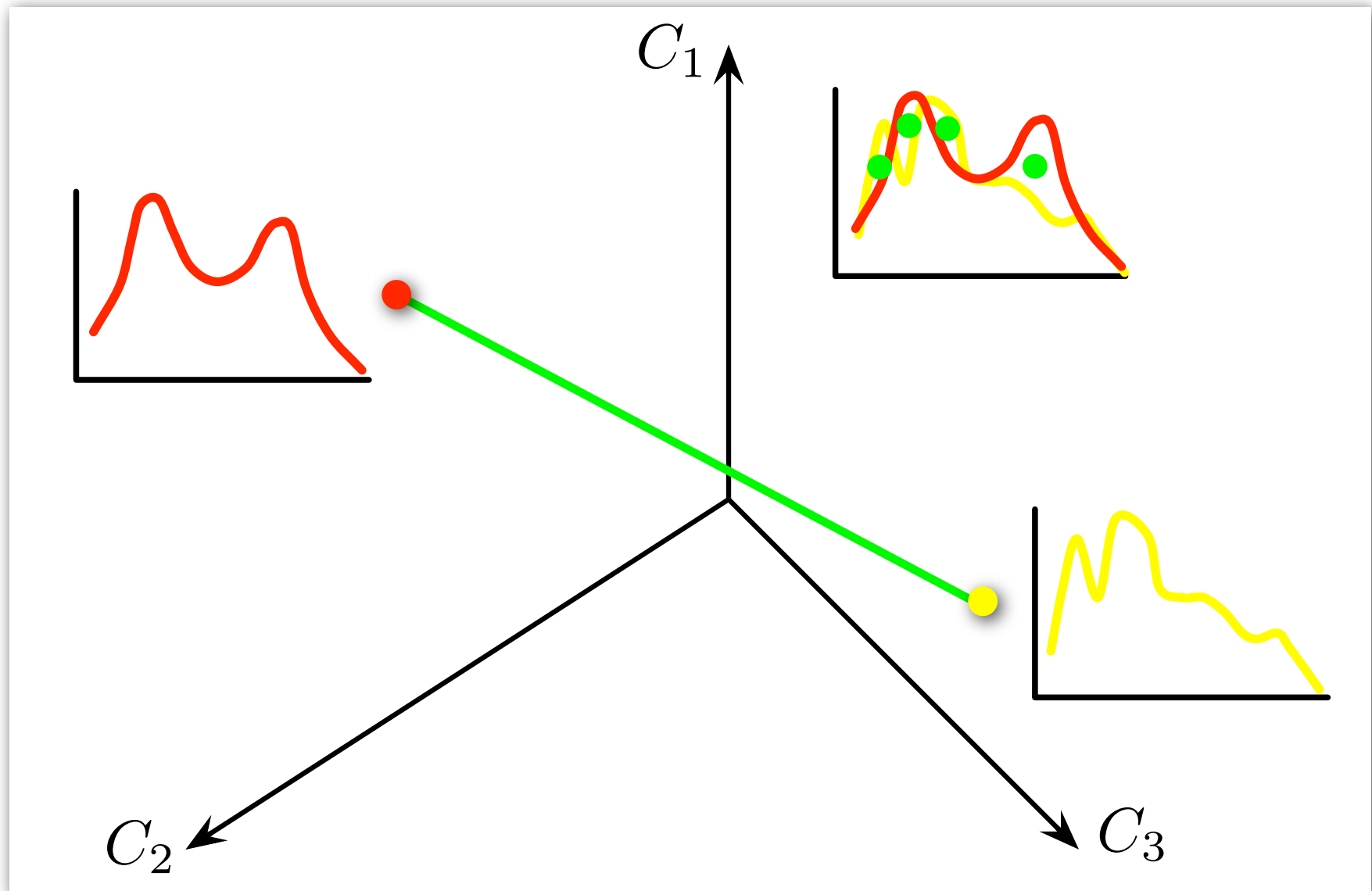
逆フーリエ変換

→ $(c_0, c_1, \dots, c_N)$

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

幾つか確認

ケプストラム空間における「点」と「点間距離」

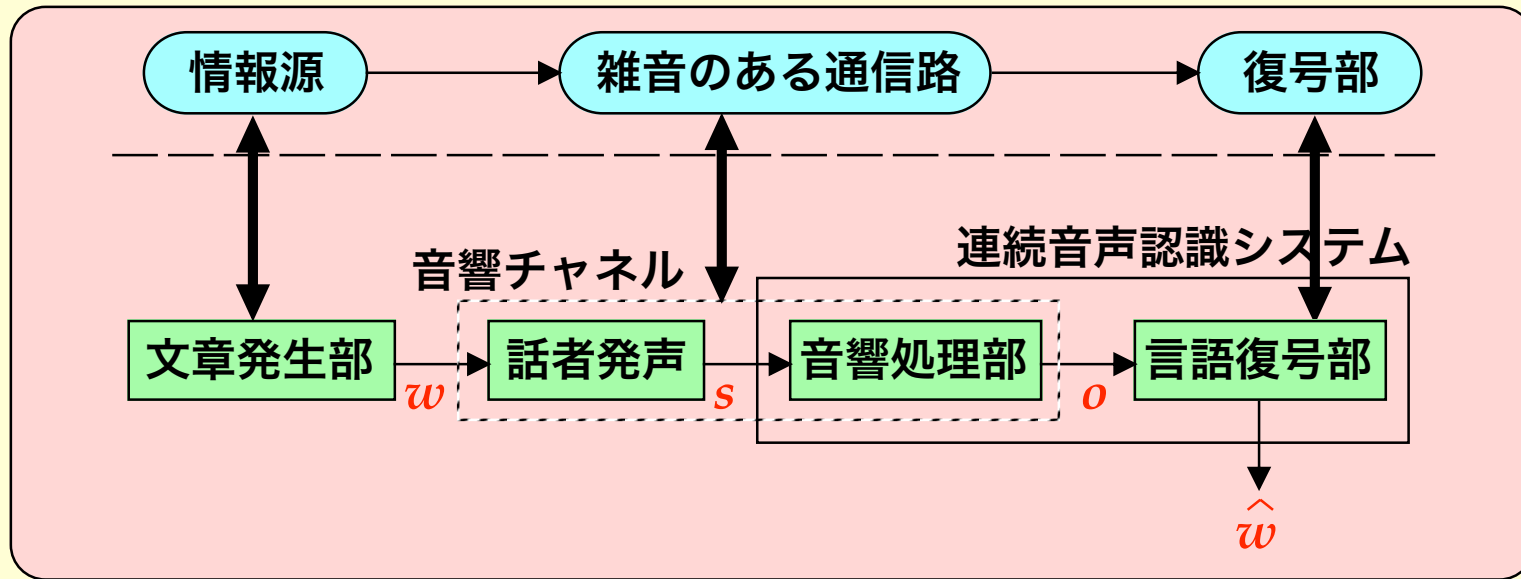


点 = スペクトル包絡, 点間距離 = スペクトル間差異

本講演の流れ

- 音声認識の情報理論的な枠組み
- 音声データベース / テキストデータベース
- 音響モデル / 「音」の統計的なモデリング
- 言語モデル / 「ことば」の統計的モデリング
- デコーダ / 2つのモデルと入力音声との照合, 認識エンジン
- 現状の音声認識精度 / 種々の条件下での認識結果の比較
- まとめと今後の課題

音声認識の情報論的な定式化



□ 音声認識の情報論的な定式化

✓ 音声認識 = $\arg \max_i P(w_i|o)$

✓ $P(w_i|o) = \frac{P(o|w_i)P(w_i)}{P(o)}$, $P(w_i)$: w_i の事前確率, $P(o)$: o の事前確率

✓ 音声認識 = $\arg \max_i P(o|w_i)P(w_i)$

✓ $P(w_i)$: w_i の事前確率 = 言語モデル, $P(o|w_i)$: w_i が与えられた時の o の事後確率 = 音響モデル

✓ 言モ = テキストデータベースより, 音モ = (書き起こし付きの)音声データベースより構築

音声データベース / テキストデータベース

- 言語モデル $P(w_i)$ = あるドメインにおいて如何なる言語事象(単語)が出現し易いのか？
 - ✓ 以前は, 規則で書くことが多かった(CFG)。
 - 音響モデル $P(o_i|w_i)$ = ある言語事象を前提とすると, 如何なる音響現象が観測され易いのか？
 - ✓ 以前は, テンプレートとして用意することが多かった(DPマッチング)。
- 
- 数理統計的な枠組みで各モデルを構築する方式へ (N-gram & HMM)
 - ✓ まず, データベースありき, , , , , の時代
 - ✓ テキストデータベース = 新聞データベース
 - ✓ 音声データベース = 各研究機関が独自で収録 / 公開を目的として複数機関が協力して収録
 - ✓ JNAS = { 男性150人, 女性150人 } x { 新聞記事 100文 + 音素バランス文 50文 }
 - ✓ IxM社では, 一社で更に大きな音声データベースを持っているらしい, , , 。
 - ✓ The larger, the better.....

音響モデル — HMM — (#1)

□ 音響モデリングの単位

✓ 単語を単位とすると単位数が爆発

➔ 音素, 音節を単位に

a i u e o a: i: u: e: o: N w y j m y k y b y g y d y n y h y r y p y
p t k t s c h b d g z m n s s h h f r q s p s i l E s i l B

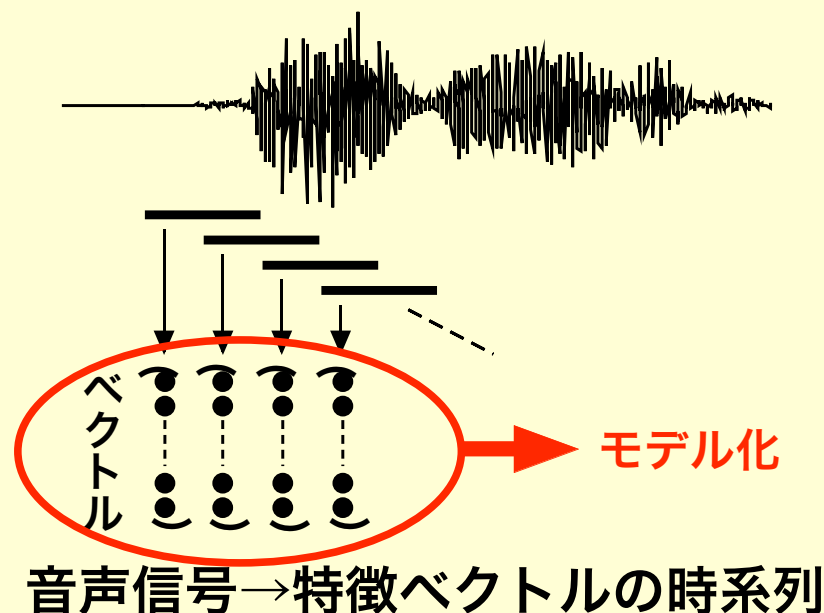
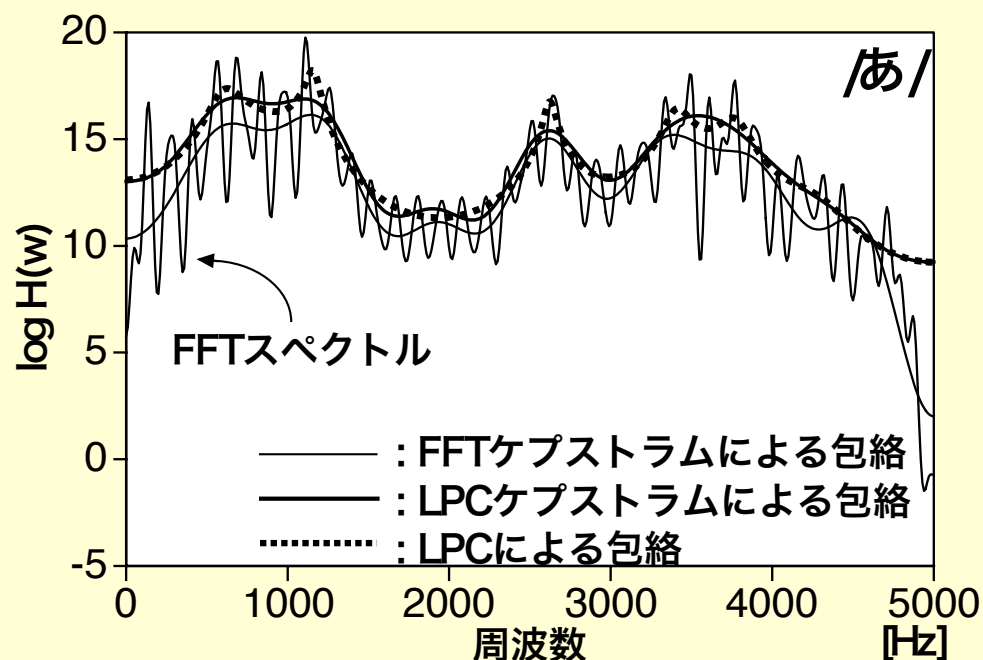
□ 音響特徴パラメータ

✓ /あ/ の音響的差異はどこに観測されるのか？

➔ 音声のスペクトル包絡

✓ スペクトル包絡の情報をどう符号化するか？

➔ ケプストラム



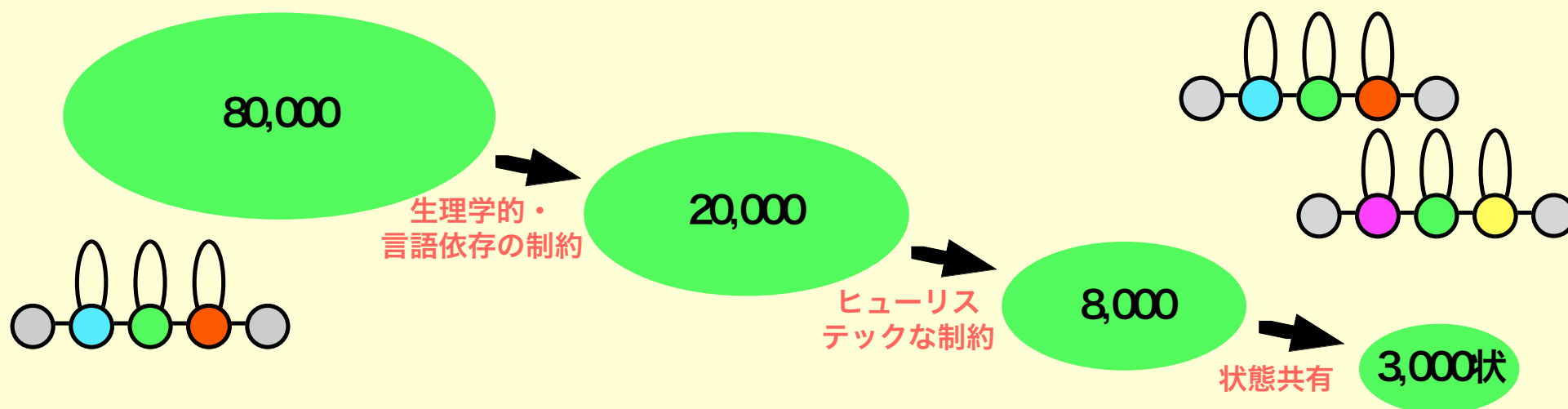
音響モデル — HMM — (#?)

□ 対象音韻の音響パラメータが持つ前後音韻への依存性

- ✓ 音響パラメータの分布は前後音韻に影響を受ける
- ✓ kAr, mEi, sAz, , , ,
- ✓ #triphone = [#monophone]^3 = approx. 80,000 ?

→ 前後音韻別のモデル

→ triphone



□ モデル精度向上に向けて

- ✓ 混合分布化
- ✓ 状態数, モデルのトポロジーの最適な配置
- ✓ 継続時間長制御の導入
- ✓ 特徴パラメータの速度成分, 加速度成分の導入
- ✓ セグメント化
- ✓ その他色々色々色々色々色々色々色々, , , , , , , , ,

言語モデル — N-gram — (#1)

□ 現在までの単語履歴から次の単語を予測するモデル

- $P(w_i | w_1, \dots, w_{i-1})$
- i の増加と共に求めるべきパラメータ数が激増, $N-1$ 重マルコフ過程を仮定
- $P(w_i | w_1, \dots, w_{i-1}) = P(w_i | w_{i-N+1}, \dots, w_{i-1})$, 過去の $N-1$ 個の単語履歴にのみ依存
- $P(w_i | w_{i-N+1}, \dots, w_{i-1}) = C(w_{i-N+1}, \dots, w_i) / C(w_{i-N+1}, \dots, w_{i-1})$: N -gram の最尤推定値
- $P(w_i | w_{i-1}) = C(w_{i-1}, w_i) / C(w_{i-1})$: **バイグラム**
- $P(w_i | w_{i-2}, w_{i-1}) = C(w_{i-2}, w_{i-1}, w_i) / C(w_{i-2}, w_{i-1})$: **トライグラム**

□ N -gram 推定時の問題 — データが足りない —

- ✓ 全語彙を対象とするのではなく, 頻出 M 単語を対象語彙として限定する
- ✓ 単語のトライグラム(語彙数 $20K$)を精度良く推定するためには数年間の新聞データでは不足
- $C(w_{i-2}, w_{i-1}, w_i) = 0$, あるいは, $C(w_{i-2}, w_{i-1}) = 0$ の場合どうする???

□ Back off スムージングによる N -gram 確率の推定(未知 N -gramへの対処)

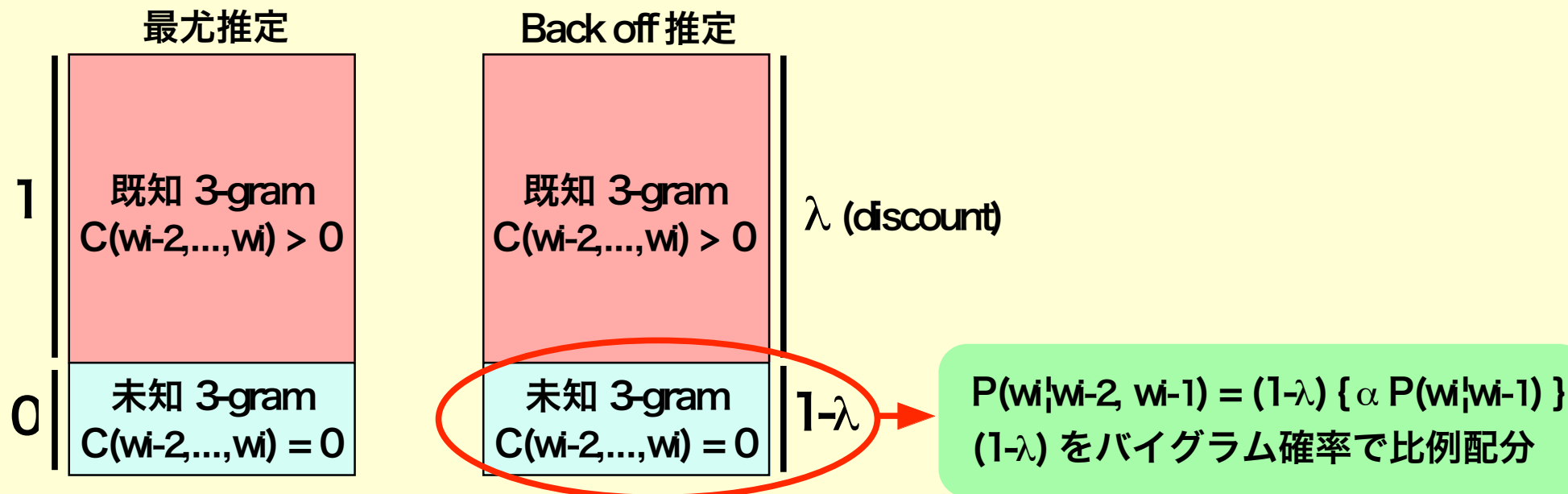
- ✓ $C(w_{i-2}, w_{i-1}) = 0$ の時
 - $P(w_i | w_{i-2}, w_{i-1}) = P(w_i | w_{i-1})$: バイグラム確率を流用
- ✓ $C(w_{i-2}, w_{i-1}, w_i) = 0$ (但し, $C(w_{i-2}, w_{i-1}) > 0$) の時
 - **既存**の N -gram 確率の一部を**未知**の N -gram の確率として分配する

言語モデル — N-gram — (#2)

□ Back off スムージングによる 3-gram 確率の推定 (未知 3-gram への対処)

✓ $C(w_{i-2}, w_{i-1}, w_i) = 0$ (但し, $C(w_{i-2}, w_{i-1}) > 0$) の時

➔ 既存の 3-gram 確率の一部を未知の 3-gram の確率として分配する



□ 種々の Back off スムージング

✓ λ を w_i に依存させ, 様々な知見に基づいた Back off スムージングが提案されている

□ カットオフの導入

✓ $C(w_{i-2}, w_{i-1}, w_i) = 0$ のみならず, $C(w_{i-2}, w_{i-1}, w_i) \leq C_0$ 時にも Back off により推定する

□ 単語クラス (品詞, 上位概念) を利用した N-gram

言語モデル — 辞書 / 語彙サイズ — (#3)

□ 認識対象語彙 = テキストデータベースにおける頻出単語

□ 語彙サイズと被覆率(coverage)

✓ 45 か月分の新聞データを元に算出

✓ 「句読点」も語彙の一部。該当する音素表記は無音記号(sp, silB, silE)

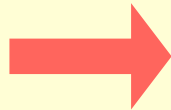
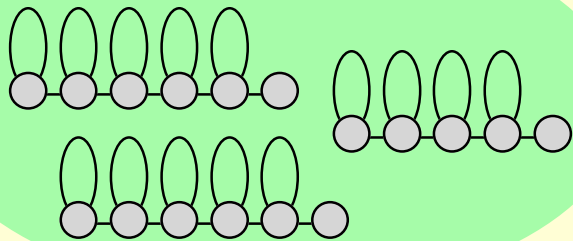
語彙サイズ	被覆率
5,000	88.2 %
6,135	90.0 %
20,000	96.5 %
22,959	97.0 %

デコーダ(認識エンジン, #1)

□ 基本的なアルゴリズム(one-pass アルゴリズム)

✓ バイグラム言語モデルを使ったデコーディング

単語 HMM セット
(音素 HMM の連結)

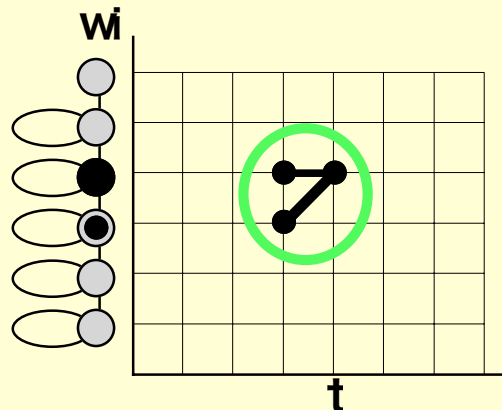


時刻 t の入力フレームは**何れの単語の何れの状態とも**照合を行わなければならない

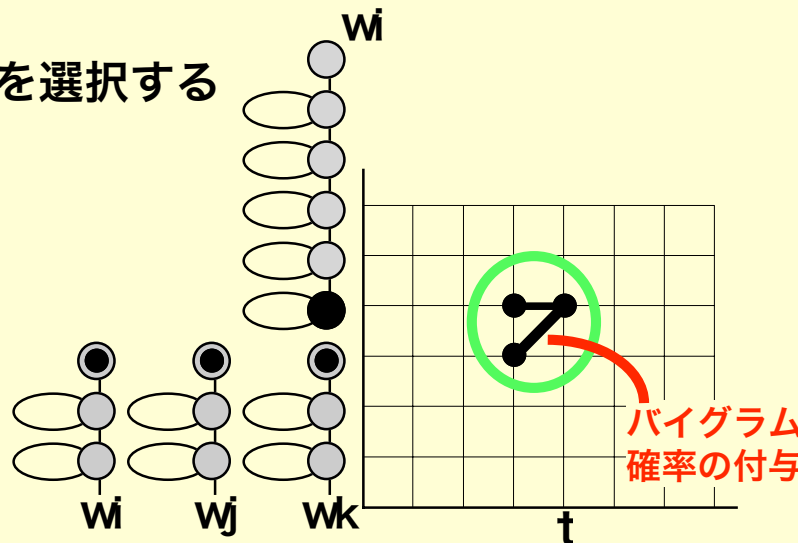
入力フレーム t は**全状態**と照合される

[時刻 t が w_i の状態 s と対応する場合の最適パス]を各単語, 各状態ごとに保持

○ : 累積スコアが高いパスを選択する

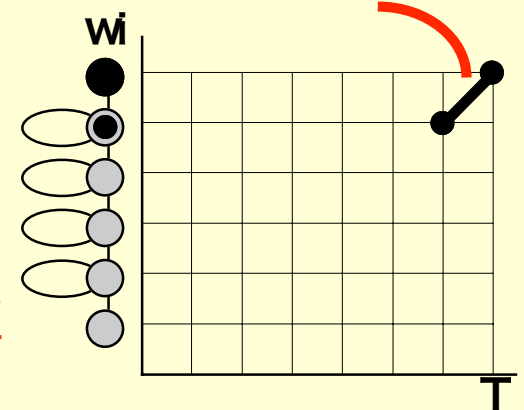


単語の非第一状態との照合



単語の第一状態との照合

最適パスを文尾から
バックトレースする



時刻Tのフレームの照合

デコーダ(認識エンジン, #2)

□ 語彙数の増加 / 状態数の増加 → 記憶容量 / 計算量の増加

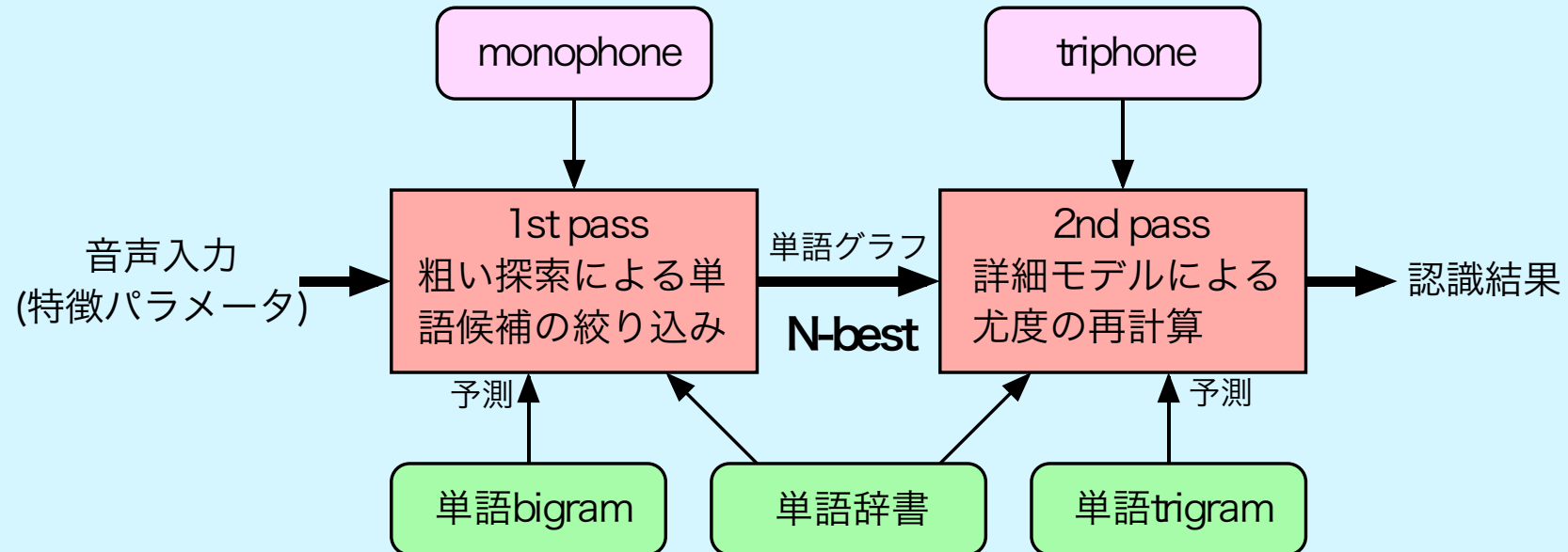
✓ **ビームサーチ**

→ 各時刻において累積尤度の上位 M 個を残し, それ以外は切り捨てる

□ 音響モデル / 言語モデル / デコーダの統合

✓ **マルチパス探索**

→ 1st pass で粗いモデルによる予備探索, 2nd pass で詳細なモデルによる再評価



デコーダ(認識エンジン, #3)

□ 種々のテクニック

✓ 枝刈りに関して

探索時点 t までの累積尤度と、それ以降の時刻 T までのフレームに対する評価値を使って刈る枝を決定する。時刻 T までのフレームに対する評価値は heuristics により決定する。

✓ 木構造化辞書

prefix が同じ単語を木構造化し、照合処理の効率を図る。

✓ 言語モデル確率のファクタリング

単語尾と単語頭の接続部において言語モデル確率は付与されるが、これが単語内部での照合においても付与されるよう、言語モデル確率を分割して各ノードに割り振る。

✓ 言語尤度重み

音響尤度に比べて言語尤度はそのレンジが小さいため、言語尤度に重みをかけて両者を足し合わせるにより、認識精度の向上を図る。

✓ インサクションペナルティー

短い単語の挿入誤りを防ぐため、照合中に新たな単語へと遷移する度にペナルティーを付与する。言語尤度重みの大きさに依存して設定される。

音声認識性能 (IPA'98)

□ 連続音節認識結果

- ✓ 音響モデル：状態共有の triphone, 総状態数 = 3,000
- ✓ 言語モデルは使用せず。純粹に音響モデルの性能評価
- ✓ 評価文数 = 約 930

特徴パラメータ	Acc.	Cor.	Sub.	Ins.	Del.
MFCC + Δ MFCC + Δ pow	66.5	80.6	16.7	14.1	2.8
上記 + Δ^2 MFCC + Δ^2 pow	70.8	83.6	14.4	12.8	2.0

正解率(Cor.) + 置換率(Sub.) + 脱落率(Del.) = 100

正解精度(Acc.) = 正解率(Cor.) - 挿入率(Ins.)

□ 大語彙連続音声認識(連続単語認識)結果

- ✓ 音響モデル：状態共有の triphone, 総状態数 = 3,000
- ✓ 言語モデルは45ヶ月分の新聞記事から作成 / 語彙数 = 20 K
- ✓ 評価文数 = 約 100

デコーダ	音響モデル	言語モデル	計算時間	Acc.	Cor.
高速版	monophone	圧縮triphone	2.2 x RT	84.6	85.7
高精度版	triphone	triphone	8.5 x RT	92.6	93.7

種々の条件下における認識結果例

□ 連続音韻認識結果(triphone の任意連結)

SILQbe:kokudeoNobetonamukita:Nhe:einokokumiNnomewachimetakuSILSILQayag
adoo:ojoo:watsunerumadeiniwaSILSILtsukanarinosaigetsohichiootoshita

□ 連続音節認識結果(上記+音節構造の知識導入)

SILがいこくでおんおべとなむきたんへいのこくみんのめわちめたくSILSILっあやがどおじょお
わっねるまでいになSILSILっかなりのさいげっおひちおおとしたSIL

□ 連続単語認識結果(上記+単語の知識導入, 語彙数=20K)

1st pass 米穀 ネオン ベトナム 機関 平 残っ 区民 度目 月 目立っ 句。 ? カヤ 花道 王女 大和 詰める まで
なり なさい えっ 消費 治療 落とし 他

2nd pass 米穀 ネオン ベトナム 帰還 平 残っ 区民 度目 月 目立っ く、 、 カヤ 門 王女 大和 詰める まで 庭
り なさい れ 曹 陽 治療 落とした

□ 大語彙連続音声認識結果(上記+単語間の連鎖知識導入)

1st pass 米国のベトナム帰還兵の国民の目が冷たく、彼らは同情を集めるまでには、かなりの歳月を
必要 落とした。

2nd pass 米国のベトナム帰還兵の国民の目は冷たく、彼らが同情を集めるまでには、かなりの歳月を
必要 とした。

□ 正解文

米国でもベトナム帰還兵への国民の目は冷たく、彼らが同情を集めるまでには かなりの歳月を必要と
した。

まとめ

□ 音声認識の情報論的な定式化

□ 音声言語データベース / テキストデータベース

- ✓ 現在は「まず、データありき」のご時世, , ,

□ 音響モデル $P(o|w)$ — HMM —

- ✓ 音声の生成モデルとしての HMM, 特徴ベクトル = observed, 状態系列 = hidden
- ✓ Viterbi 演算による尤度 $P(o|w)$ 計算
- ✓ 環境依存型の triphone

□ 言語モデル $P(w)$ — N-gram —

- ✓ マルコフ過程を仮定した単語予測モデル
- ✓ Back off スムージングによるトライグラムの推定

□ デコーダ(認識エンジン)

- ✓ one-pass アルゴリズム
- ✓ 音響モデル / 言語モデル / デコーダの統合: マルチパスによる構成

□ 現状の認識精度(技術レベル)とその例

- ✓ IPA'98 の性能評価
- ✓ 種々の条件下における認識結果の比較

本日のメニュー

音声認識セミナーより

- 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

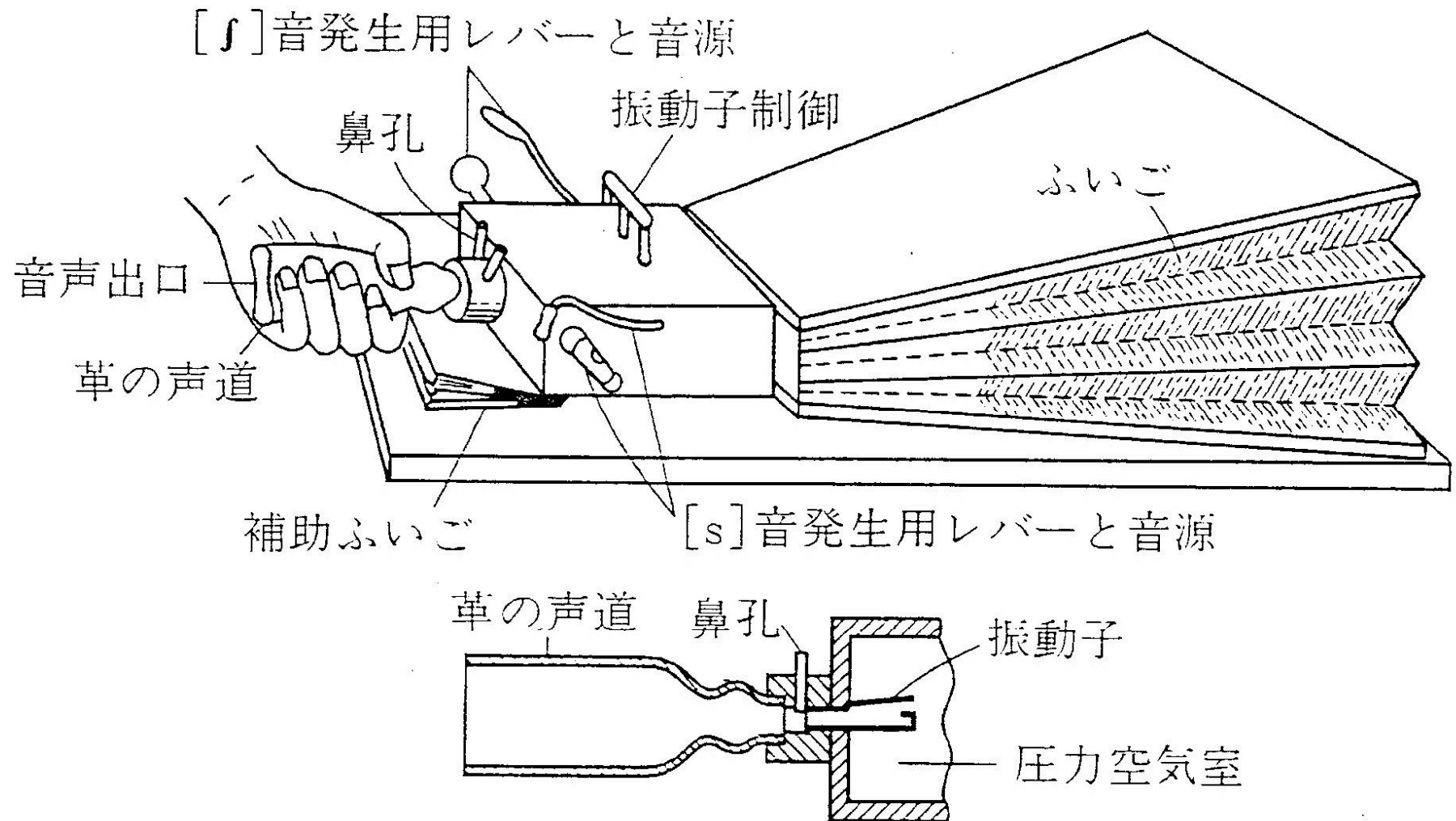
テキスト音声合成の原理

- 言語処理／音韻処理／波形生成
- 音韻処理：音素から単音へ／適切な韻律生成に向けて
- 波形生成：波形編集型，分析合成型（HMM音声合成）

まとめ

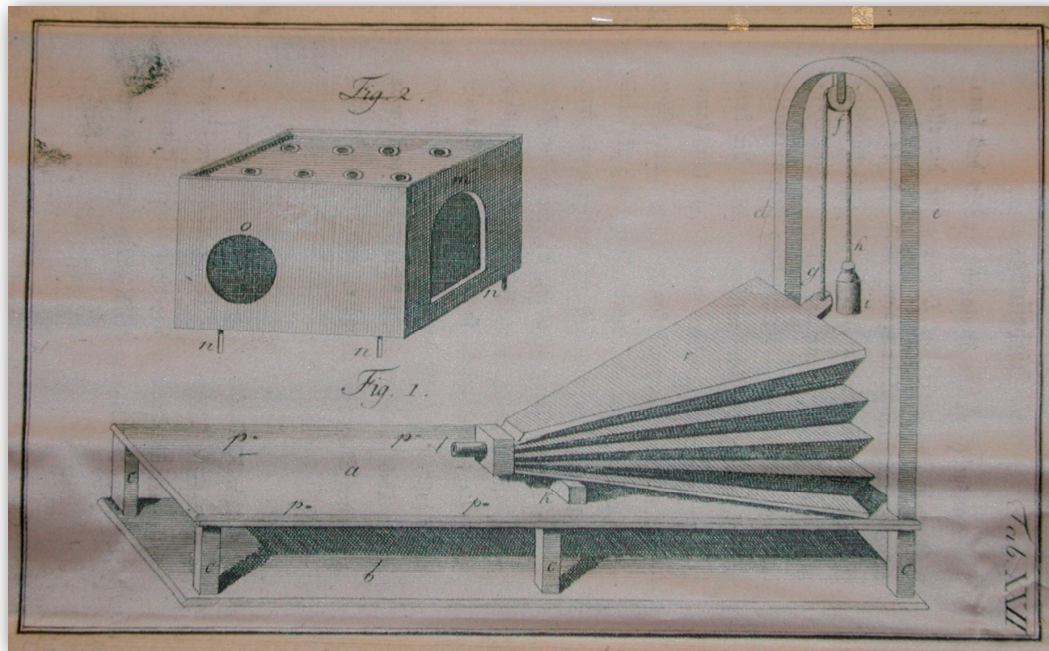
世界最古の音声合成器

von Kempelenの機械式音声合成器



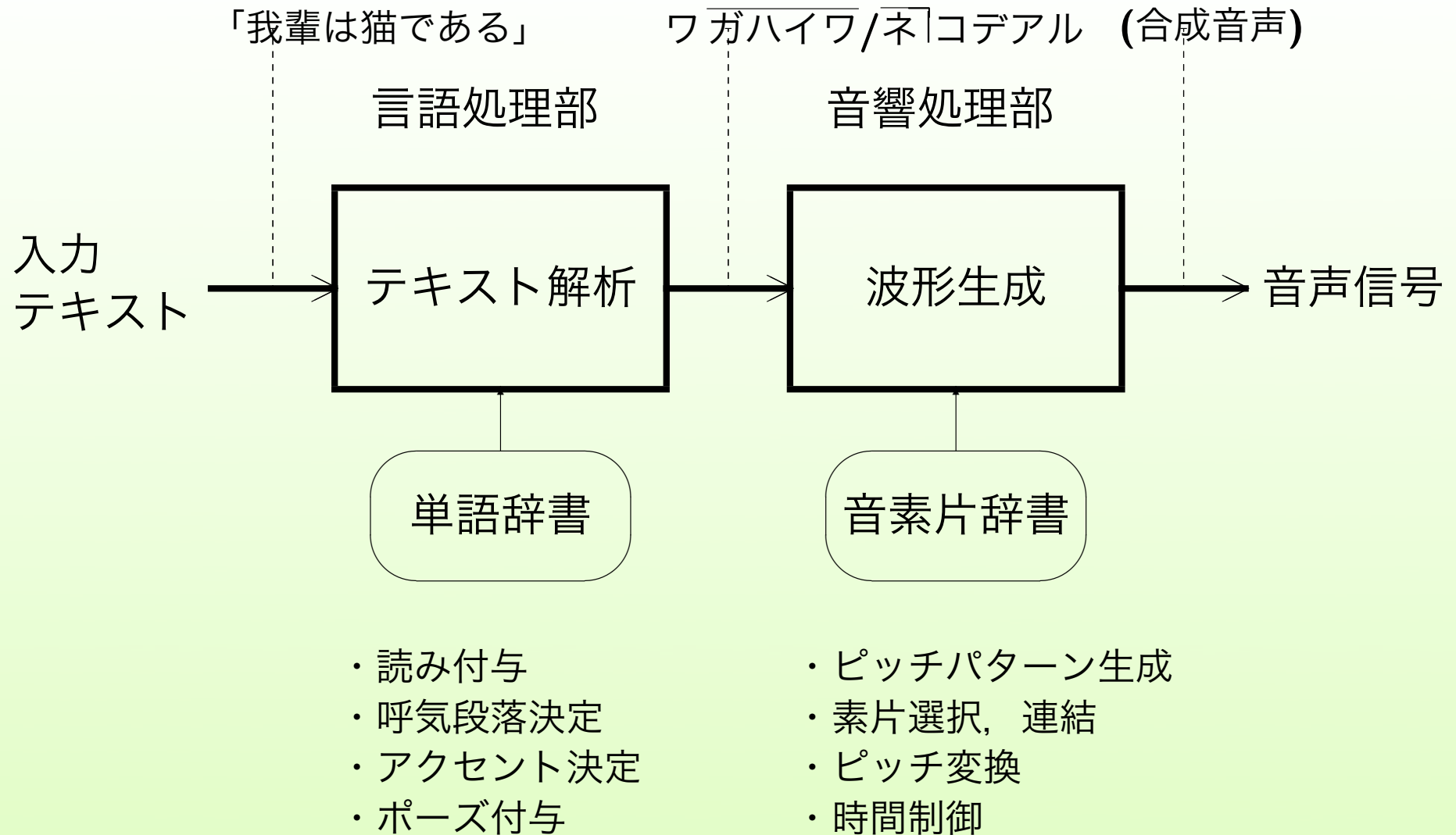
世界最古の音声合成器

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

単語分割後, 波形生成前に必要な（無意識的な）処理

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）

- 一卷，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，ペン先

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

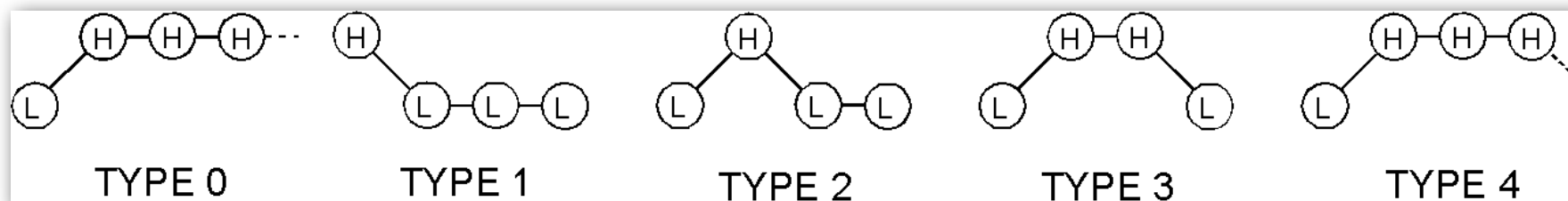
記号が入力されても某の音にしなければならない

- %, 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：接続後に伴い現れる核の位置を接続境界からのモーラ数で表現したもの。接続位置に核（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 $\tilde{M}_2$ 価 $\rightarrow N_c$ 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	$N_1 + \tilde{M}_2$	M_1
(F3) 融合型	M_1	$N_1 + \tilde{M}_2$
(F4) 支配型	$N_1 + \tilde{M}_2$	
(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	$N_2 \geq 2, M_2 \neq 0, N_2^\dagger$	$N_1 + M_2$
(C2) 生起型	$N_2 \geq 2, M_2 = 0, N_2^\dagger$	$N_1 + 1$
(C3) 標準型	$N_2 \leq 2$	N_1
(C4) 平板型	$N_2 \leq 2$	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

単語分割後, 波形生成前に必要な（無意識的な）処理

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）

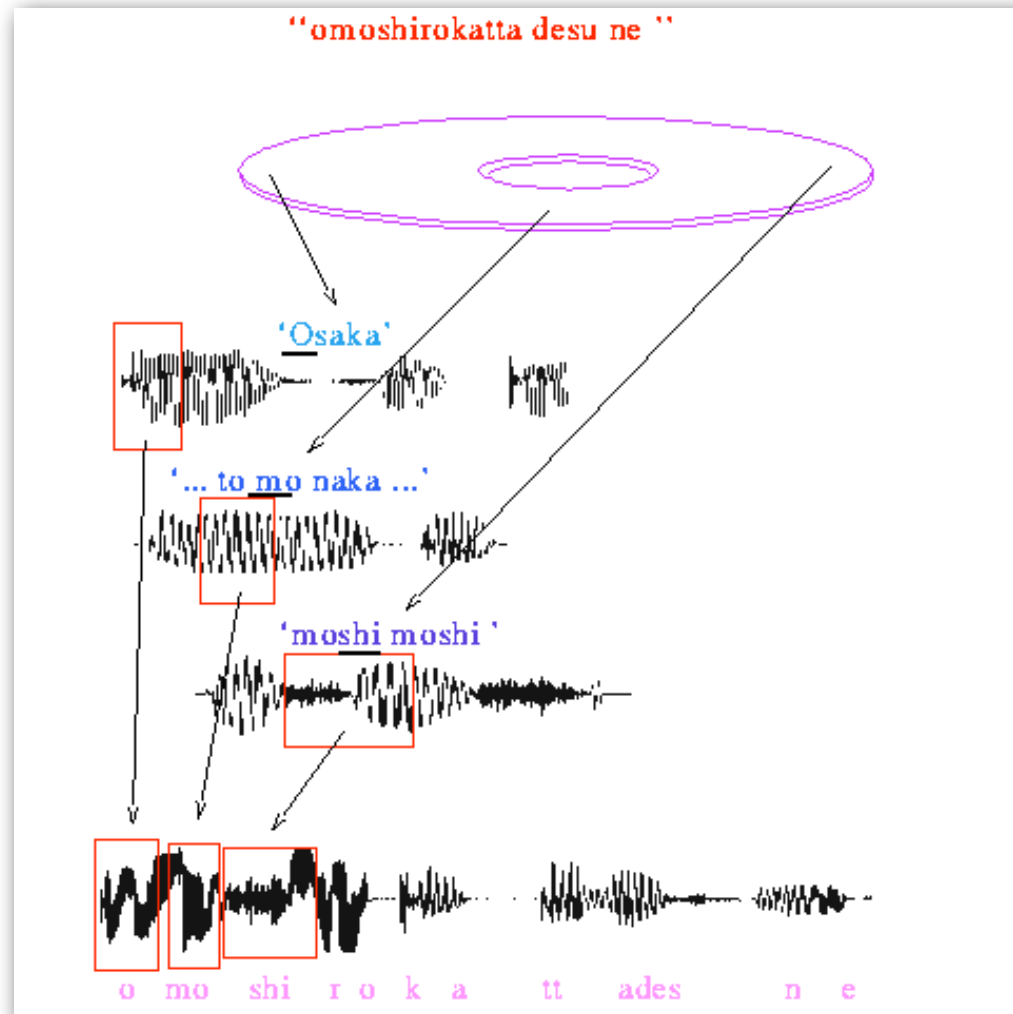
- 周波数領域での音声生成過程の模擬

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

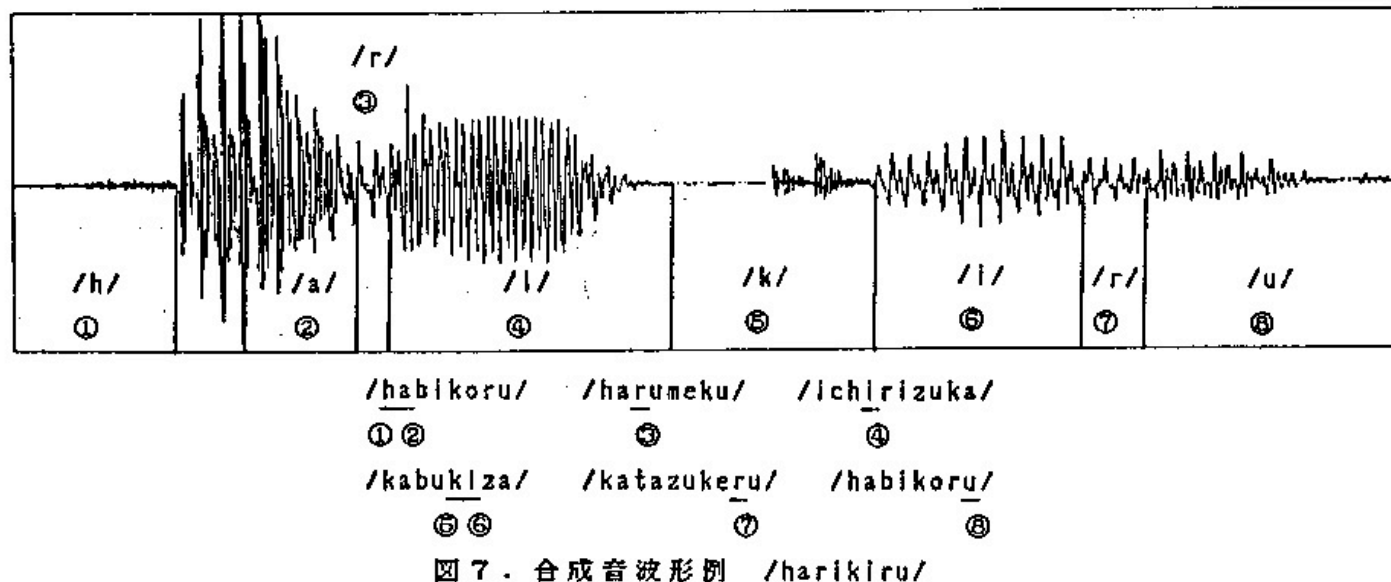
● 音素, 音節, 半音節, どれを使う?

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

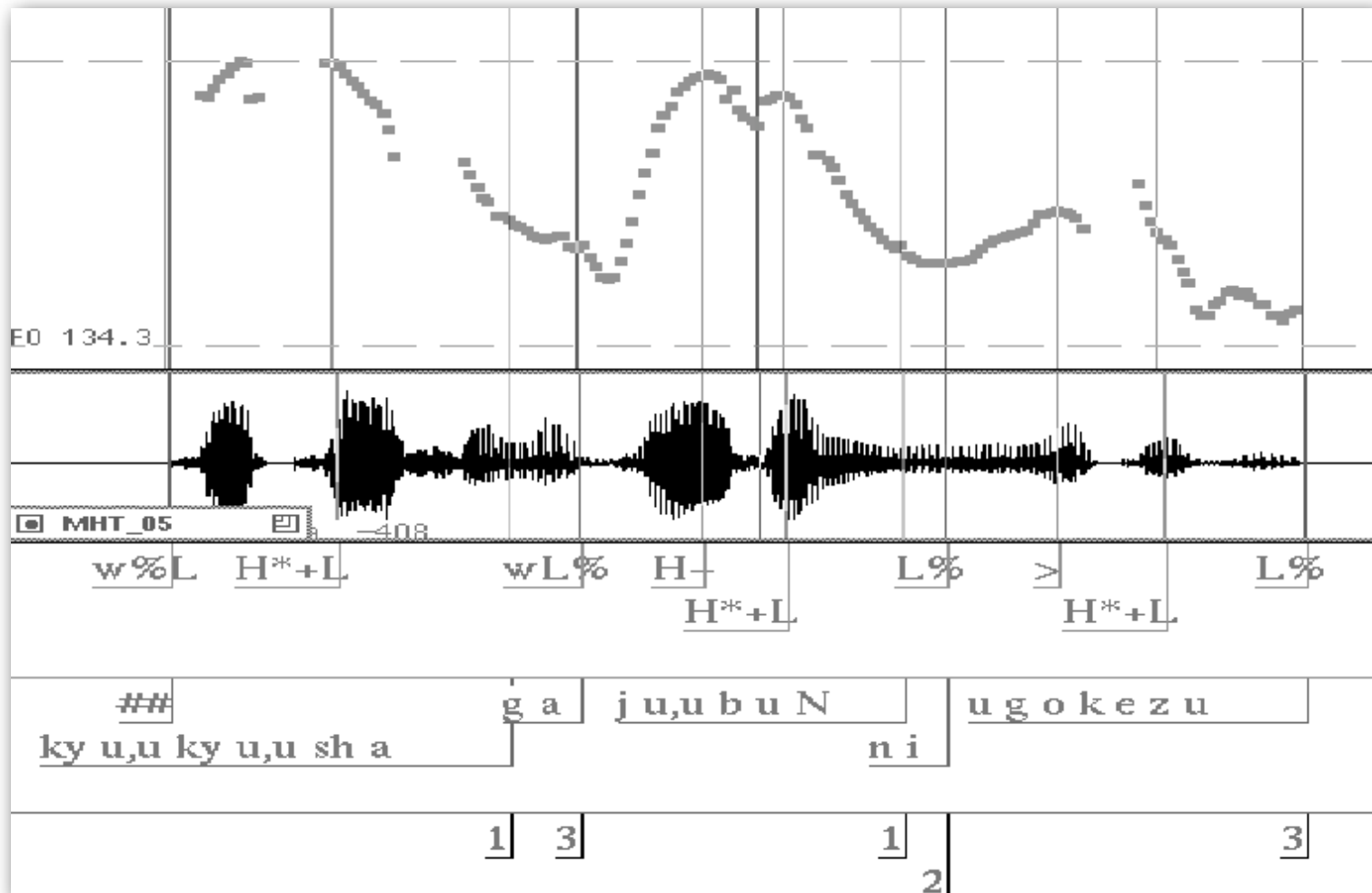
● 例えば欲しい高さの音がなかった場合どうする?



波形編集方式

音声DBに対するラベリング・素片DB化

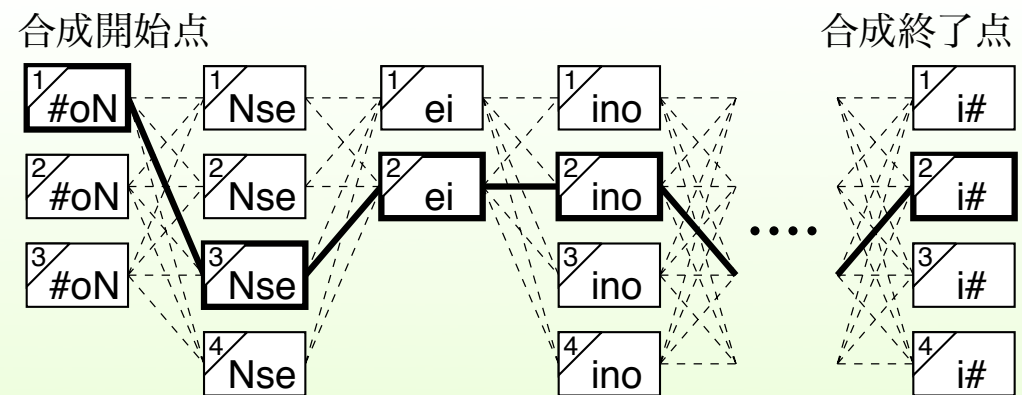
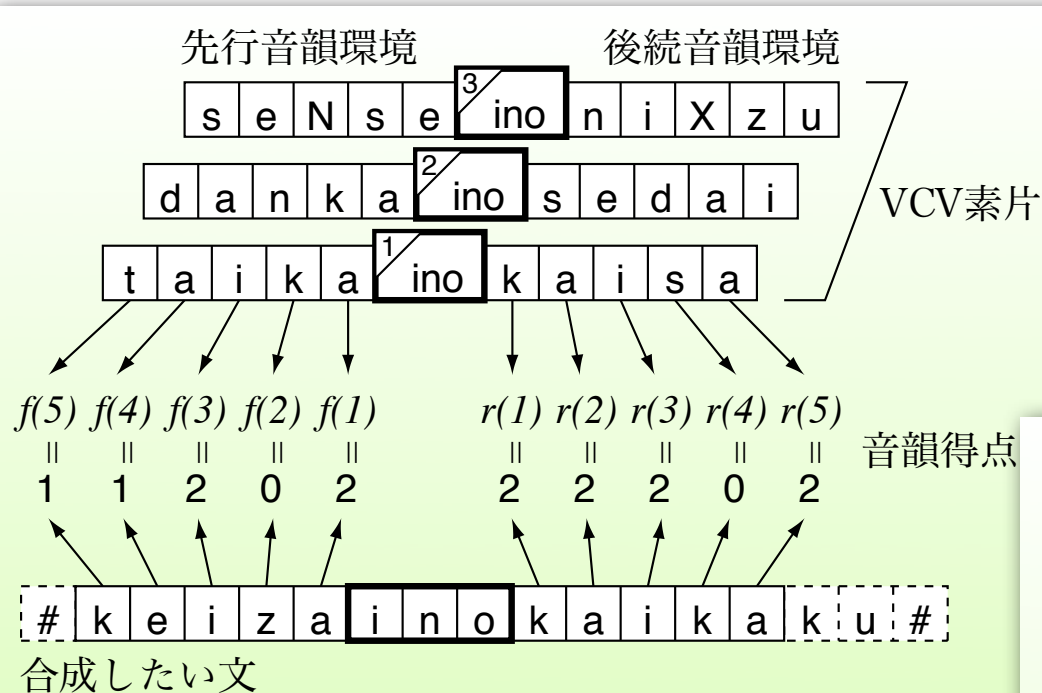
- 音韻ラベリングと韻律ラベリング
- 手動ラベリング／音声認識技術を使った自動ラベリング



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



VCV素片番号
VCV素片 **1/ Nse** — VCV単位名

----- VCV素片の接続経路
(LSP距離の重み付き)

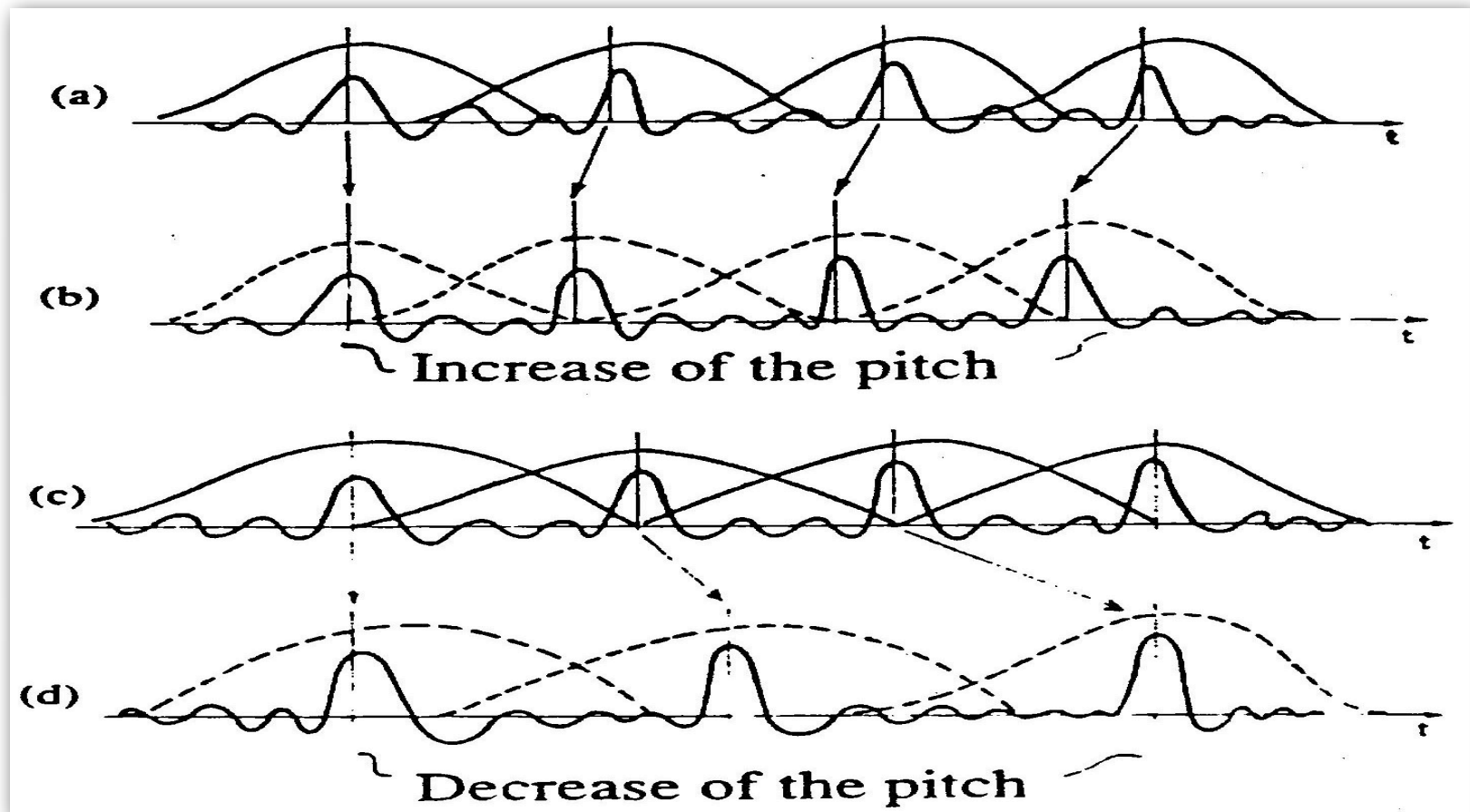
—— LSP距離の総和が最小の経路

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

The screenshot shows the NTT IT Voice Communication website. The header features the NTT IT logo and the text "Voice Communication". A navigation menu on the left lists various services, including "音声合成" (Voice Synthesis) and "音声合成ASP". The main content area displays the breadcrumb "HOME > 製品 > CTI・音声 > 音声合成ソフトウェア" and the title "音声合成ソフトウェア". Below the title, it states "単純音声合成から音声編集ツールまで豊富なラインアップをご提供". A highlighted box contains information about the NTT Research Institute's voice synthesis engine "FLUET", its variants "FinalFluet" and "Cralinet", and their capabilities in handling a wide range of voice data and applications.

NTT IT Voice Communication

HOME > 製品 > CTI・音声 > 音声合成ソフトウェア

音声合成ソフトウェア

単純音声合成から音声編集ツールまで豊富なラインアップをご提供

NTT研究所の音声合成エンジン「FLUET」、
「FinalFluet」、「Cralinet」搭載
豊富な音声データによる滑らかで高品質な音声合成を実現
最新の単語や専門用語、様々な言い回しにも対応
ライブラリ、サーバ、ツールなど用途に応じた製品ラインアップ

HMM合成

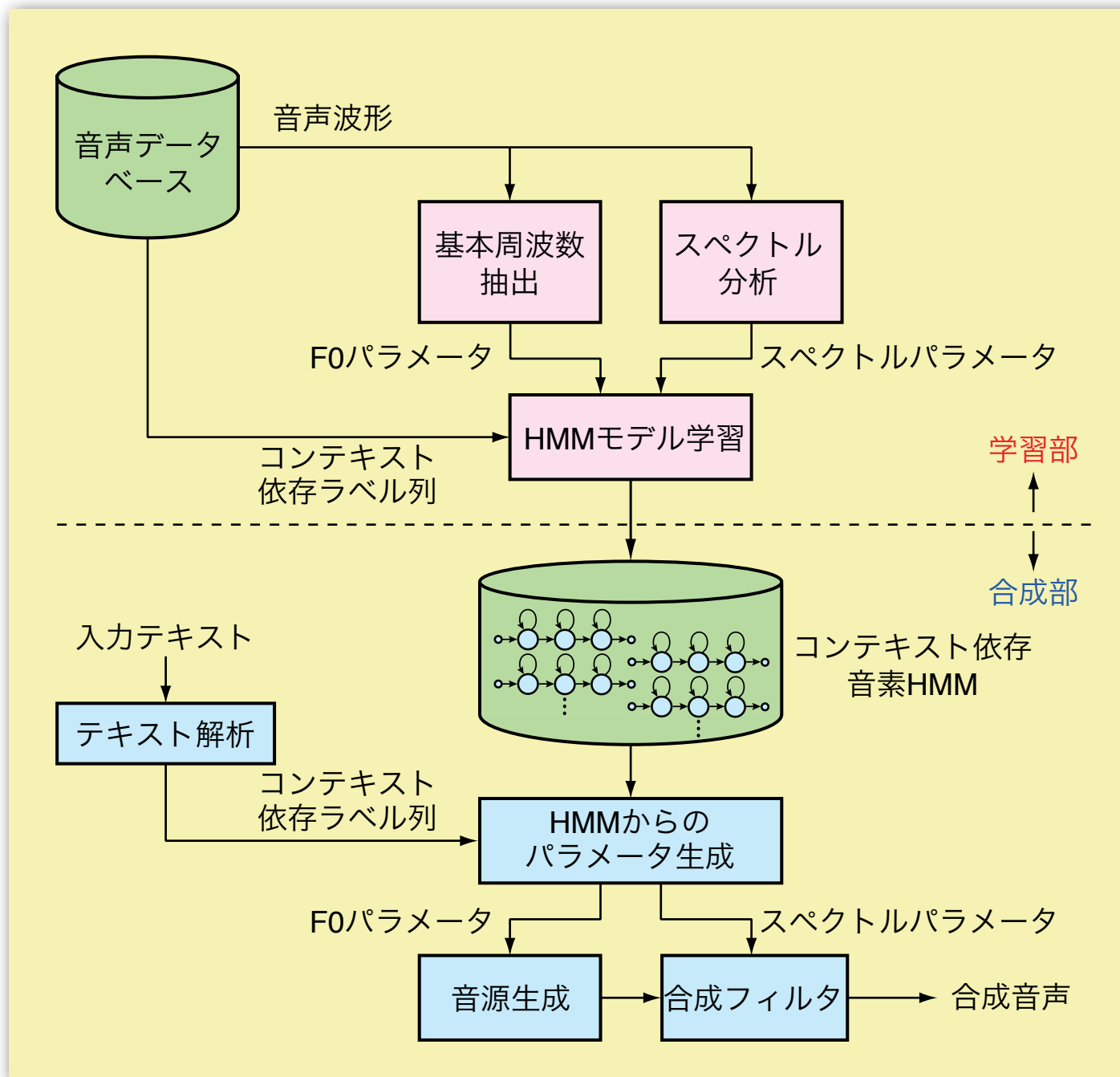
LMAフィルタ

- Log Magnitude Approximation
- ケプストラム係数から直接デジタルフィルタを構成
- 音源（パルス or 白色雑音）+LMAフィルタで音は出る

テキスト→ケプストラム系列の変換

- HMMを用いた解決
- テキスト→HMM状態系列→尤度最大基準でケプストラム系列を生成→音源駆動→音声波形へ
- デルタケプストラムの利用により滑らかなスペクトル系列へ
- F0とスペクトルの同時モデル化を実現
- 比較的少数のデータ（500文ほど）で品質確保
- 話者適応など音声認識系の技術が直接利用可

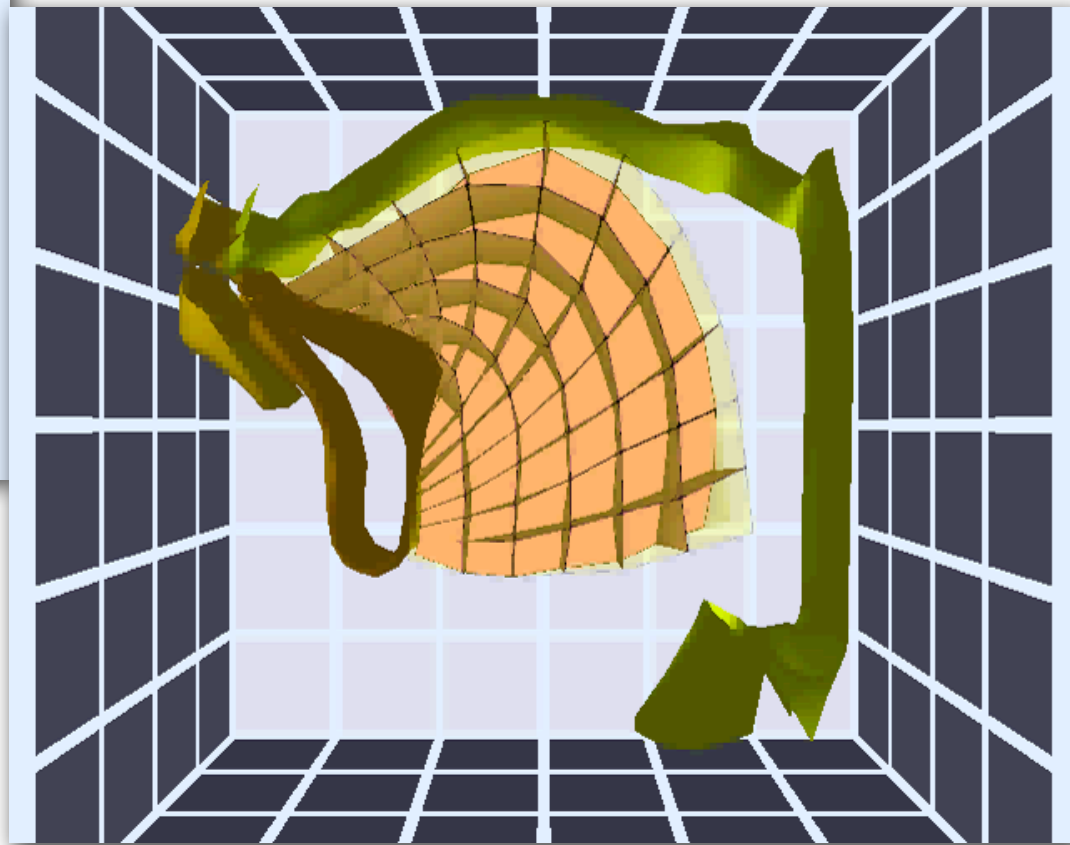
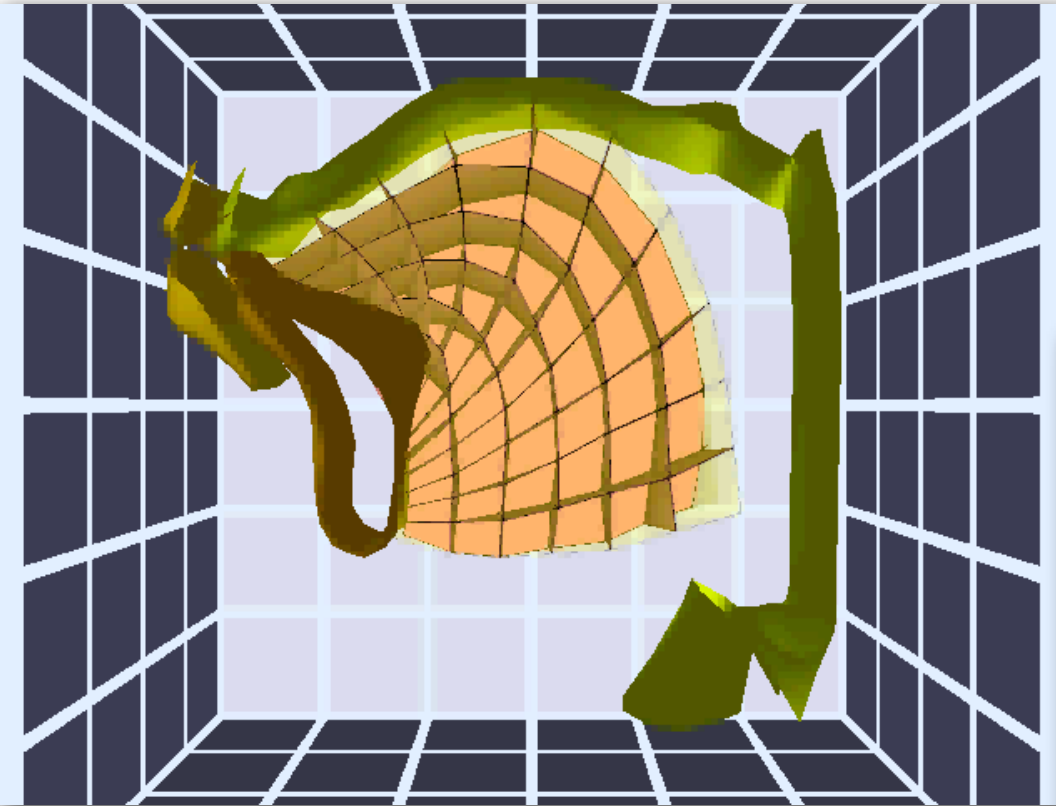
HMM合成



Text-to-Speech ではないけれど

調音音声合成

- 舌の動き，唇の動きをシミュレートして音源を入れる



本日のメニュー

音声認識セミナーより

- 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

テキスト音声合成の原理

- 言語処理／音韻処理／波形生成
- 音韻処理：音素から単音へ／適切な韻律生成に向けて
- 波形生成：波形編集型，分析合成型（HMM音声合成）

まとめ