

# 「大容量時代の音声情報処理手法とその応用」

産業技術総合研究所

知能システム研究部門

知的インターフェース研究グループ

児島 宏明

(知的コミュニティ基盤研究センター 客員研究員)

## 内容

1. 音声情報処理の手法と研究の展開
2. 中間符号系に基づく音声認識と音声検索  
(音声検索のデモビデオ)
3. コーパス型サブバンド方式に基づく音声合成と音声対話  
(音声対話エージェントシステムのデモビデオ)

# 音声情報処理手法の歴史的な流れ

少数の標準パターン

統計的モデル推定

事例ベース(?)

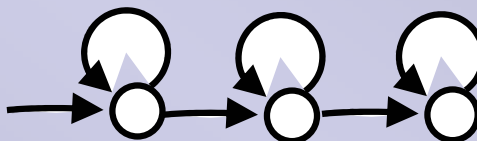
音声認識

DPマッチング  
(DTW)

HMM

中間符号系

?



パラメータ  
合成方式

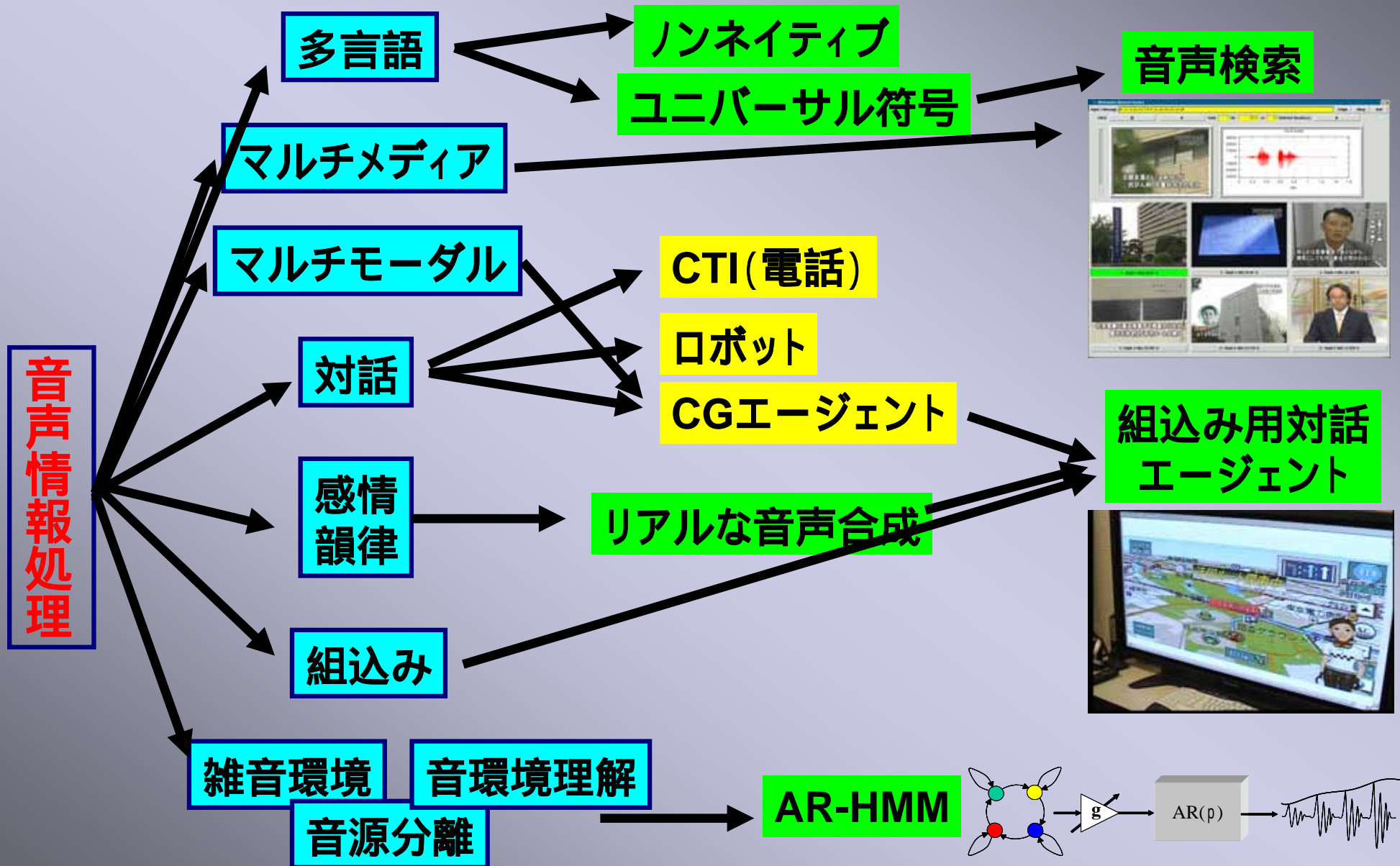
HMM合  
成方式

素片合成方式

コーパスベース方式

音声合成

# 研究の展開



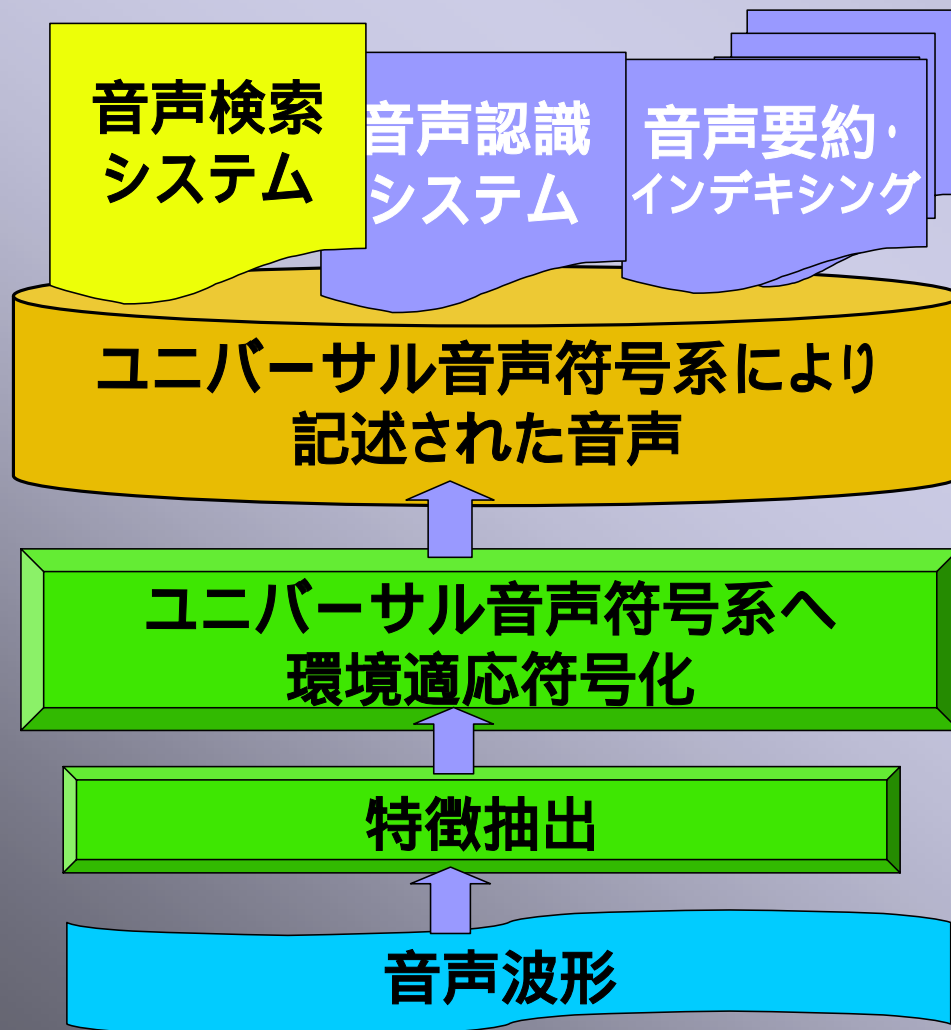
# 中間符号系に基づく音声処理

## 背景

- 現在の(大語彙)音声認識手法
  - 音響モデル: HMM (Hidden Markov Model)
  - 音響モデルの単位: 音素(前後の環境依存)
  - 言語モデル(文法): 単語連鎖統計
  - 適用範囲: 統計的に標準的な発声、書き言葉、静環境
- 問題点
  - 学習時のサンプルと異なる特徴の入力に対応できない  
(子供、老人、ノンネイティブなど)
  - 学習の際に大量のサンプルが必要  
(音声サンプル、テキストコーパス)
  - モデルを精密にしようとする、学習に必要なサンプルが増大

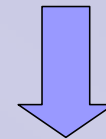
# 中間符号系に基づく音声処理

## 枠組



### ● 音声を記述するための符号 【従来】

- 音響的符号  
(ベクトル量子化、圧縮符号)
- 音韻的記号  
(音声記号、発音記号)



### 【提案法】

- 中間的符号  
(ユニバーサル音声符号系)

# 中間符号系に基づく音声処理

## 特徴

### ● 利点

- 認識や検索におけるマッチングの際の処理効率が向上。
- 非母語話者や老人・幼児などへの適応が容易。
- データベースの整備が充分でない少数派言語への対応が容易。

### ● 問題点

- 最適な符号系の決定手法が確立されていない。
- 特定の言語や環境に特化して学習したモデルと比べて精度が低下する。

# 中間符号系に基づく音声処理

## 記述例(英語)

[i] 英文サンプル from TIMIT-DB

*She had your dark suit in greasy wash water all year.*

[ii] 上の英文音声サンプルの発音表記(XSAMPA 記号系)

(TIMIT-DBで専門家が付与したもの)

*h# S i h E dcl dZ @ r dcl d A kcl k s u q N gcl g r i z i w  
OO S PAU w OO dcl d @ q OO l j I @ r h#*

[iii] 上の表記を提案表記(SPS系列)に規則で変換

*h# #S SS Si ii ih hh hE EE EdZ dcl ddZ dZ@ @@ @r rr rd dcl dd dA AA Ak kcl  
kk ks ss su uu uq qq qN NN Ng gcl gg gr rr ri ii iz zz zi ii iw ww wO OOO OS SS  
S# PAU #w ww wO OOO Od dcl dd d@ @@ @q qq qO OOO Ol ll lj jj jI II I@  
@@ @r rr r# h#*

[iv] SPS系列へ自動符号化した結果(自動ラベリング)

*h# #S SS Si ii ih hh hE EE EdZ dcl ddZ dZi ii id dcl dd dA AA Ak kcl kk ks ss si  
ii iI II IN NN Ng gcl gg gr rr ri ii iz zz zs ss sI II Iw ww wO OOO OS SS Sw ww  
wO OOO Od dcl dd dA AA A@ @@ @q qq qO OOO Ol ll lj jj jI II Ir rr r# h#*

# 中間符号系に基づく音声処理

## 符号系の生成手法

### ■ トップダウンな生成手法

- IPA (国際音声記号)などを基にして、これを細分化・精密化する。

### ■ ボトムアップな生成手法

- 音声サンプルから、統計処理や機械学習により帰納的に符号系を生成する。

### ■ 両者を融合した生成手法

- 調音的特徴を利用して、音韻的制約を加えた上で、そのパラメータをサンプルから学習。



# 中間符号系に基づく音声処理

## 動機:

- 音韻体系や音声モデルを予め与えるのではなく、音声対話を通じて自動的に形成された符号系が最適  
(幼児が音声言語を習得する過程をシミュレート)

## タスクの設定:

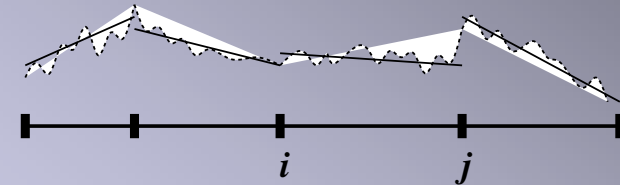
- 音声サンプルの音韻表記による内容を教えず、単語の語彙上の区別のみを与えて、認識率を評価関数として、音声単位を決定。  
(音韻体系の自動形成タスク)

## 実装方式:

- 区分線形セグメントラティスモデル  
(学習や構造の変更が容易なモデルの枠組み)

# ボトムアップな音声セグメントの生成

$$\hat{y}(t) = a(i, j) (x(t) - \bar{x}(t)) + b(i, j)$$



セグメントモデルのパラメータ：

ケプストラムの線形回帰係数

分割手法：

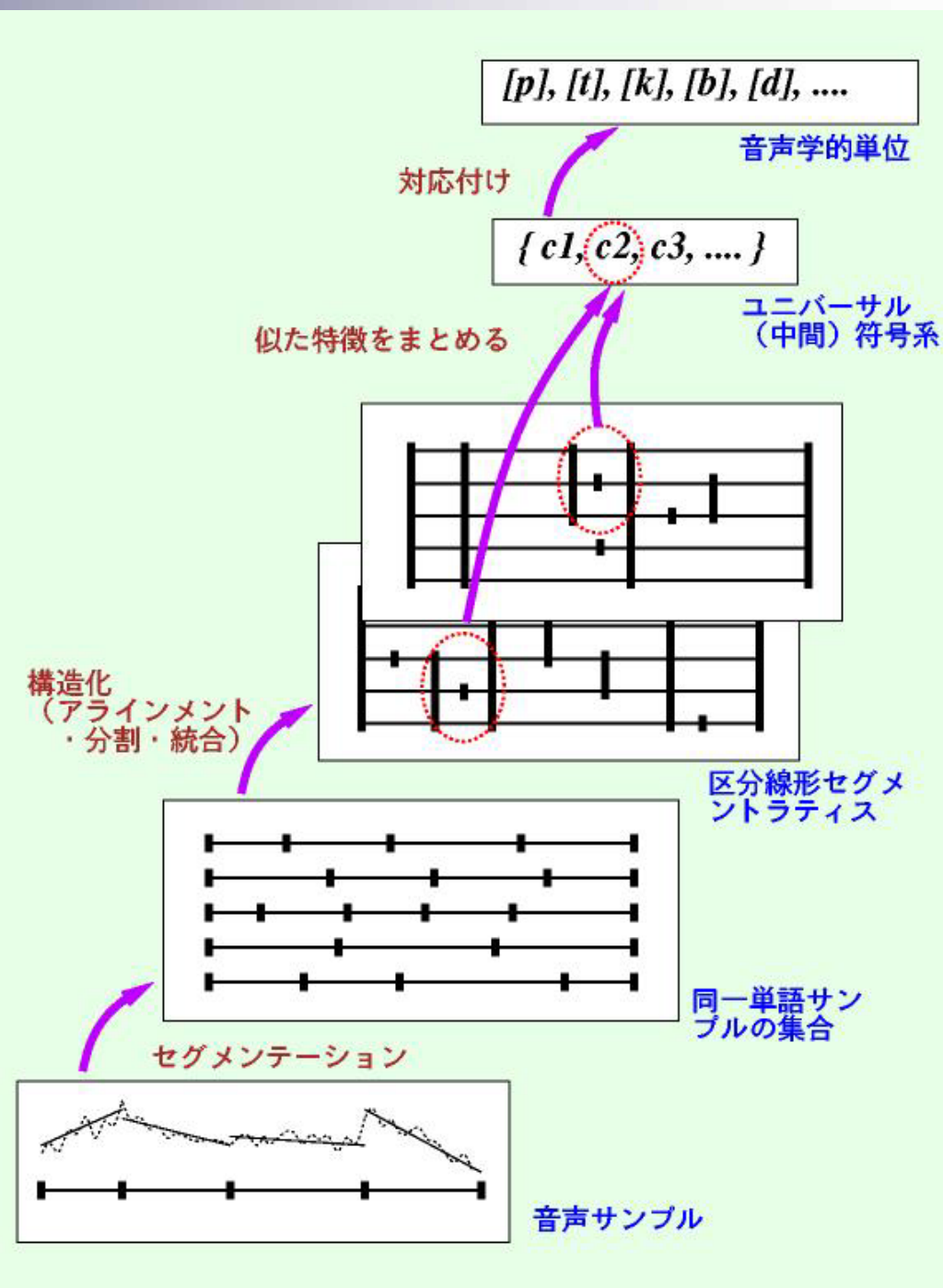
回帰の近似誤差を最小化するような分割  
(DPにより効率的に計算可能)

分割数の決定：

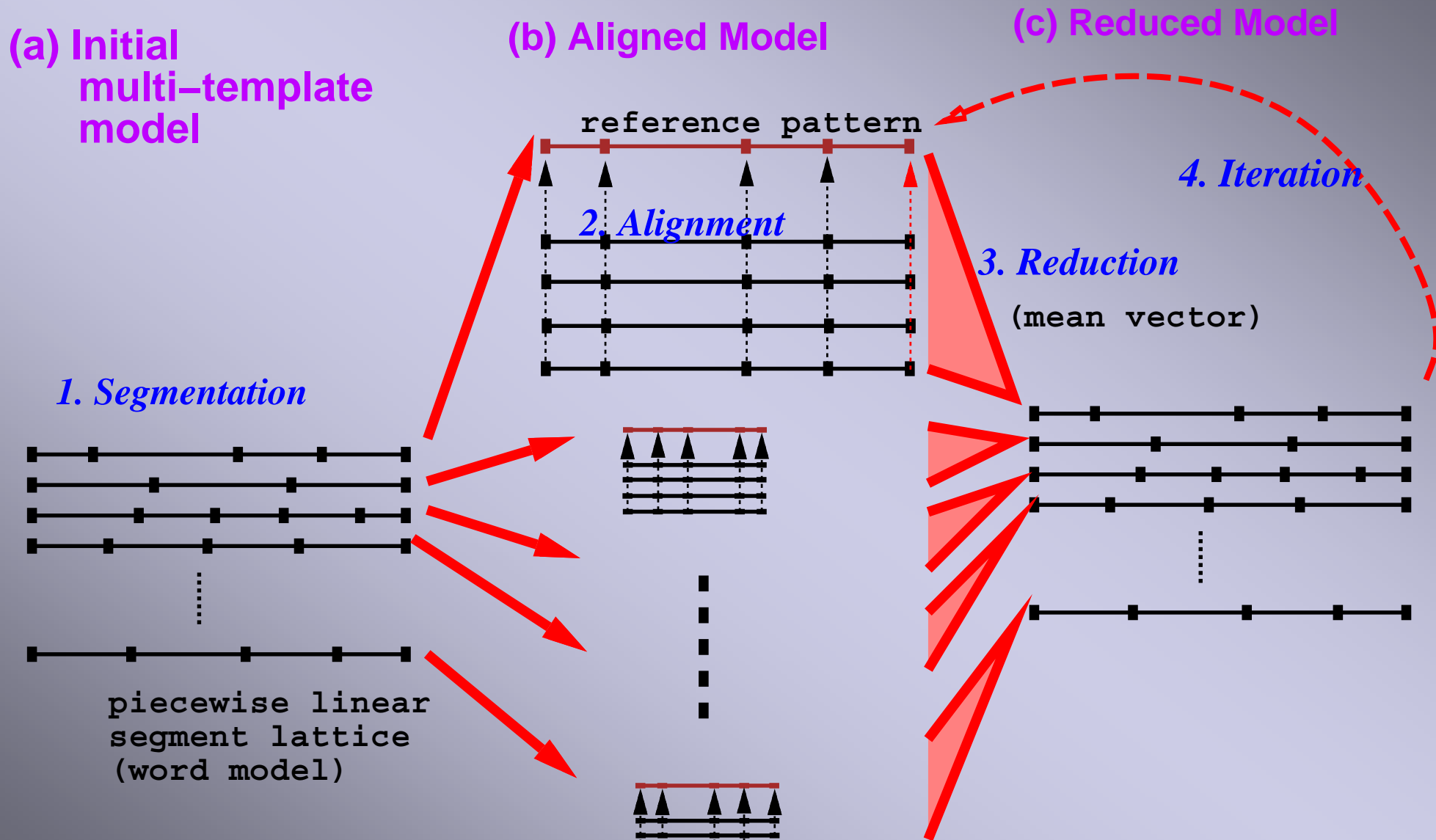
MDL規準に基づいて決定

$$L_{MDL} = g(N, T) + \alpha \cdot N \cdot \beta \cdot \log T$$

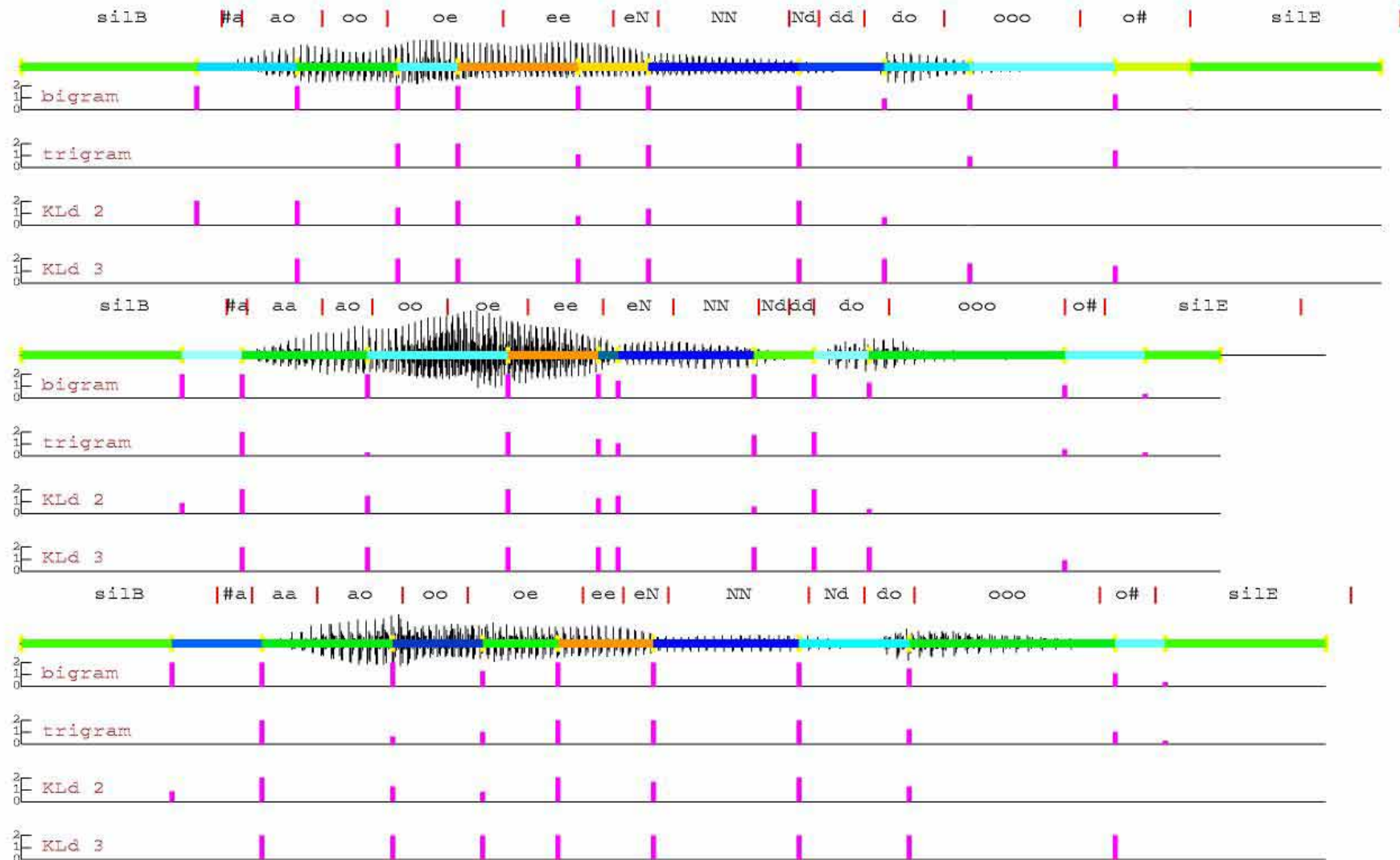
# ボトムアップな音声 符号系の生成



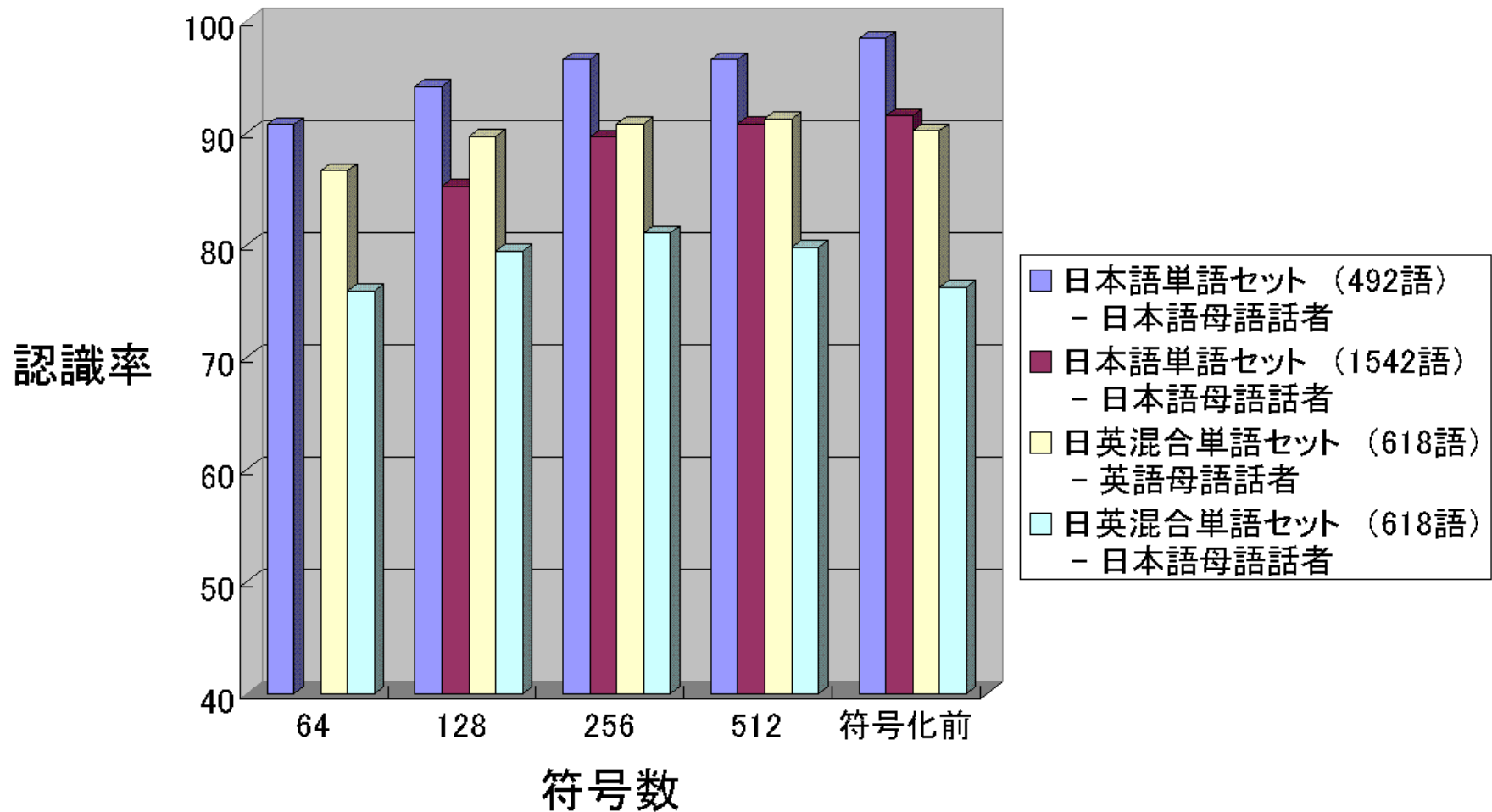
# ボトムアップな音声符号系の生成手法



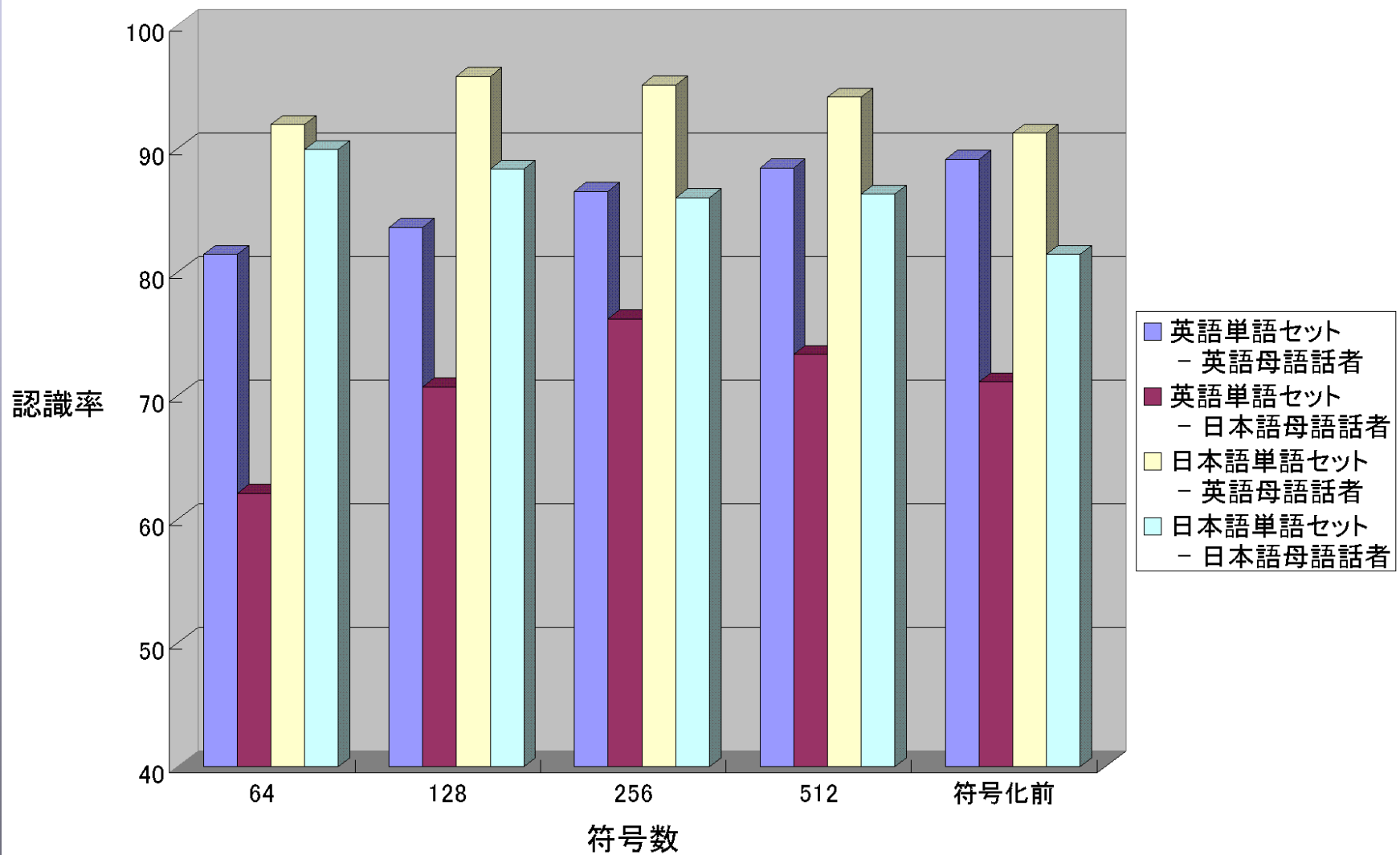
# ボトムアップな音声符号系の生成例



# 認識結果（日本語・日英混合音声）



# 認識結果の内訳（日英混合音声）



# ユニバーサル符号系に基づく音声検索システム

## 【音声検索】

音声により音声を含むコンテンツを検索する。

## 【手法の分類】

- 音声コンテンツ(検索対象)とキーワード(検索語)を音声認識によりテキスト化し、テキスト検索を行う。
- 音声信号のレベルでマッチングを行う。
- 検索対象も検索語も、中間的な記号に変換し、記号レベルでマッチングを行う。

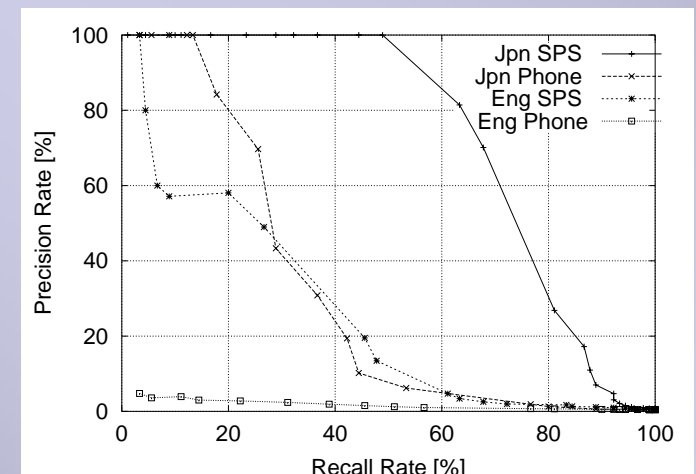
採用した手法

## 【本手法の特長】

- 話者やネイティブ言語による音声の特徴の違いを、記号レベルで吸収できる。
- 検索やマッチングの計算効率が高い。
- 辞書に登録されていないような新語や未知語など任意の検索語を入力可能。

## 【応用例】

- ・テレビ放送の検索
- ・ビデオライブラリ / ビデオメールの検索
- ・会議の発言内容の検索
- ・博物館などの案内





# 中間符号系に基づく音声検索

## [方式の比較]

- 検索キーもDBも音声認識を行いテキスト検索
  - 認識誤りが致命的（特に検索キー）。
  - 非母語話者や老人・幼児などへの適応が困難。
  - 大量DBの認識処理に計算コストがかかる。
- 検索キーとDBを音響的特徴レベルでマッチング
  - マッチングの処理に計算コストがかかる。
  - 話者や環境による特徴量の違いを吸収できない。



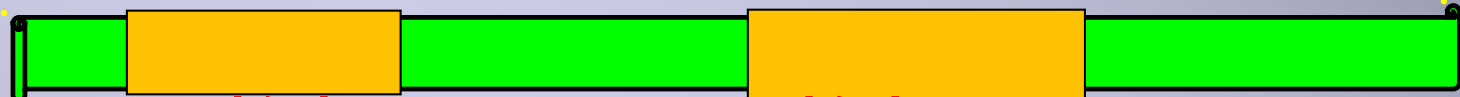
検索キーもDBも中間符号系に変換

# 中間符号系に基づく音声検索

「1週間ばかりニューヨークを取材した」 /iqshuukaNbakari nyuuyookuo shuzaishita/

音声符号列の例: #i, ii, ish, shq, ssh, shu, uk, kcl, kk, ka, aN, Nb, ba, aa, ak, kcl, kk, ka, ar, ri, ii, i#, sil2, sil1, sil1, #n, nn, ny, yu, uy, yy, yo, oo, ok, kcl, kk, ku, uo, oo, osh, shi, iz, zzz, za, aa, ai, ish, ssh, shit, tcl, tt, ta, aa, a#, silE.

DB 側音声



検出

検出

検索キー音声



「1週間ニューヨーク」

ii, ish, ssh, shu, uuu, uk, kcl, kk, ka, aa, aN, NN, N#, sil1, sil2, #n, ny, yu, uuu, uy, yy, yo, ooo, ok, kcl, kk, ku,..

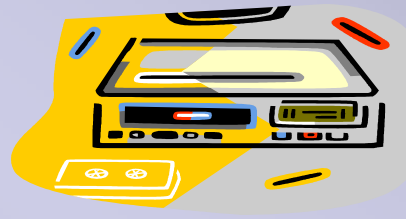
音声符号の任意区間の部分系列同士の最適整合が計算されるので、決まった単語や言い廻しなどの制限は無い。(ShiftCDP法)

# ユニバーサル符号系に基づく音声検索システム

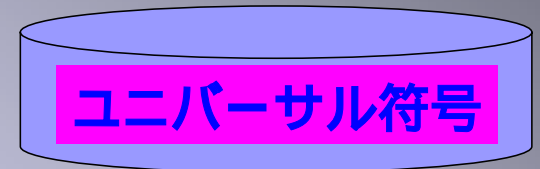
マルチメディアコンテンツ



アーカイブ



データベース



検索システム



検索キーワード



ユニバーサル符号

検索結果

# 音声合成の主な用途と方式

## 用途

- 自動メッセージ(「いらっしゃいませ」)
- 駅(電車)の案内
- 電話による自動応答システム (CTI/IVR)
  - コールセンター
  - ボイスポータル
- 音声対話ロボット
- PCテキストの音声による朗読
  - 視覚障害者用インターフェース

## 方式

- **なめらかで自然な声**
- 単語間の接続が不自然。
- 自由な文を発声できない。

録音再生方式

単語接続方式

素片接続方式

- **自由な文章で合成可能**
- 肉声感の乏しいロボットのな声



# コーパス型サブバンド方式に基づく音声合成

## 本方式の特徴

### ■ コーパス方式

#### □ 録音再生に近いリアルな声

収録した音声のデータベース(コーパス)から、  
文脈的に最適な部分を任意の単位で選択(CHATR)

#### □ (問題点)

- 接続部分が不自然になりやすい。
- データベース用に大規模な装置が必要になり、用途が限られる。



### ■ サブバンド方式



#### □ なめらかな接続

精密なピッチ(基本周期)分析手法

#### □ 高い圧縮効率


ピッチとサブバンド(周波数帯)ごとに正規化

### 合成音声サンプル

女性  子供 

### 圧縮効率の比較

オリジナル 

MP3 (1/30) 

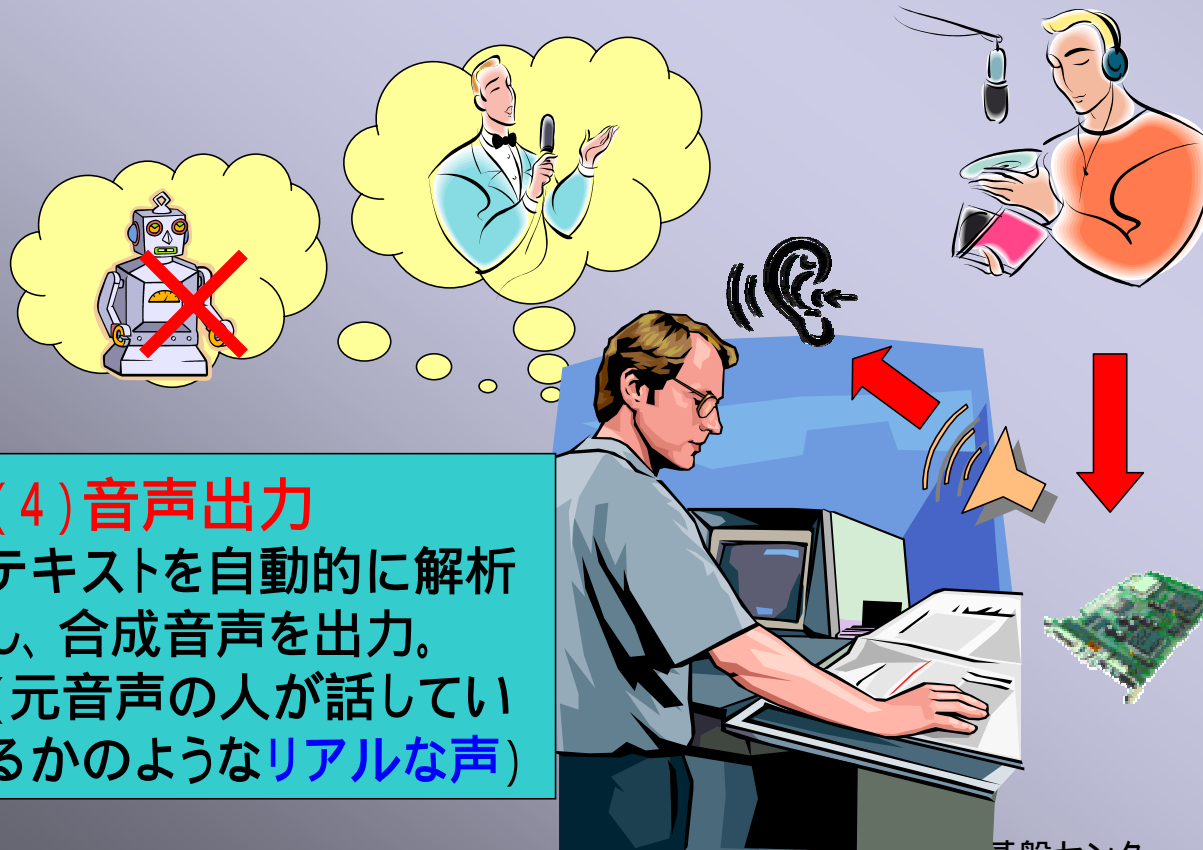
サブバンド (1/30) 

# コーパス型サブバンド方式に基づく音声合成

## 実装の手順

### (1) タスクの決定

用途(タスク)を決定し、それに応じた収録用の文章リストを作成。



### (4) 音声出力

テキストを自動的に解析し、合成音声を出力。  
(元音声の人が話しているかのようなリアルな声)

### (2) 収録

元音声となる人(アナウンサーや声優など)に、予め、いくつかの簡単な文章を朗読してもらう。

**収録時間:** 発声内容がある程度限定できる場合は1時間程度、限定がない場合は5時間程度。ただし収録時間が短い場合でも、つなぎ目やイントネーションが多少不自然になったり、肉声感が減少することはあっても、文章の全体の発声は可能。

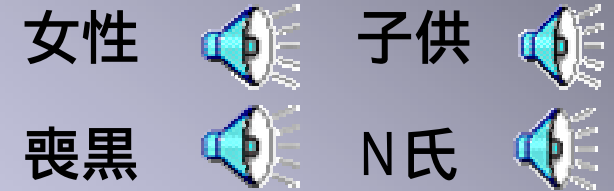
### (3) データベースの作成

収録した音声に音素(ローマ字)ラベルを付ける。(音声認識を利用することにより、部分的に自動化可能)

# コーパス型サブバンド方式に基づく音声合成

## 新たな用途

### 合成音声サンプル



#### ■ 録音再生に近い音質

- 列車やバスの案内(追加や変更が容易)
- 防災放送(正確に聴き取れる)
- カーナビ(聴いて疲れない)

#### ■ 収録話者本人のような音声

- 有名人やアイドルの声による合成 (携帯電話などに配信も可能)
- バーチャル俳優(声優)による映画 (故人の声で新作を)
- 喉頭癌など声帯手術前の本人の声による音声合成装置 (福祉)

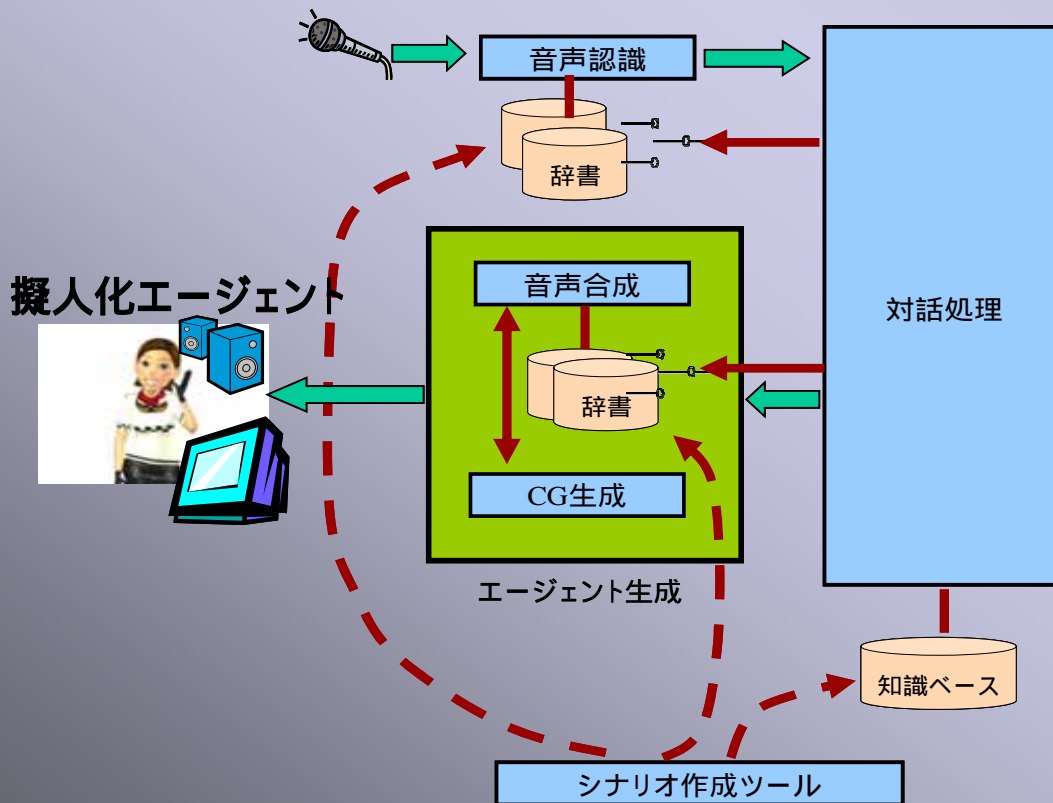
#### ■ リアルな音声

- リアルな音声対話による家電製品等のコントロール
- バーチャルキャラクターとの対話システム



# リアルな音声合成を使った音声対話システム

## システムの構成

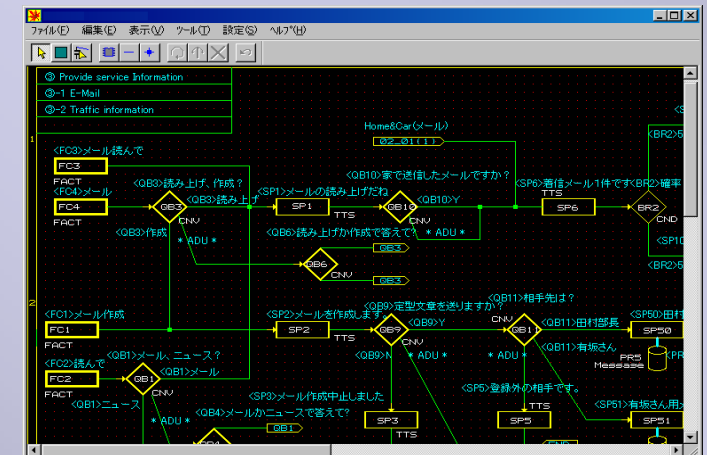


## システム応用例



擬人化エージェント

## シナリオ作成ツール





# リアルな音声合成を使った音声対話システム

## 音声対話エージェントシステム

