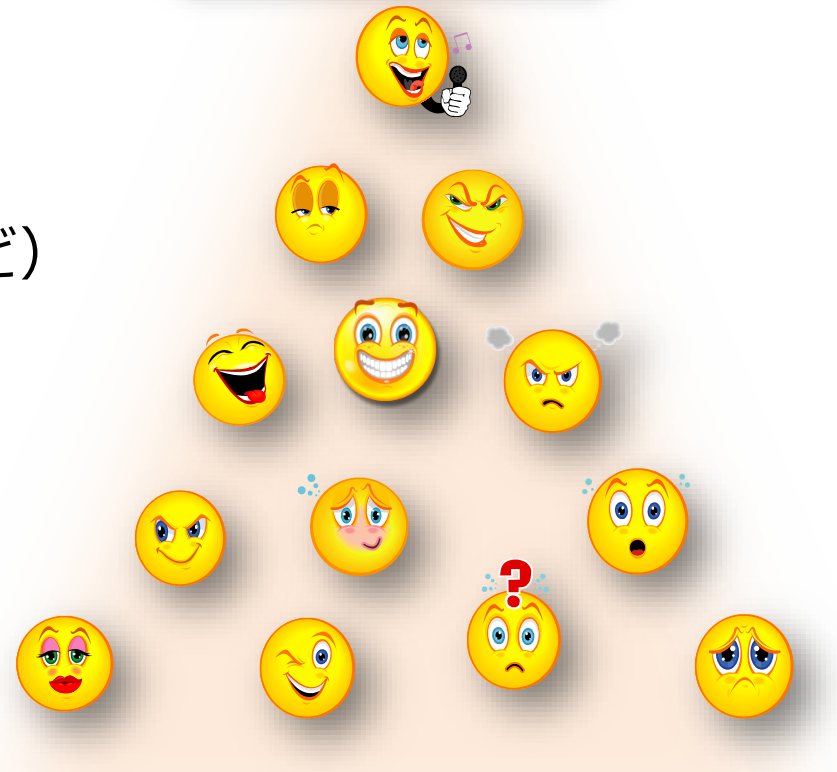
統計的音声合成技術の現在・過去・未来

名古屋工業大学
徳田恵一

人間のように喋る機械の実現

- 音声インタフェースの普及
  1. コミュニケーションチャネルの確立
  2. もっと自然に・魅力的に・快適に！

- 人間のように喋る機械
  - 任意の話者の声質
  - 様々な発話スタイル（読み上げ調・会話調など）
  - 感情表現（楽しそうに・悲しそうになど）
  - 強調
  - その他，様々な非言語情報
  - しかもあらゆる言語で！
  - 更には歌も！！



音声合成の歴史（超簡略版）

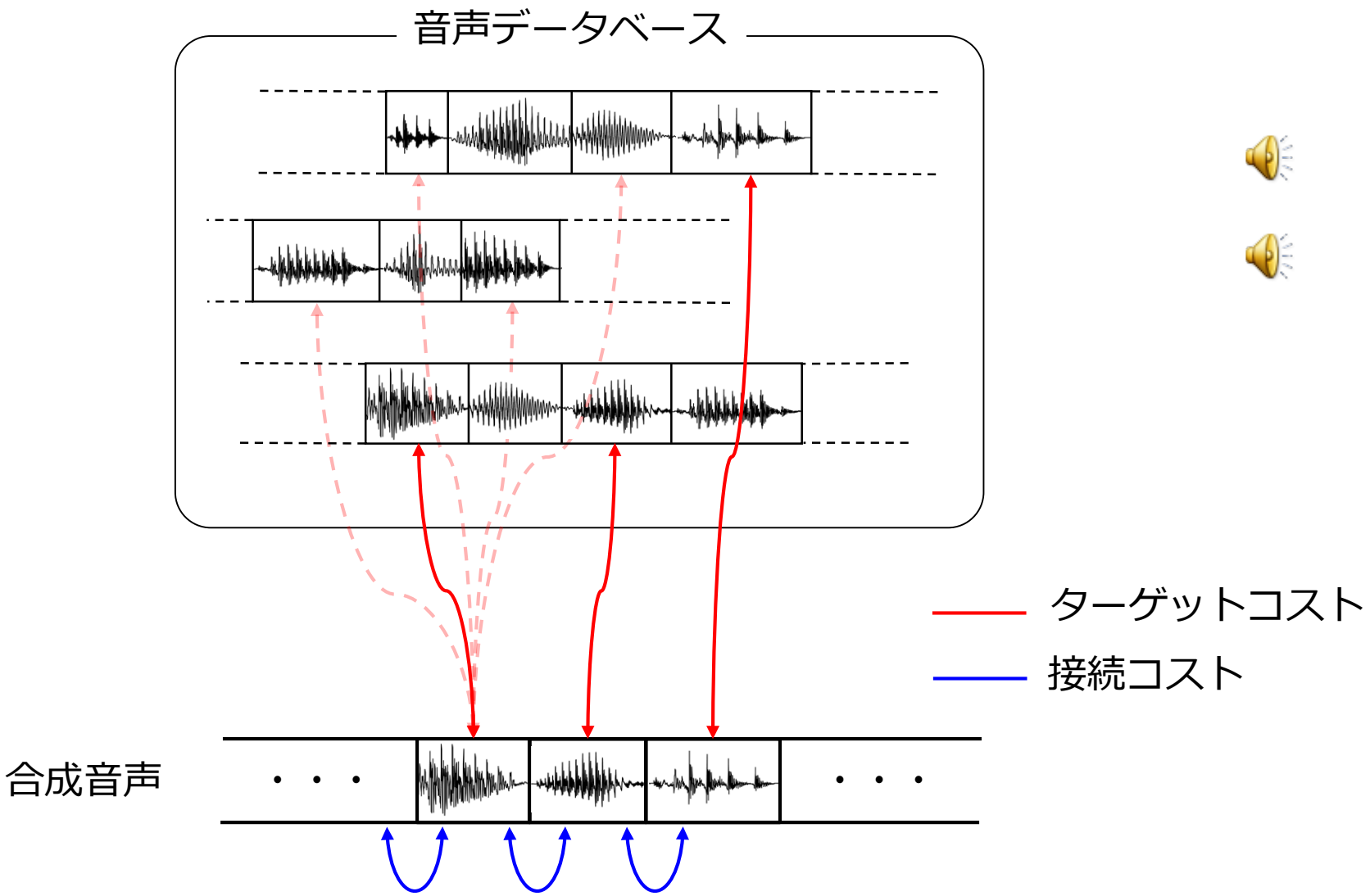
- ルールベース: フォルマント音声合成 (～'90)

- コーパスベース: 波形接続型音声合成 ('90s～)
  - 単一インベントリ: ダイフォン音声合成
  - 複数インベントリ: 単位選択型音声合成

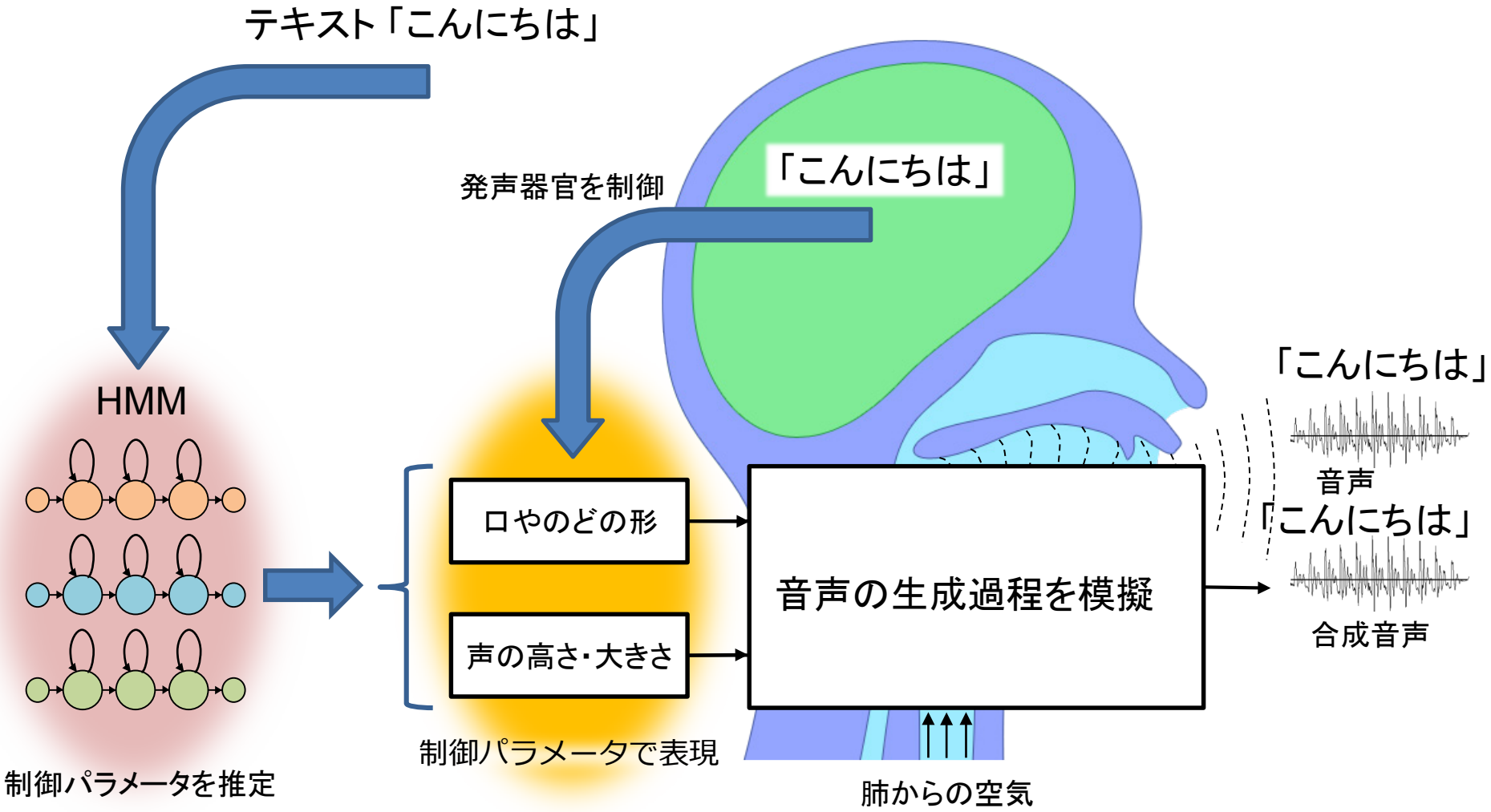
- コーパスベース: 統計的パラメトリック音声合成 ('95～)
  - 隠れマルコフモデルによる音声合成（HMM音声合成）
  - DNN, LSTM等に基づく音声合成

→

単位選択型音声合成



統計的音声合成の概要


テキスト「こんにちは」
HMM
制御パラメータを推定
発声器官を制御
「こんにちは」
口やのどの形
声の高さ・大きさ
制御パラメータで表現
音声の生成過程を模擬
肺からの空気
「こんにちは」
音声
「こんにちは」
合成音声

統計的音声合成の特徴

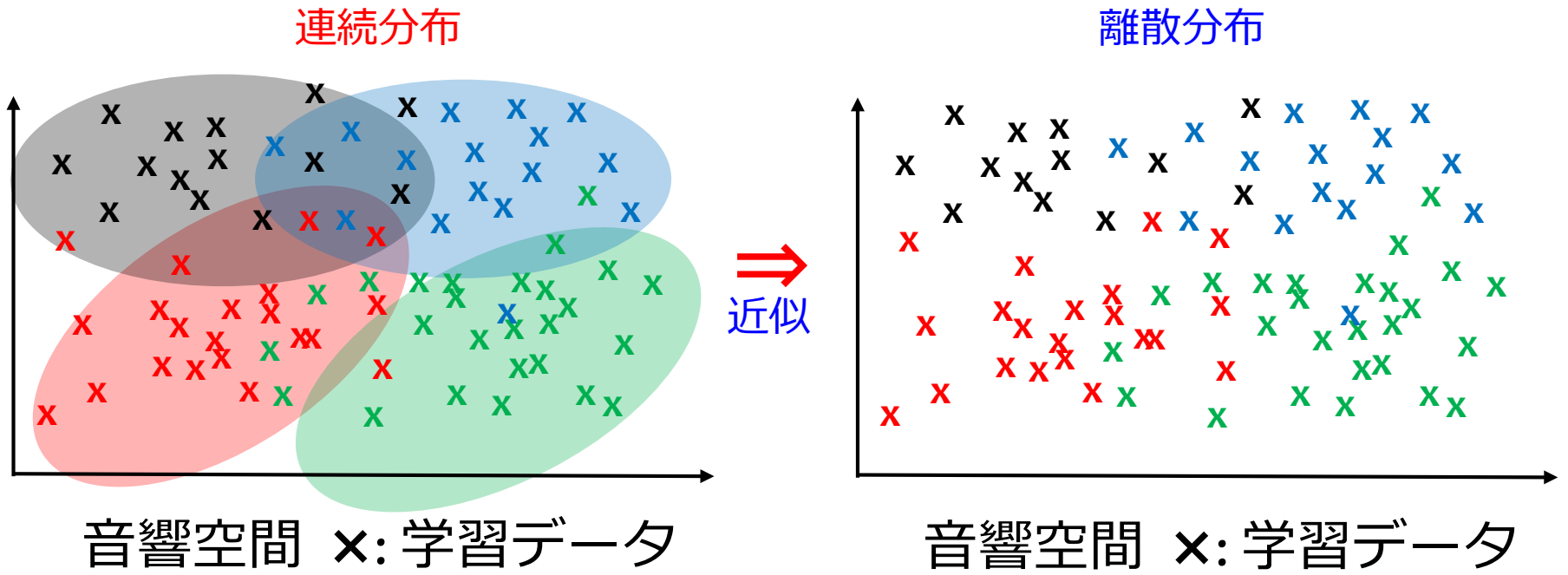
- 自動学習 → システムの自動構築
- 小メモリサイズ → 携帯デバイスでも容易に動作
- 低い言語依存性 → 多言語化が容易

- 柔軟性（話者性，感情表現等々）
  - 声を真似る・声を混ぜる・声をつくる

←

統計 vs. 波形

- 状態出力分布を学習データそのものからなる離散分布で近似



- 動的特徴量付パラメータ生成はフレームベースのDPに帰着
  → 動的特徴量付きパラメータ生成は単位選択の「アナログ版」

本発表のあらまし

- 背景
- 音声合成の基本問題
- 音声合成の課題
- 評価
- 音声合成の社会的役割
- まとめ

余談をはさみながら

音声合成の基本問題 (1/7)

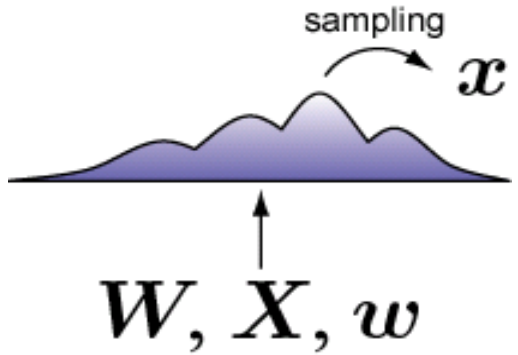
テキストとそれに対応する音声波形の組の集合があるとき，
任意に与えられたテキストに対応する音声波形を求めよ。

- $\boldsymbol{W}$: テキスト
- $\boldsymbol{X}$: 音声波形

音声データベース

既知

- $\boldsymbol{w}$: 任意のテキスト ($\boldsymbol{w} \notin \boldsymbol{W}$)
- $\boldsymbol{x}$: 合成音声波形 ← ?



$$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$$

音声合成の基本問題 (2/7)

- 予測分布の推定は簡単ではない → パラメトリック表現を導入

$$p(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{X},\boldsymbol{W}) = \int p(\boldsymbol{x}|\boldsymbol{w},\lambda)p(\lambda|\boldsymbol{X},\boldsymbol{W})d\lambda$$

λ: モデルパラメータ

音声合成の基本問題 (3/7)

- 補助変数に関する積分は容易ではない → 同時最大化で近似

$$p(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{X},\boldsymbol{W}) = \int p(\boldsymbol{x}|\boldsymbol{w},\lambda)p(\lambda|\boldsymbol{W},\boldsymbol{X})d\lambda \approx p(\boldsymbol{x}|\boldsymbol{w},\hat{\lambda})p(\hat{\lambda}|\boldsymbol{W},\boldsymbol{X})$$

但し $\hat{\lambda} = \arg\max_{\lambda} p(\boldsymbol{x}|\boldsymbol{w},\lambda)p(\lambda|\boldsymbol{W},\boldsymbol{X})$

- 更に $p(\boldsymbol{x}|\boldsymbol{w},\lambda)p(\lambda|\boldsymbol{W},\boldsymbol{X})$ の最大化を $p(\lambda|\boldsymbol{W},\boldsymbol{X})$ の最大化で近似

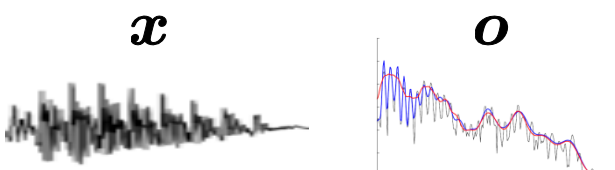
$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W},\boldsymbol{X})$ ← 学習

$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w},\hat{\lambda})$ ← 生成

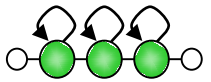
音声合成の基本問題 (4/7)

- 通常，生成モデルは部分モデルに分解される

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$



特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

言語的特徴（ラベル/コンテキスト）

- 当該音素
- 先行・後続音素
- 当該音素のアクセント句でのモーラ位置
- ｛先行，当該，後続｝の品詞，活用形，活用型
- ｛先行，当該，後続｝のアクセント句の長さ，アクセント型
- 当該アクセント句の位置，前後のポーズの有無
- ｛先行，当該，後続｝の呼気段落の長さ
- 当該呼気段落の位置
- 文の長さ

・・・・

膨大な組み合わせ数 ⇒ 全コンテキストについて推定することは困難

音声合成の基本問題 (5/7)

- 生成モデルを部分モデルに分解すると

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W}, \boldsymbol{X}) \leftarrow \text{学習}$$

$$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w}, \hat{\lambda}) \leftarrow \text{生成}$$

$$\{\hat{\lambda}_A, \hat{\lambda}_L\} = \arg\max_{\lambda_A, \lambda_L} \int \sum_{L} p(\boldsymbol{X}|\boldsymbol{O})\, p(\boldsymbol{O}|\boldsymbol{L}, \lambda_A)\, P(\boldsymbol{L}|\boldsymbol{W}, \lambda_L)\, d\boldsymbol{O}\, p(\lambda_A) p(\lambda_L)$$

$$\boldsymbol{x} \sim \int \sum_{l} p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l}, \hat{\lambda}_A)\, P(\boldsymbol{l}|\boldsymbol{w}, \hat{\lambda}_L)\, d\boldsymbol{o}$$

音声合成の基本問題 (6/7)

- 積分と総和を同時最大化で近似

$$\{\hat{\boldsymbol{O}}, \hat{\boldsymbol{L}}, \hat{\lambda}_A, \hat{\lambda}_L\} = \arg\max_{\lambda_A, \lambda_L} \max_{\boldsymbol{O}, \boldsymbol{L}} p(\boldsymbol{X}|\boldsymbol{O})\, p(\boldsymbol{O}|\boldsymbol{L}, \lambda_A) P(\boldsymbol{L}|\boldsymbol{W}, \lambda_L) p(\lambda_A)\, p(\lambda_L)$$

$$\{\hat{\boldsymbol{o}}, \hat{\boldsymbol{l}}\} = \arg\max_{\boldsymbol{o}, \boldsymbol{l}} p(\boldsymbol{x}|\boldsymbol{o})\, p\left(\boldsymbol{o}\middle|\boldsymbol{l}, \hat{\lambda}_A\right) P\left(\boldsymbol{l}\middle|\boldsymbol{w}, \hat{\lambda}_L\right)$$

$$\boldsymbol{x} \sim p(\boldsymbol{x}|\hat{\boldsymbol{o}})$$

音声合成の基本問題 (7/7)

- 同時最大化は容易ではない → 逐次最大化で近似 ← ←

$\hat{\lambda}_L$: 事前学習されたテキスト解析モジュールのパラメータ

$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X}|\boldsymbol{O})$ ← 音声特徴抽出

$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L}|\boldsymbol{W}, \hat{\lambda}_L)$ ← ラベリング

$\hat{\lambda}_A = \arg\max_{\lambda_A} p(\hat{\boldsymbol{O}}|\hat{\boldsymbol{L}}, \lambda_A)p(\lambda_A)$ ← 音響モデリング

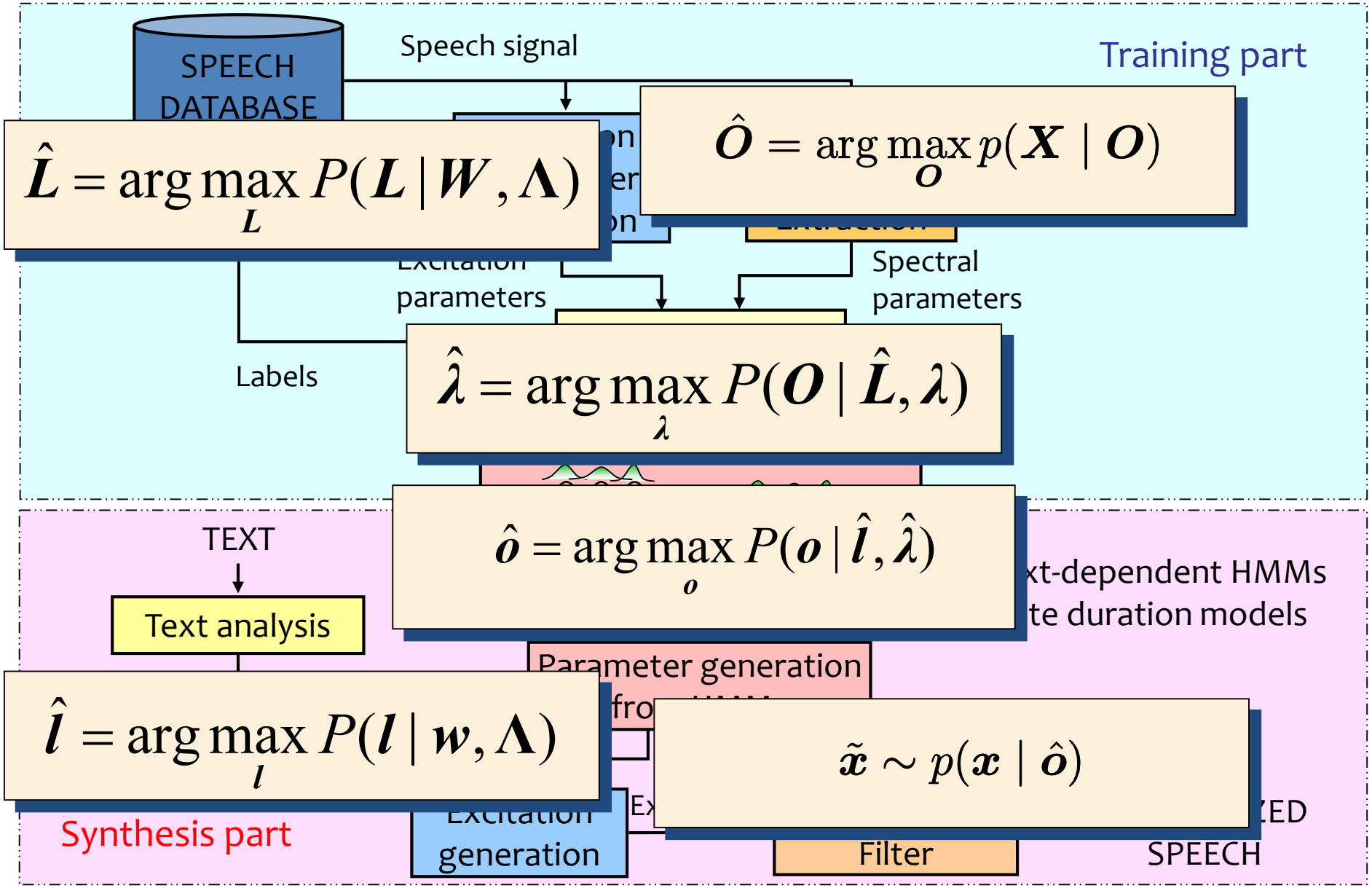
↑ 学習

$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l}|\boldsymbol{w}, \hat{\lambda}_L)$ ← テキスト解析

↓ 合成

$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{o}|\hat{\boldsymbol{l}}, \hat{\lambda}_A)$ ← 音声パラメータ生成

$\boldsymbol{x} \sim p(\boldsymbol{x}|\hat{\boldsymbol{o}})$ ← 音声波形生成

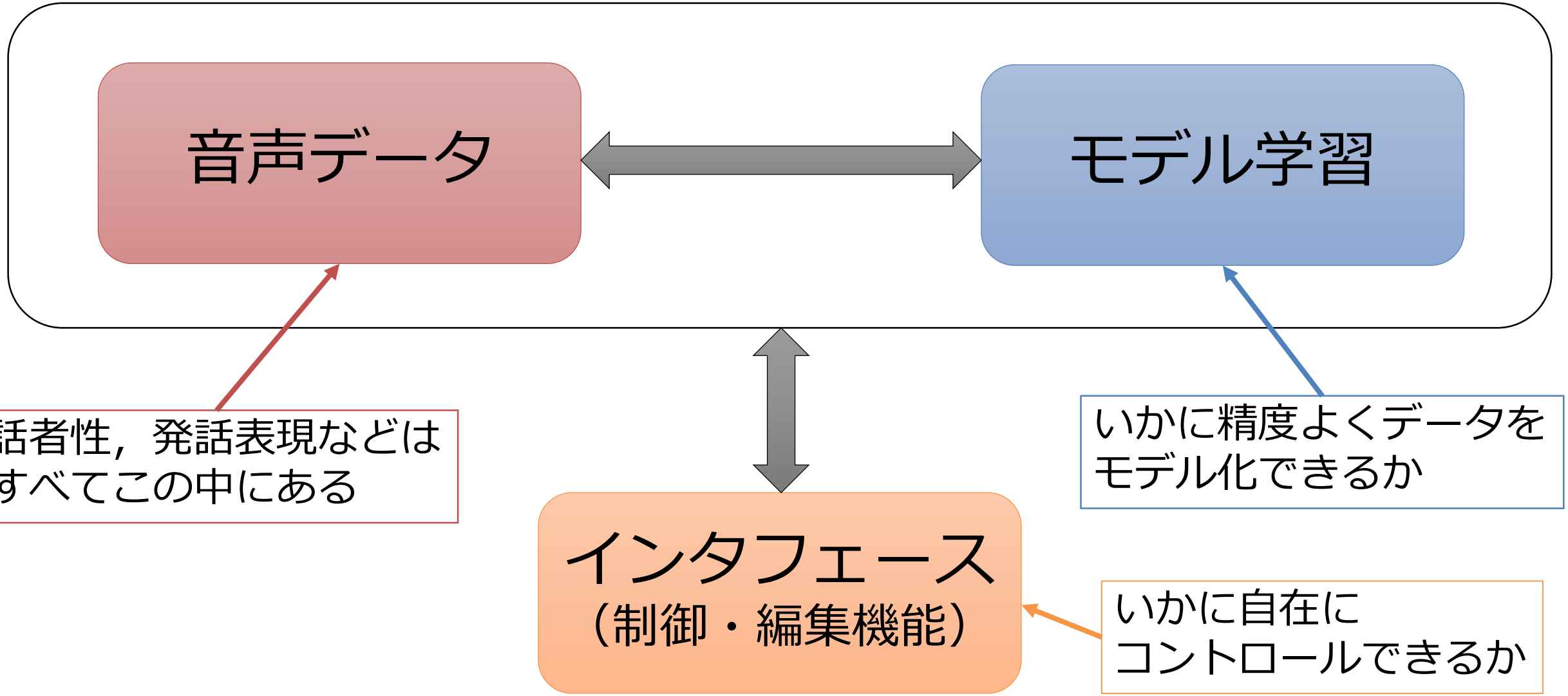

SPEECH DATABASE
Speech signal
Training part
$\hat{L} = \arg\max_{L} P(L \mid W, \Lambda)$
$\hat{O} = \arg\max_{O} p(X \mid O)$
Excitation parameters
Spectral parameters
Labels
$\hat{\lambda} = \arg\max_{\lambda} P(O \mid \hat{L}, \lambda)$
$\hat{o} = \arg\max_{o} P(o \mid \hat{l}, \hat{\lambda})$
TEXT
Text analysis
$\hat{l} = \arg\max_{l} P(l \mid w, \Lambda)$
Synthesis part
Excitation generation
Filter
SPEECH
$\tilde{x} \sim p(x \mid \hat{o})$

本発表のあらまし

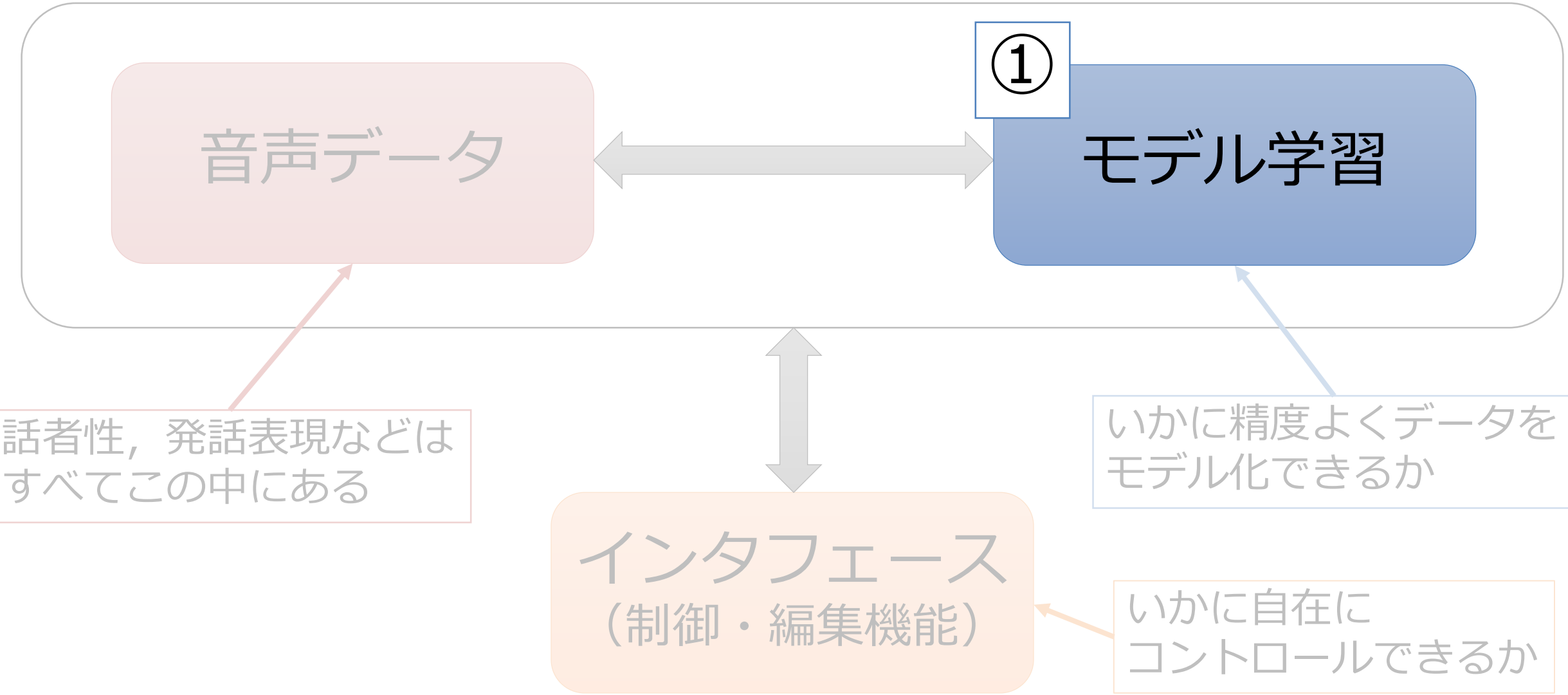
- 背景
- 音声合成の基本問題
- 音声合成の課題
- 評価
- 音声合成の社会的役割
- まとめ

余談をはさみながら

音声合成の課題

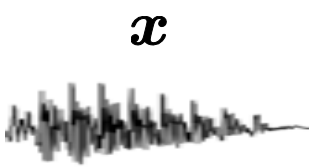

音声データ
モデル学習
話者性，発話表現などは
すべてこの中にある
いかに精度よくデータを
モデル化できるか
インタフェース
（制御・編集機能）
いかに自在に
コントロールできるか

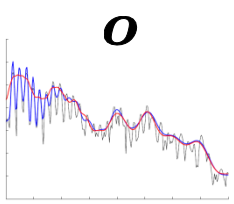
音声合成の課題


①
音声データ
モデル学習
話者性，発話表現などは
すべてこの中にある
いかに精度よくデータを
モデル化できるか
インタフェース
（制御・編集機能）
いかに自在に
コントロールできるか

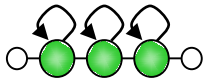
音響モデル

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$





特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

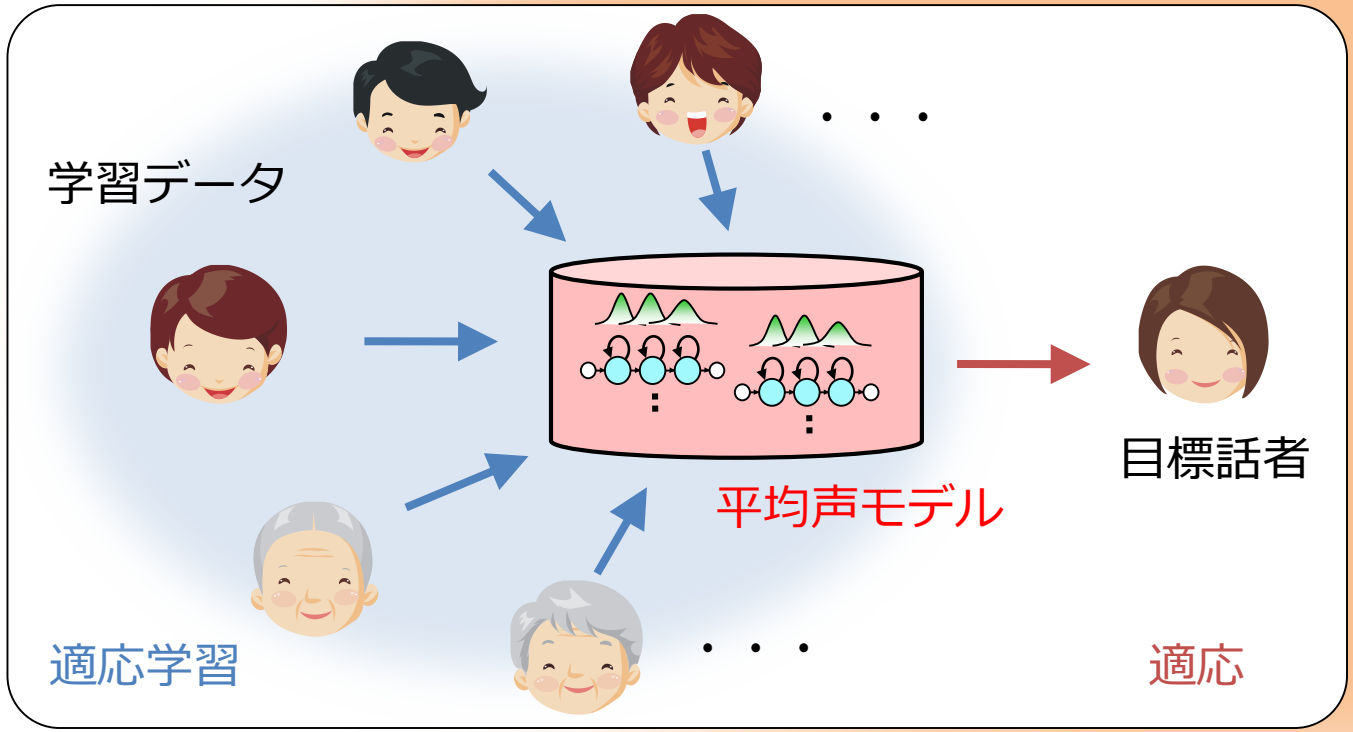
$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

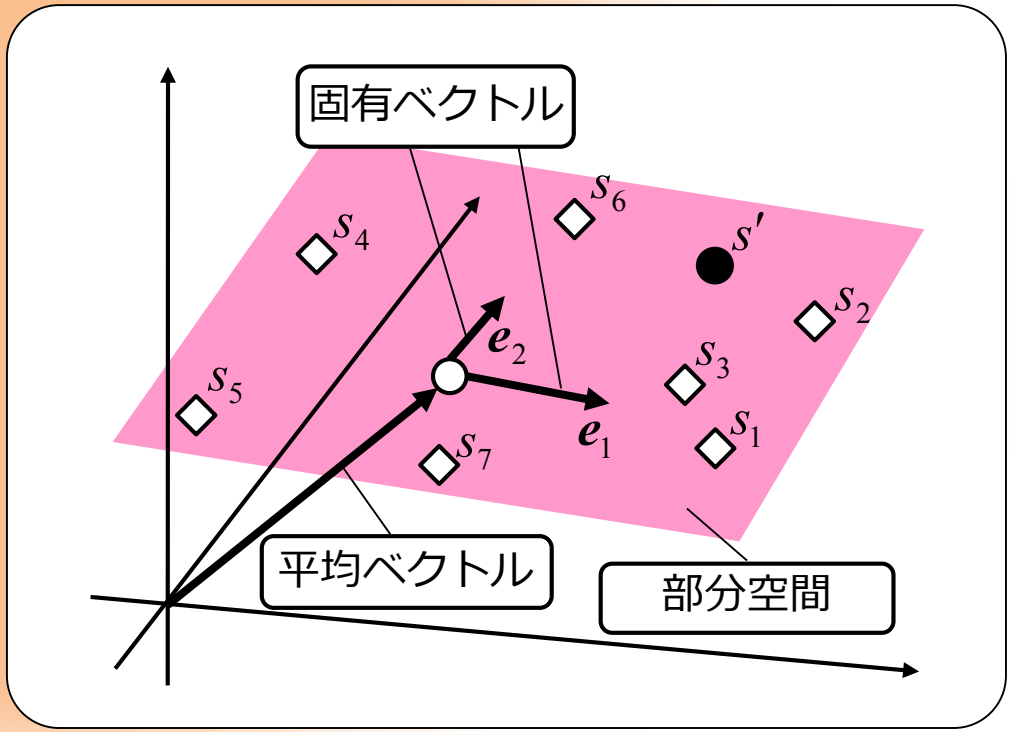
ユニバーサル音声モデル

あらゆる声質，発話スタイル，感情表現，言語等を
自在にモデル化・制御可能な音声モデル

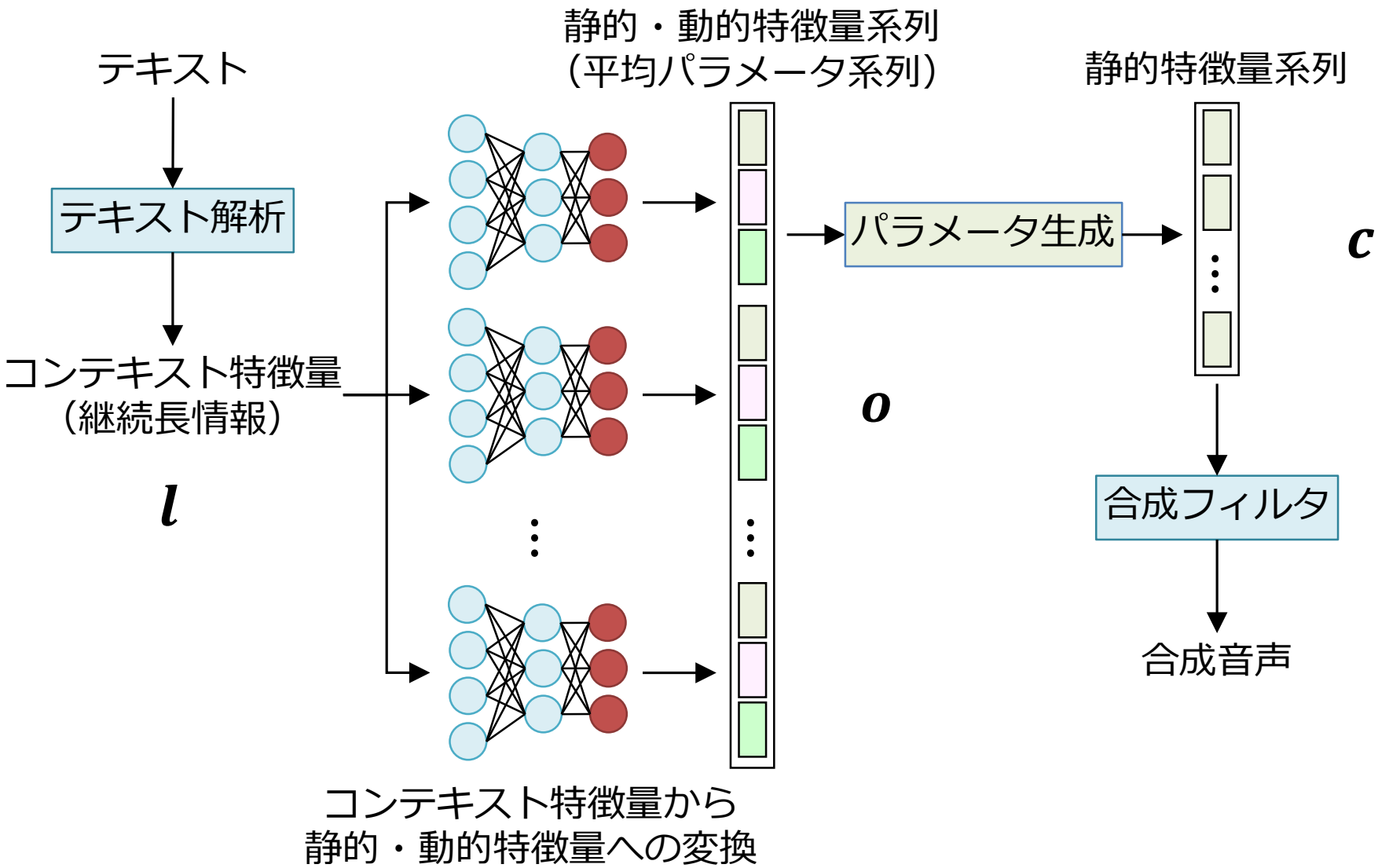
「平均声」



「固有声」 (→)



ニューラネットワークによる音響モデル



DNN vs. HMM

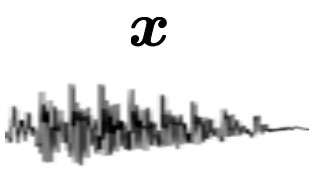
DNN
- データ量比較的大？
- フラットな構造
  - トラブル解決が困難
  - 実装が容易
- 事前情報や知識が初期化や学習手順に埋め込まれる
- 並列・分散処理しやすい
- 連続空間における最適化

HMM
- データ量比較的小？
- 意味付けのある構造
  - トラブル解決が容易
  - 実装が困難
- 事前情報や知識を明確な形で与えやすい
- 並列分散処理しにくい
- 離散空間における最適化

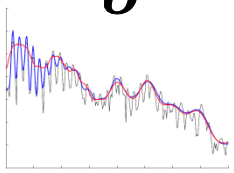
いずれもダイナミクスのモデル化構造は必要そう

特徴抽出・波形生成（ボコーダ）

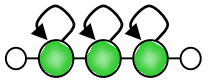
$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$



$\boldsymbol{x}$



$\boldsymbol{o}$



特徴抽出・波形生成
（ボコーダ）

音響モデル

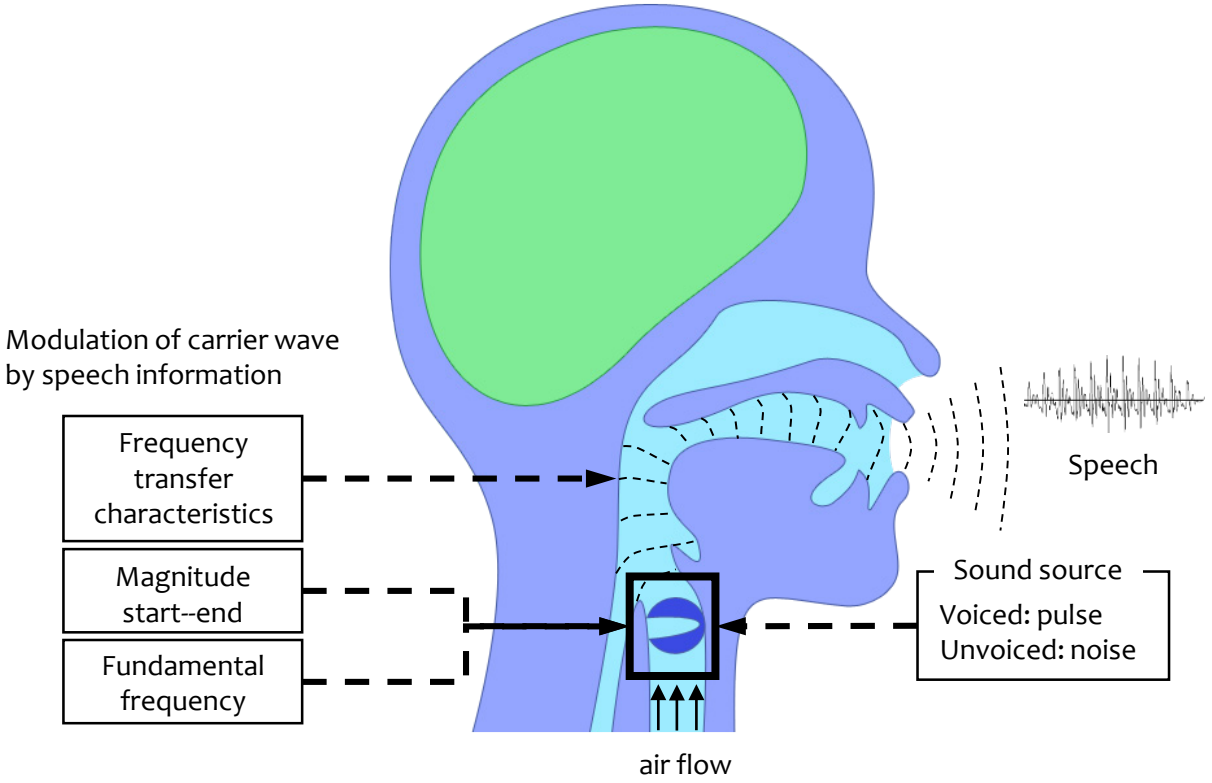
テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

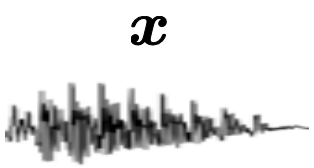
物理的なシミュレーションによる合成

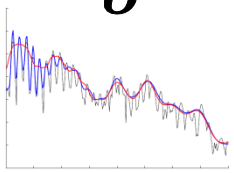
- まだ難しい？
  - 計算機能力・計測技術
  - 解剖学的な知識とモデル
- $p(\boldsymbol{x}|\boldsymbol{o}) \rightarrow p(\boldsymbol{x}|\boldsymbol{o}, \lambda_s)$
  - λs: 物理形状の個人モデル？
  - $\boldsymbol{o}$ は何に対応？
  - 藤崎Foモデル，LFモデルなどとの関係
- ほとんど無限のデータが得られる場合の有効性は？
  - モデルの高度化 vs. データ量
  - データ量のアドバンテージを活かせるシンプルなモデル？



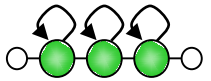
二つのモジュールの再結合

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$





特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

関係する論文

- Log spectral distortion-version of minimum generation error training (MGELSD) [Wu et al., 2009] 1 2 3
- Factor analyzed trajectory HMM (STAVOCO) [Toda et al., 2008] 1 2 3
- Joint estimation of acoustic and excitation models [Maia et al., 2010] 2
- Mel-cepstral analysis-integrated hidden Markov modeling [Nakamura et al., 2014] 1 2 3
- Direct modeling of speech waveforms by neural networks [Tokuda et al., 2015] 3

1. Shifting and overlapping short segments (called 'frame')
   and/or, not measuring likelihoods for speech waveforms directly
2. Fixed decision trees borrowed from standard HMM state clustering
3. Only for voiced or unvoiced sounds

定常的音声信号モデル

- 音声信号 $\boldsymbol{x}$ は往々にして平均ゼロの定常ガウス過程と仮定される

$$p(\boldsymbol{x}|\boldsymbol{c}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{\Sigma_c})$$

$$\boldsymbol{x} = [x(0), x(1), \ldots, x(N-1)]^{\mathrm{T}} \leftarrow \text{フレームに対応}$$

$$\boldsymbol{\Sigma_c} = \begin{bmatrix} \sigma(0) & \sigma(1) & \cdots & \sigma(N-1) \\ \sigma(1) & \sigma(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma(1) \\ \sigma(N-1) & \cdots & \sigma(1) & \sigma(0) \end{bmatrix} \leftarrow \text{共分散 (Toeplitz)}$$

ケプストラム表現

- 自己相関は $\sigma(k)$ はパワースペクトル $\left|H(e^{j\omega})\right|^2$ の逆フーリエ変換で与えられる

$$\sigma(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2 e^{j\omega k}\, d\omega \quad \text{(Wiener–Khinchin theorem)}$$

- ここでは最小位相システム $H(e^{j\omega})$ をケプストラム表現する

$$H\left(e^{j\omega}\right) = \exp\sum_{m=0}^{M} c(m)\, e^{-j\omega m}$$

$\boldsymbol{c} = [c(0), c(1), \ldots, c(M)]^{\mathrm{T}}$ ← ケプストラム

非定常性のモデル化 (1/2)

- 共分散行列 $\boldsymbol{\Sigma_c}$ と精度行列 $\boldsymbol{\Sigma_c^{-1}}$ は，それぞれ $H(e^{j\omega})$ および $H^{-1}(e^{j\omega})$ のインパルス応答によって以下のように表される

$$\boldsymbol{\Sigma_c = H_c H_c^{\mathrm{T}}} \qquad \boldsymbol{\Sigma_c^{-1} = A_c^{\mathrm{T}} A_c}$$

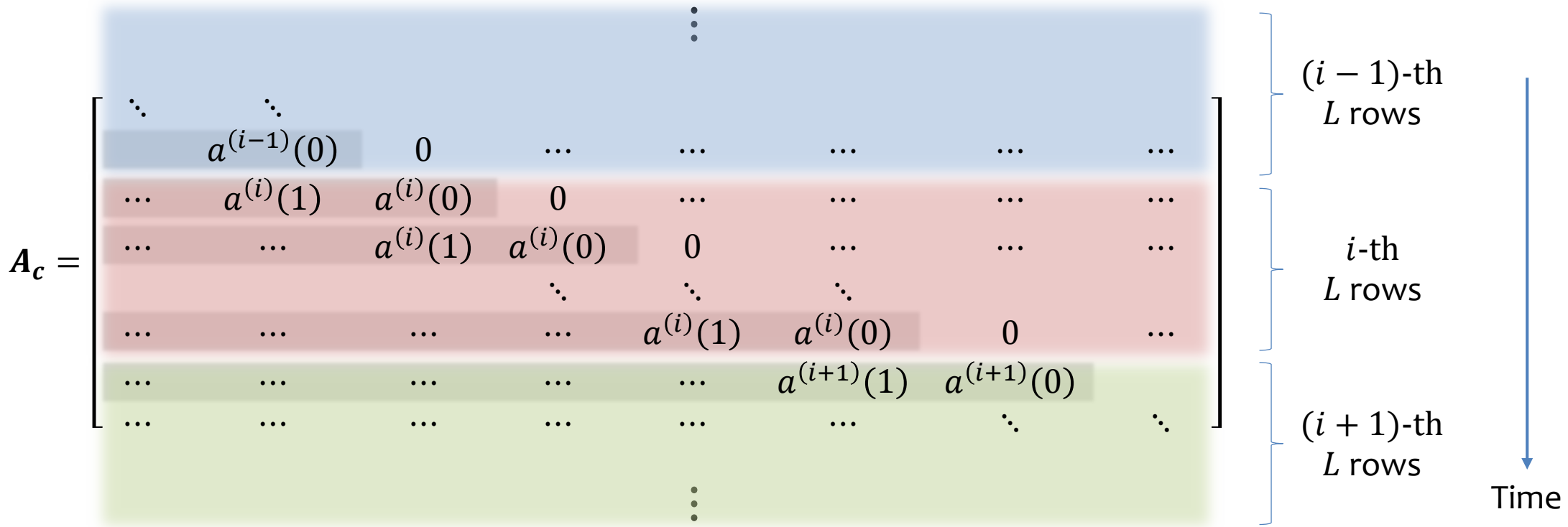
$$\boldsymbol{H_c} = \begin{bmatrix} h(0) & 0 & \cdots & 0 \\ h(1) & h(0) & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h(N-1) & \cdots & h(1) & h(0) \end{bmatrix} \quad \boldsymbol{A_c} = \begin{bmatrix} a(0) & 0 & \cdots & 0 \\ a(1) & a(0) & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a(N-1) & \cdots & a(1) & a(0) \end{bmatrix}$$

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})\, e^{j\omega n}\, d\omega \qquad a(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H^{-1}(e^{j\omega}) e^{j\omega n}\, d\omega$$

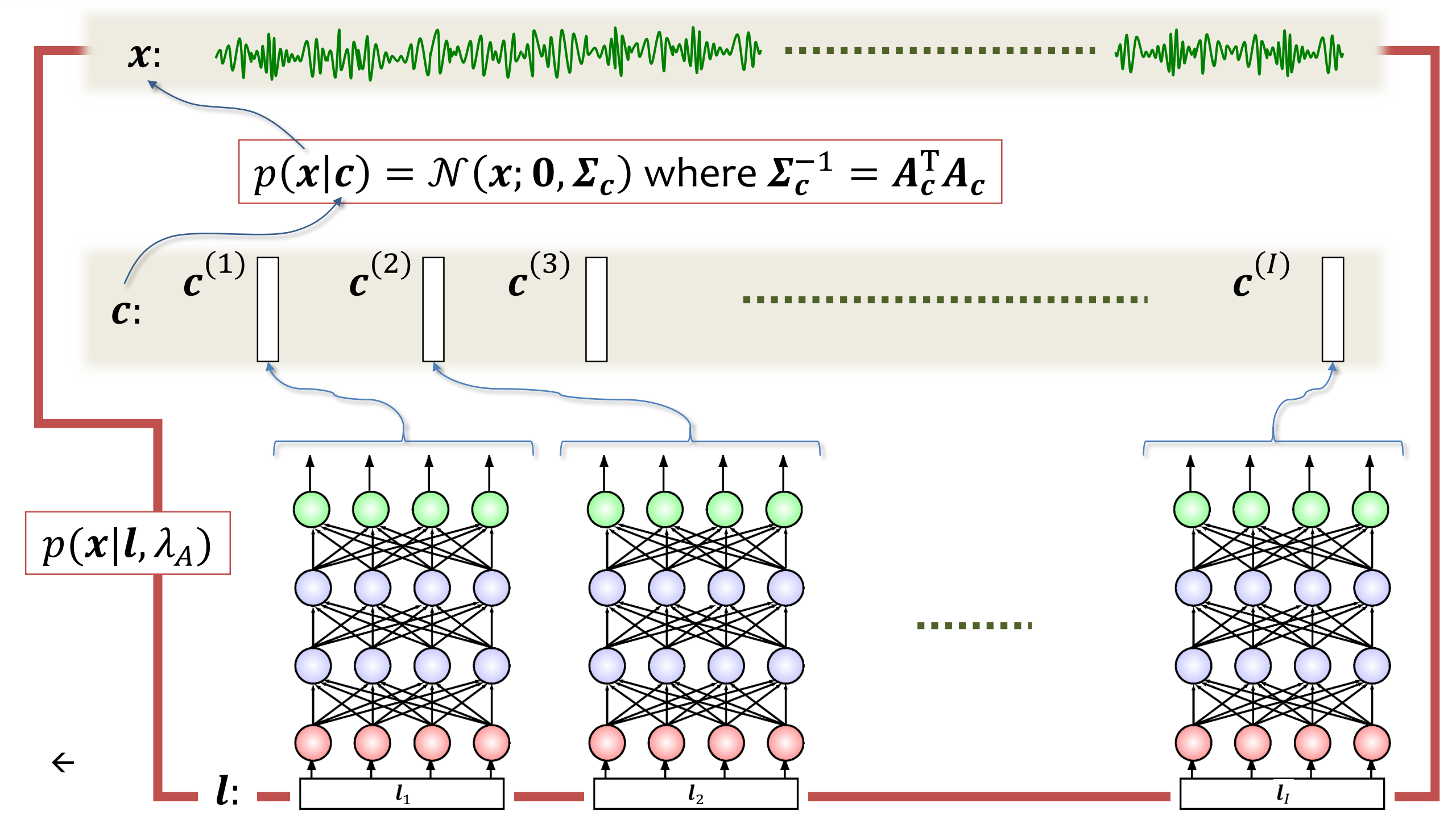
注）行列のサイズは無限大だが，簡単のため有限で記述している

非定常性のモデル化 (2/2)

- $H^{-1}(e^{j\omega})$ がひとつの発声の中で徐々に変化していると過程する



注）行列のサイズは無限大だが，簡単のため有限で記述している


$\boldsymbol{x}$:
$p(\boldsymbol{x}|\boldsymbol{c}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{c}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{c}}^{-1} = \boldsymbol{A}_{\boldsymbol{c}}^{\mathrm{T}} \boldsymbol{A}_{\boldsymbol{c}}$
$\boldsymbol{c}$:
$\boldsymbol{c}^{(1)}$
$\boldsymbol{c}^{(2)}$
$\boldsymbol{c}^{(3)}$
$\boldsymbol{c}^{(I)}$
$p(\boldsymbol{x}|\boldsymbol{l}, \lambda_A)$
$\boldsymbol{l}$:
l_1
l_2
l_I

微分

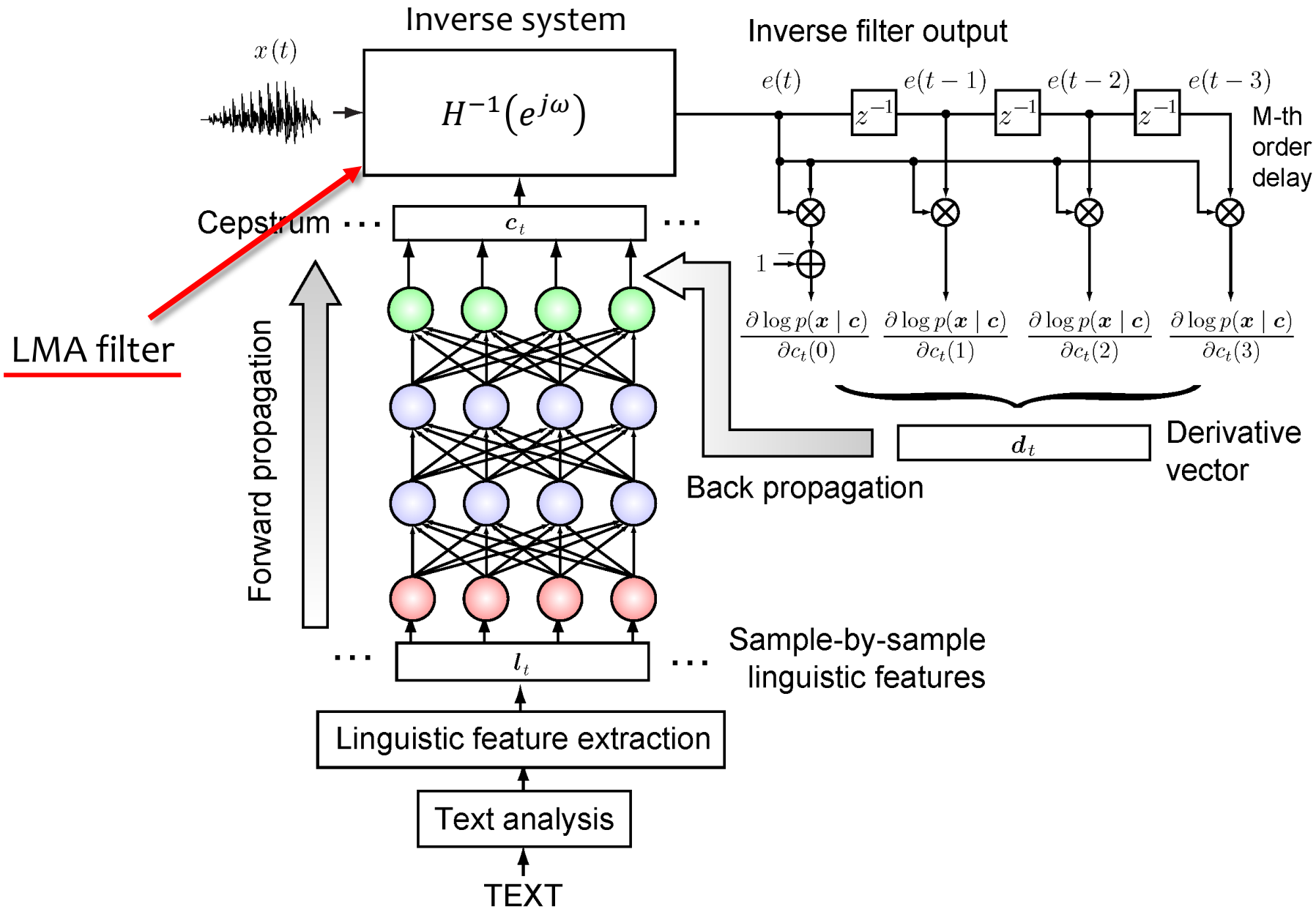
$$\frac{\partial \log p(\boldsymbol{x}|\boldsymbol{c})}{\partial \boldsymbol{c}^{(i)}} = \boldsymbol{d}_i = \begin{bmatrix} \sum_{k=0}^{L-1} \left\{e^{(i)}(Li+k)\right\}^2 - L \\ \sum_{k=0}^{L-1} e^{(i)}(Li+k)e^{(i)}(Li+k-1) \\ \vdots \\ \sum_{k=0}^{L-1} e^{(i)}(Li+k)e^{(i)}(Li+k-M) \end{bmatrix} \approx \begin{bmatrix} e^2(t)-1 \\ e(t)e(t-1) \\ \vdots \\ e(t)e(t-M) \end{bmatrix}$$

$$L = 1, i = t$$

$$\text{where } e^{(i)}(t) = \sum_{n=0}^{\infty} a^{(i)}(n)x(t-n)$$

微分を計算できたので，ニューラルネットに接続することが可能

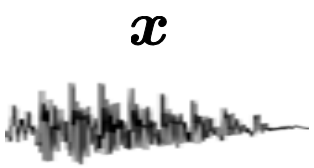
$L = 1, M = 3$, with some approximation assumptions

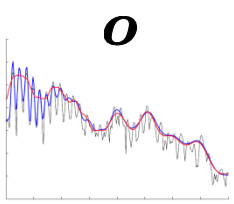


The acoustic model is illustrated as a feed-forward neural network rather than LSTM-RNN.

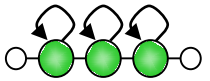
テキスト解析

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$





特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

テキスト解析

- 言語識別
- テキスト正規化
- 形態素解析，POS tagger
- 構文解析，係り受け解析
- 発音，発音変形の推定
- アクセント，アクセント結合，強勢，声調等の推定
- ポーズ位置・長さの推定
- などなど

通常のテキスト解析の問題

音声合成特有の問題

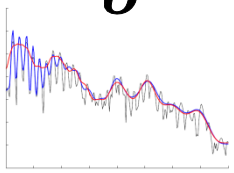
発音の変動等は音響モデルで吸収？テキスト解析部で推定？

二つのモジュールの再結合

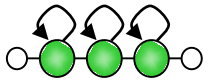
$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

$\boldsymbol{x}$

$\boldsymbol{o}$



特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現
$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の言語的特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ
λ_A: 音響モデルパラメータ
λ_L: テキスト解析部パラメータ

テキスト解析部と音響モデルの同時学習

- 学習データの発音変形，アクセント/強勢，ポーズの有無などは，人手で付与（あるいは修正）する必要がある
  - 大変なので，音響モデルを使う:

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} p(\boldsymbol{O}|\boldsymbol{L}, \lambda_A) P(\boldsymbol{L}|\boldsymbol{W}, \lambda_L)$$

- 更に，発音変形，アクセント/強勢，ポーズの有無などを隠れ変数とみなす
  - 音響モデルとの統合学習:

$$\boldsymbol{L} = \arg\max_{\lambda_A, \lambda_L} \sum_{L} p(\boldsymbol{O}|\boldsymbol{L}, \lambda_A)\, P(\boldsymbol{L}|\boldsymbol{W}, \lambda_L)$$

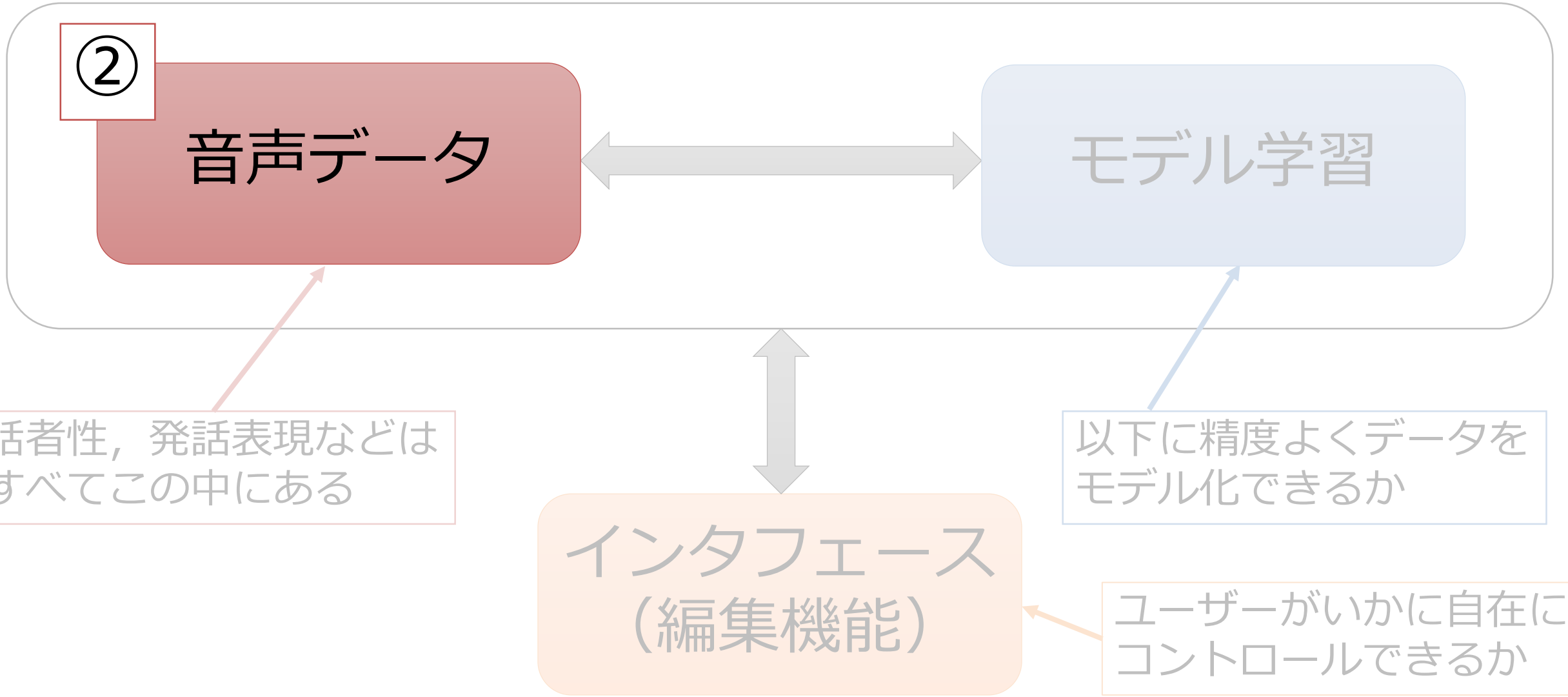
- テキストコーパスは沢山あるけど，音声付きのものは少ない

関係する論文

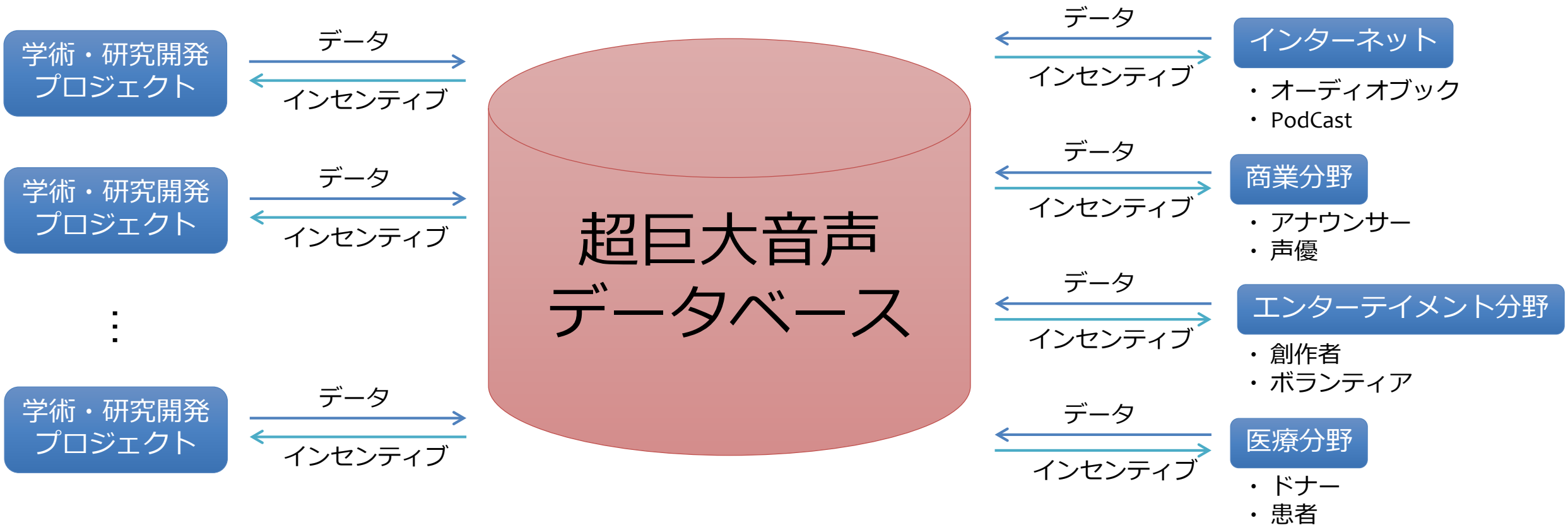
- Simultaneous Acoustic, Prosodic, and Phrasing Model Training for TTS Conversion Systems [Oura et al., 2008]
- H/L 型アクセント推定と音響モデリングを統合したHMM音声合成の検討 [神谷他, 2014]

⋮

音声合成の課題


②
音声データ
モデル学習
話者性，発話表現などは
すべてこの中にある
以下に精度よくデータを
モデル化できるか
インタフェース
（編集機能）
ユーザーがいかに自在に
コントロールできるか

音声データを集積・共有する仕組み



LDC, ELRA/ELDA, SRC, ALAGIN, GSK　　→ ソフトウェアに関しては

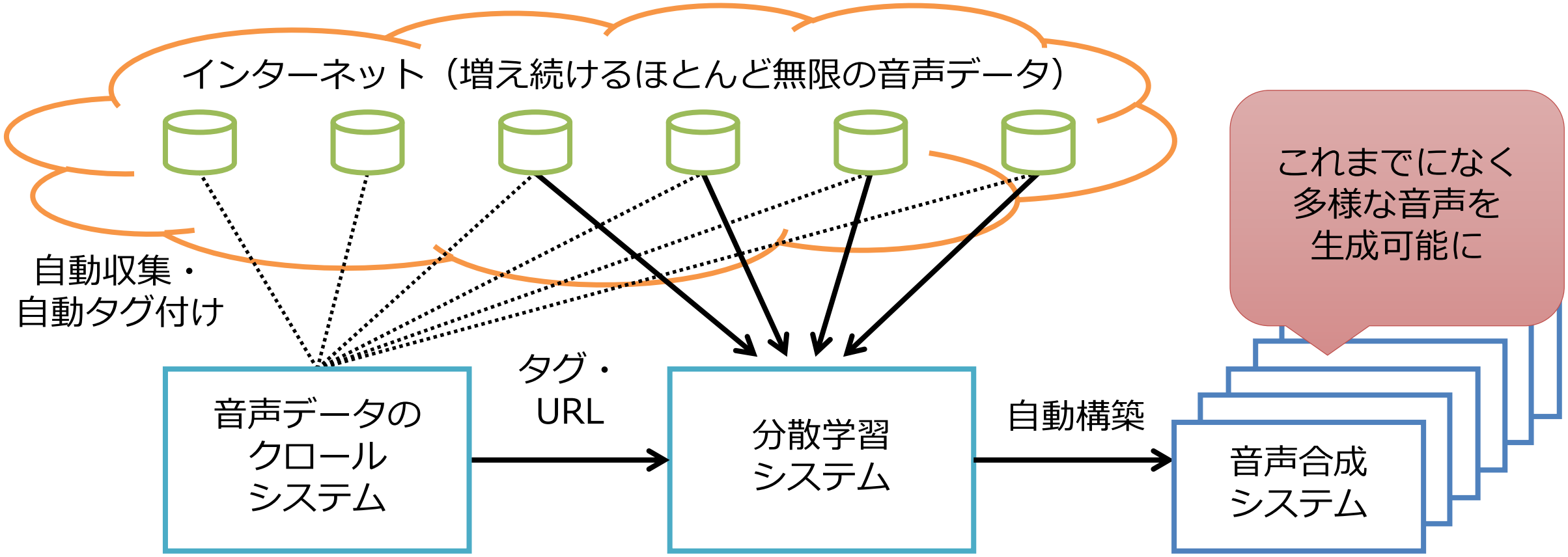
オープンソース・ソフトウェアツール

- SPTK, Speech Tools, Snack Sound Toolkit, ESPS, STRAIGHT, SWIPE’
- HTS, HTK, KALDI, IDLAK
- hts_engine API, Flite+hts_engine, Open Jtalk, Sinsy, Flite
- MMDAgent, Julius, Festival, MARY Text-To-Speech
- MeCab他多くのテキスト解析ツール
- 言語モデル学習ツール
- FST関係ツール
- Wavesurfer, PARAAT, などなど

社会的基盤

- 分断された音声データの共有化
  - 医療（ボイスバンク），エンターテイメント，商業分野，学術分野（Blizzard Challenge等）で個別に音声データを管理・維持（あるいは喪失）
- 音声データを収集・共有する社会的な仕組み
  - 適切なインセンティブの設定
  - 共通のライセンス形態の定義
- Web上の音声データの利用
  - Podcast, Audiobook, ニュース, あらゆる動画

学習用データの分散共有と分散型モデル学習

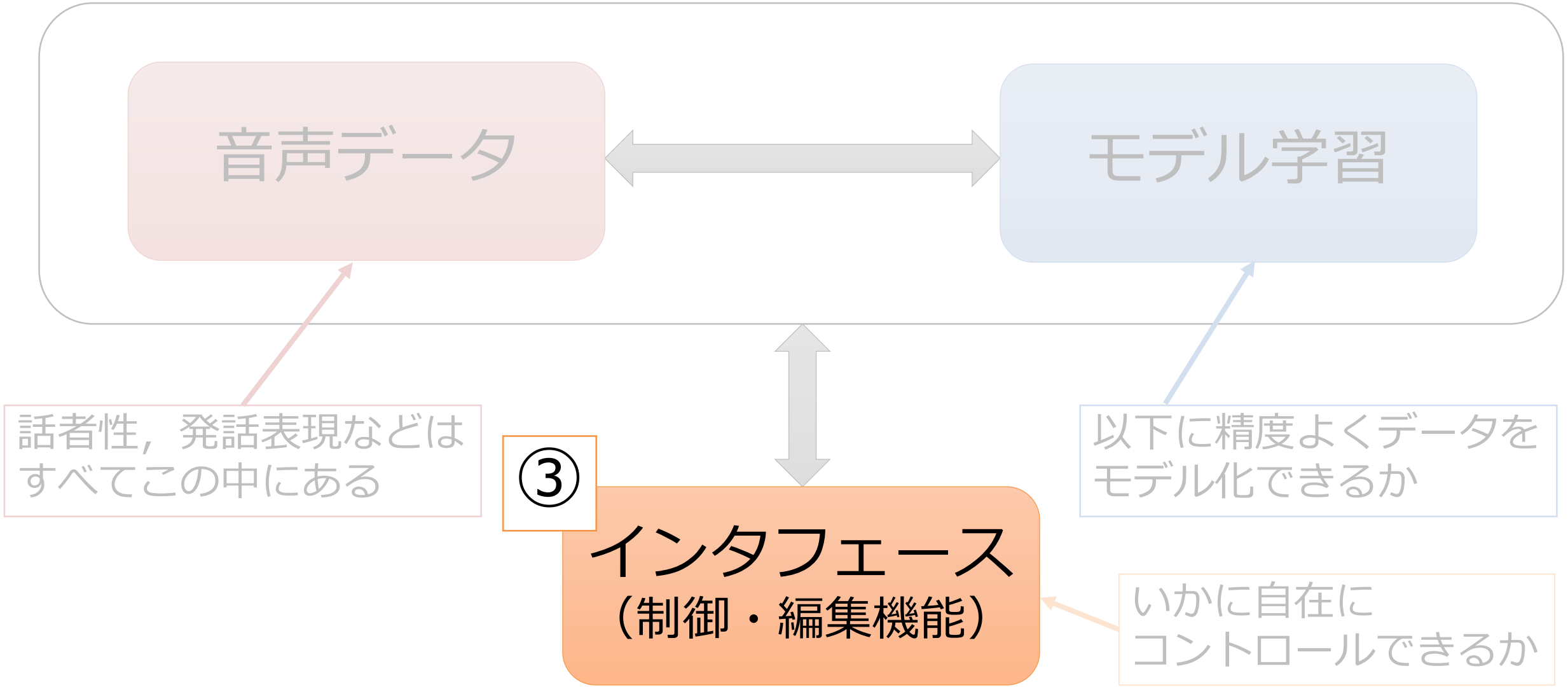


技術的基盤

- 大規模データを活かしたシステム構築法の確立
  - 多様なデータを活かしたモデル化手法の確立
  - 並列計算手法の確立
  - スケーラビリティ

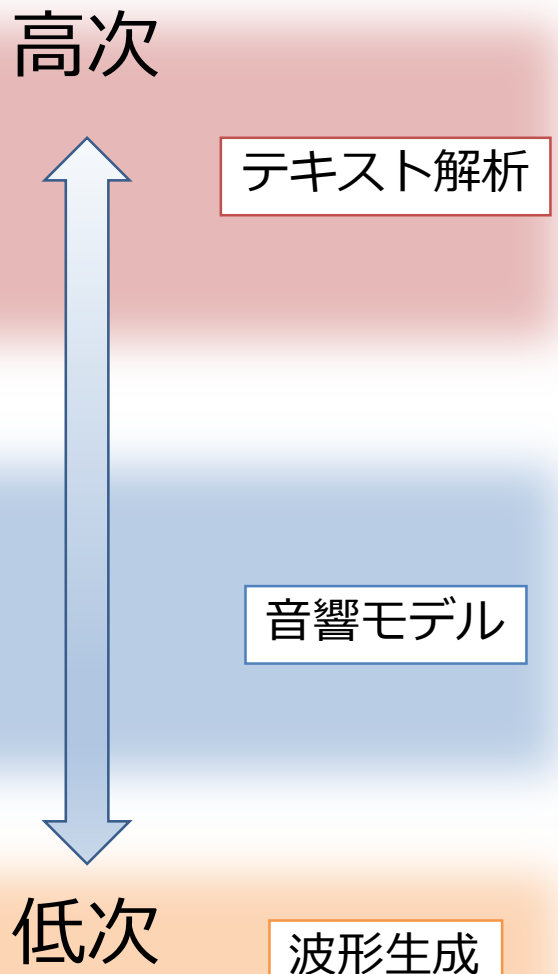
- 多様なデータを活かしたシステム構築法の確立
  - テキストとの不一致に対する対処
  - 背景雑音，言いよどみ等に対する対処
  - 動画等のマルチメディアデータ，ライフログ等からの音声データ抽出

音声合成の課題


音声データ
モデル学習
話者性，発話表現などは
すべてこの中にある
以下に精度よくデータを
モデル化できるか
③
インタフェース
（制御・編集機能）
いかに自在に
コントロールできるか

様々な階層での制御・編集機能

- 言語
  - 日本語，英語，中国語，日本語英語，…
- 読み
- ポーズ
- 発音変形
  - 音声 → /o N s e:/, します → /sh I ma s/
- 韻律変形
  - アクセント結合・変形
- 発話スタイル・感情表現等
- 単語の強調
- ノンバーバル情報・パラ言語情報
- 声質
  - 男性，女性，大人，子供
- 音声パラメータ
  - 基本周波数パターン，音量変化パターン，継続長，…



音声創作ソフトウェア「CeVIO Creative Studio」

- 感情音声合成

  ＋

- 歌声合成

- コンテンツ制作を意識した編集機能

↑　↑　(→)

本発表のあらまし

- 背景
- 音声合成の基本問題
- 音声合成の課題
- 評価
- 音声合成の社会的役割
- まとめ

余談をはさみながら

Blizzard Challenge

- 合成音声の品質は音声データベースに依存
- 音声合成技術自体の比較・評価は困難

“Blizzard Challenge”

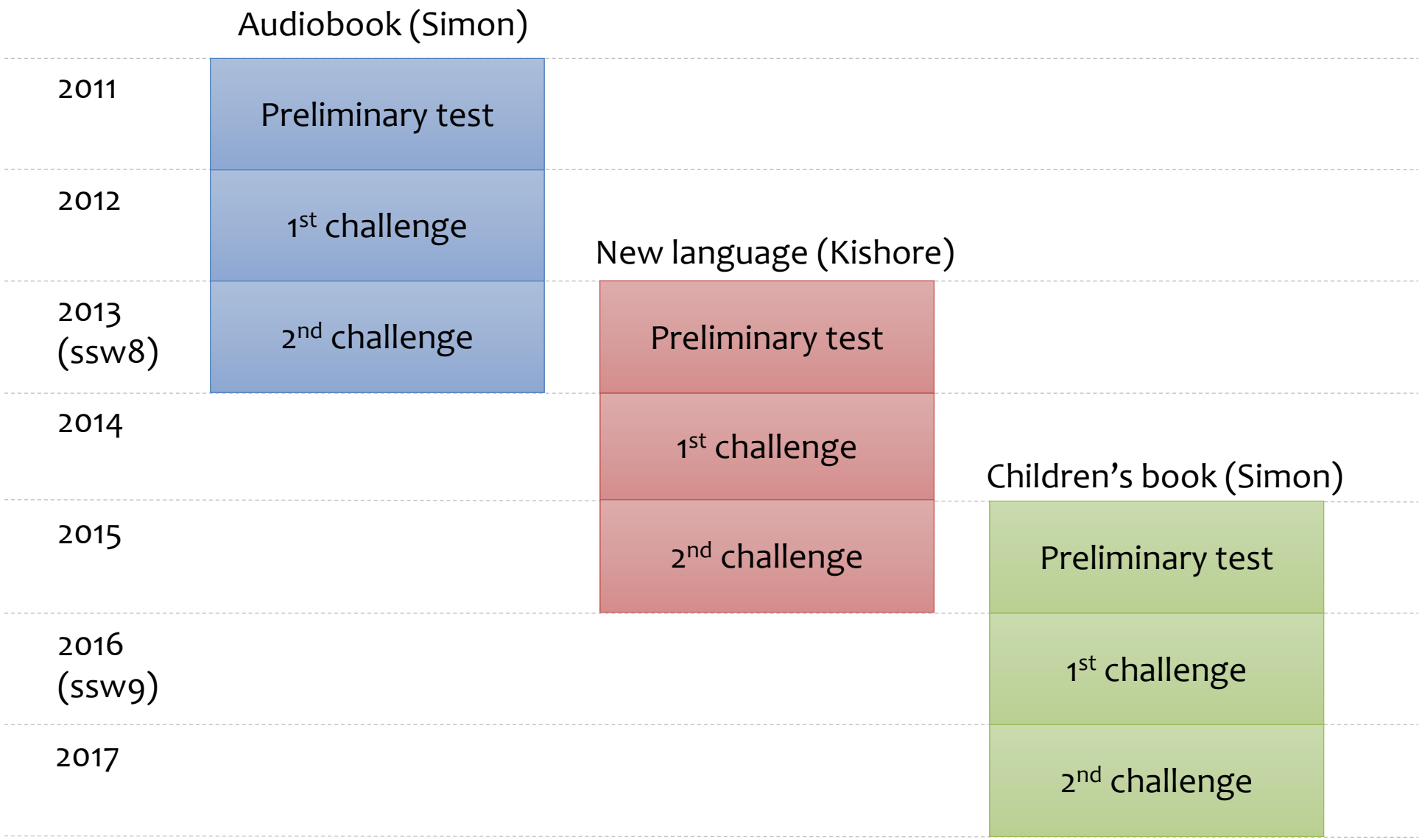
Evaluating corpus-based speech synthesis

on **common datasets**

Since 2005

→

New Blizzard Challenge Structure



今後のタスク（徐々に難しく）

- 自由発話はまだまだ難しい
  - 内在する心理状態等の変動要因が隠れ変数としてうまくモデル化できていない？
  - ラベリングは必要？少なくとも初期モデルとしては？
  - ランダムな変動は unsupervised に学習？
  - そもそも言語解析がうまく動かない
- マルチリンガル・クロスリンガル
  - 言語性と話者性の分離
- Low resource language, zero resource language
  - 数千の書記言語

- 評価法自体や主観評価結果の予測なども重要課題

本発表のあらまし

- 背景
- 音声合成の基本問題
- 音声合成の課題
- 評価
- 音声合成の社会的役割
- まとめ

余談をはさみながら

なりすまし・オレオレ詐欺に利用？

- 合成音声による詐称の懸念
  - On the security of HMM-based speaker verification systems against imposture using synthetic speech [Masuko et al., 1999]
- 合成音声の検出
  - A robust speaker verification system against imposture using an HMM-based speech synthesis system [Satoh et al., 2001]
- ASVspoof 2015
  - The First Automatic Speaker Verification Spoofing and Countermeasures Challenge
- 音声合成技術がこのような段階に来ていることを一般に広く知って貰う必要がある → 画像は既に既知（写真の偽造が簡単なのは周知）
- リアルタイムでは無理

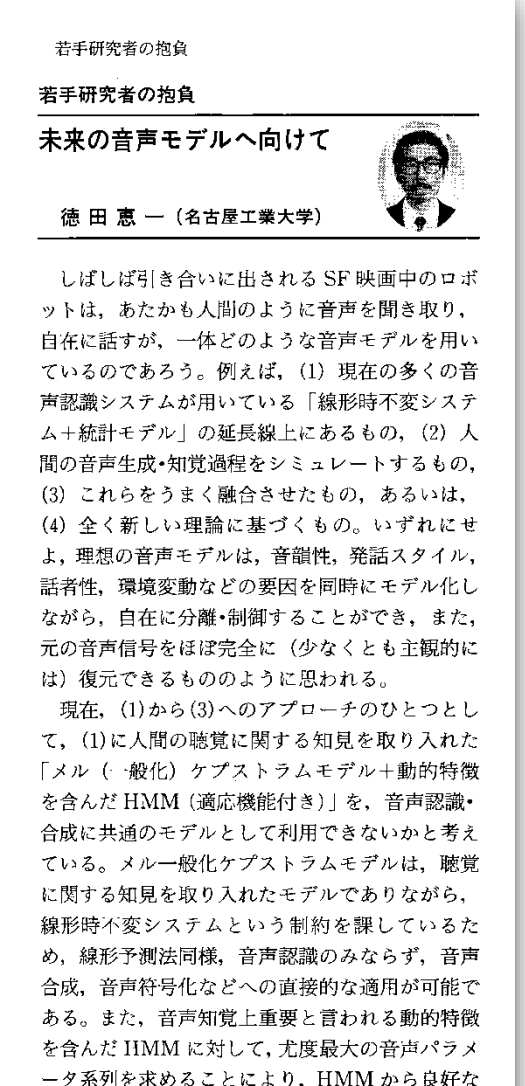
声の職業と音声合成の関係

- アナウンサー・ナレーター・声優など，声の職業がなくなる？
  - 現状では自然音声と合成音声の違いは大きい
  - リアルタイム性の必要な応用は難しい
  - コンテンツ制作には十分利用可能
- 演奏家とレコード・CD・音楽配信の関係
  - レコードを出すと演奏会に来てもらえなくなる？
  - CDからの収益
  - リアルイベントへの導入
- 音声合成システムもレコードのような役割を果たしうる？
- 誰の声でもない創造された声の権利は？
  - ビジュアルはキャラクターライセンス → 声は？

余談: “Right time, right place, right people”

- 日本音響学会誌「若手研究者の抱負」１９９６年

- 音声スペクトル分析・音声符号化・適応信号処理
- 優秀な指導者
- 優秀な同僚
- 親切な研究者たち
- HTK/Festival
- 思い込みと頭の悪さ？
- 音声合成研究者は単位選択で忙しかった
- 音声認識研究と音声合成研究は分離していた
- 大学の研究環境がまだ平穏な時代だった



若手研究者の抱負

若手研究者の抱負

未来の音声モデルへ向けて

徳田恵一（名古屋工業大学）

しばしば引き合いに出されるSF映画中のロボットは，あたかも人間のように音声を聞き取り，自在に話すが，一体どのような音声モデルを用いているのであろう。例えば，(1) 現在の多くの音声認識システムが用いている「線形時不変システム＋統計モデル」の延長線上にあるもの，(2) 人間の音声生成・知覚過程をシミュレートするもの，(3) これらをうまく融合させたもの，あるいは，(4) 全く新しい理論に基づくもの。いずれにせよ，理想の音声モデルは，音韻性，発話スタイル，話者性，環境変動などの要因を同時にモデル化しながら，自在に分離・制御することができ，また，元の音声信号をほぼ完全に（少なくとも主観的には）復元できるもののように思われる。

現在，(1)から(3)へのアプローチのひとつとして，(1)に人間の聴覚に関する知見を取り入れた「メル（一般化）ケプストラムモデル＋動的特徴を含んだHMM（適応機能付き）」を，音声認識・合成に共通のモデルとして利用できないかと考えている。メル一般化ケプストラムモデルは，聴覚に関する知見を取り入れたモデルでありながら，線形時不変システムという制約を課しているため，線形予測法同様，音声認識のみならず，音声合成，音声符号化などへの直接的な適用が可能である。また，音声知覚上重要と言われる動的特徴を含んだHMMに対して，尤度最大の音声パラメータ系列を求めることにより，HMMから良好な

余談: Googleに１年間滞在してみて

- お金があればなんでもできる？（理路整然とした金満体質）
  - 優秀な人材とリーダー（優秀なリーダーもお金があれば雇える）
  - お金があっても，硬直化した組織と評価システムではダメ

- もう太刀打ちできないのか？
  - より早く動く（俊敏性・少数精鋭）
  - 未開フィールドへ
  - より深く複合的な学術知識

まとめ

全音声空間に広がる超巨大分散データにより，あらゆる音声を
自在に生成することの可能な音声合成システムを構築する

- 超巨大データを用いたユニバーサル音声モデルの構築
  - 「平均声」および「固有声」の手法をベースに融合
  - 分散・並列処理，荒れたデータへの対応
- データの分散共有（人類共有の財産として）
  - インセンティブ・ライセンス
- 編集機能の多元化・多層化
  - 高次から低次まで，多様な操作機能

今後の課題

- 部分生成モデルの密結合
- バックエンドとの結合
- 波形の直接モデル化
- より詳細な物理モデル
- 音声の多様性を表現できるユニバーサルな音声モデル
- 大規模データの分散共有と分散学習
- マルチリンガル・クロスリンガル・low resource language
- 自由発話・本当のリアル音声
- 音声合成の社会的役割
  - オレオレ詐欺
  - アナウンサー・声優との共存

参考書籍「おしゃべりなコンピュータ」

- 最新の音声合成技術について，専門家以外にもわかりすい言葉で書かれています
- 明るい未来の技術！
  - 人に優しく，楽しく使えるインターフェース
  - 医療応用：声を失う・失った方→ボイスバンクプロジェクト
- ISBN: 978-4621053850

HTS Slides
released by HTS Working Group
http://hts.sp.nitech.ac.jp/