信号処理論特論
音声合成・変換 2

猿渡 洋・高道 慎之介

hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
情報理工学系研究科
システム情報学専攻

Dec. 20, 2016

目次

Section 1

講義予定など

講義予定

- 09/27: 第１回　統計的音声音響信号処理概論
- 10/04: 第２回　非負値行列因子分解
- 10/11: 第３回　独立因子分析（ICA, IVA, ILRMA）
- 10/18: 第４回　独立因子分析（続き）
- 10/25: 第５回　音場再現・スパース最適化
- 11/01: 第６回　音声合成・変換１
- 11/15: 第７回　【レポート課題１】
- 11/22: 第８回　話者認識
- 11/29: 休講
- 12/06: 第９回　エンハンスメント・高次統計量解析
- 12/13: 休講
- 12/20: 第 10 回　音声合成・変換２
- 01/10: 第 11 回　【レポート課題２】

講義資料と成績評価

- 講義資料
  - http://www.sp.ipc.i.u-tokyo.ac.jp/~saruwatari/
  - （システム情報第一研究室からたどれるようになってます）
- 成績評価
  - 出席点
  - レポート点（２回の提出が必須）

今日お話しすること

音声合成・変換を高品質化する技術とは？
それをどう応用する？

Section 2

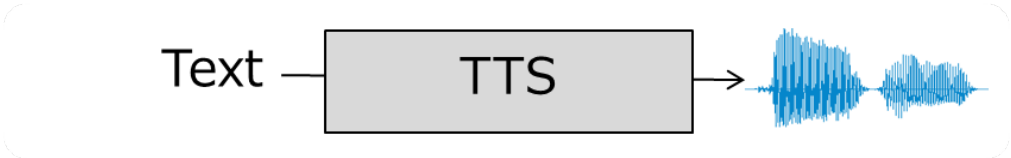
復習

音声合成：音声を人工的に作り出す技術

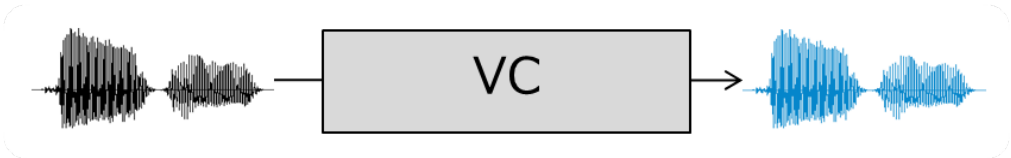
- 狭義の音声合成
  - テキスト音声合成 (**Text-To-Speech: TTS**)
- 広義の音声合成 (XX-to-speech)
  - テキスト音声合成
  - 音声変換 (**Voice Conversion: VC**) …ボイスチェンジャ
  - 概念音声合成 (Concept-To-Speech: CTS)…概念 $\rightarrow$ 言語生成 $\rightarrow$ 音声合成
  - 調音・音響間マッピング…調音機構特性と音声の変換
  - マルチモーダル音声合成…動画像などを含む音声合成

テキスト音声合成と音声変換

- **テキスト音声合成 (Text-To-Speech: TTS)**
  - テキスト等から音声を合成
  - ヒト以外のモノのコミュニケーションのため



- **音声変換 (Voice Conversion: VC)**
  - 言語情報を保持したままパラ言語・非言語情報を変換
  - ヒトの発声制約をこえたコミュニケーションのため

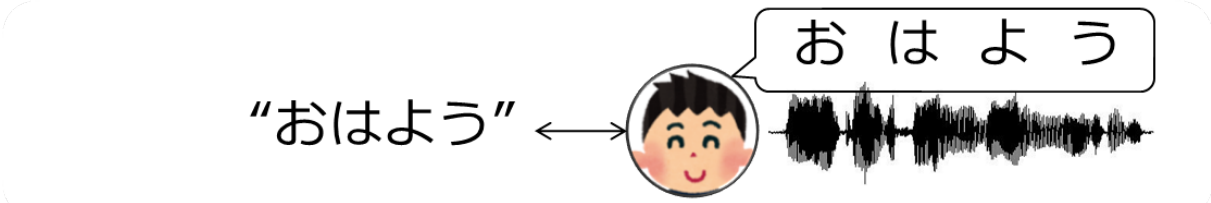


コーパスベース音声合成の種類

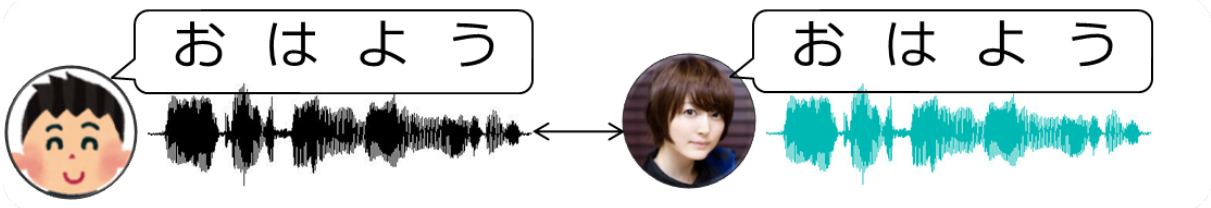
- ルールベース音声合成
  - 開発者独自の規則による音声合成
  - フォルマント音声合成 (-1990)
  - ルールベースの周波数伸縮等による音声変換
- **コーパスベース音声合成 (1990-)**
  - データドリブンで音声合成を構築
  - 知見・技術の共有が可能に
  - 入出力データの関係性を記述する逆問題

音声合成・変換の学習データ

- 教師あり … パラレルデータあり
  - 音声合成: テキスト＆音声



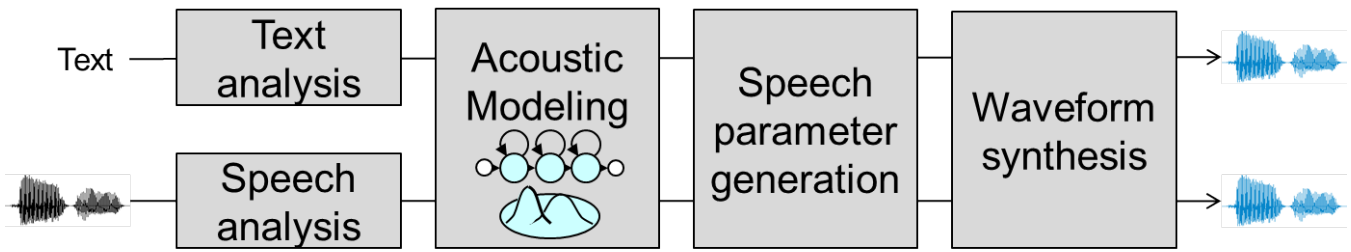
  - 音声変換 (話者変換): 異なる話者による音声対



- 教師なし… パラレルデータなし
  - 音声合成: 音声のみ or 言語知識なし など
  - 音声変換 (話者変換): 発話内容の異なる音声 など

コーパスベース音声合成方式の種類

- サンプルベース音声合成 (素片選択型合成) [1]
  - 音声波形・パラメータを保存し、その接続・加工で音声合成
  - 長所: 非常に肉声感の高い合成音
  - 短所: 声質を制御しにくい、フットプリントが大きい
- 統計的音声合成 [2]
  - 音声波形・パラメータを統計モデルでモデル化
  - 長所: 声質を制御しやすい、フットプリントが小さい
  - 短所: 低い音質 (こもった音になりやすい)

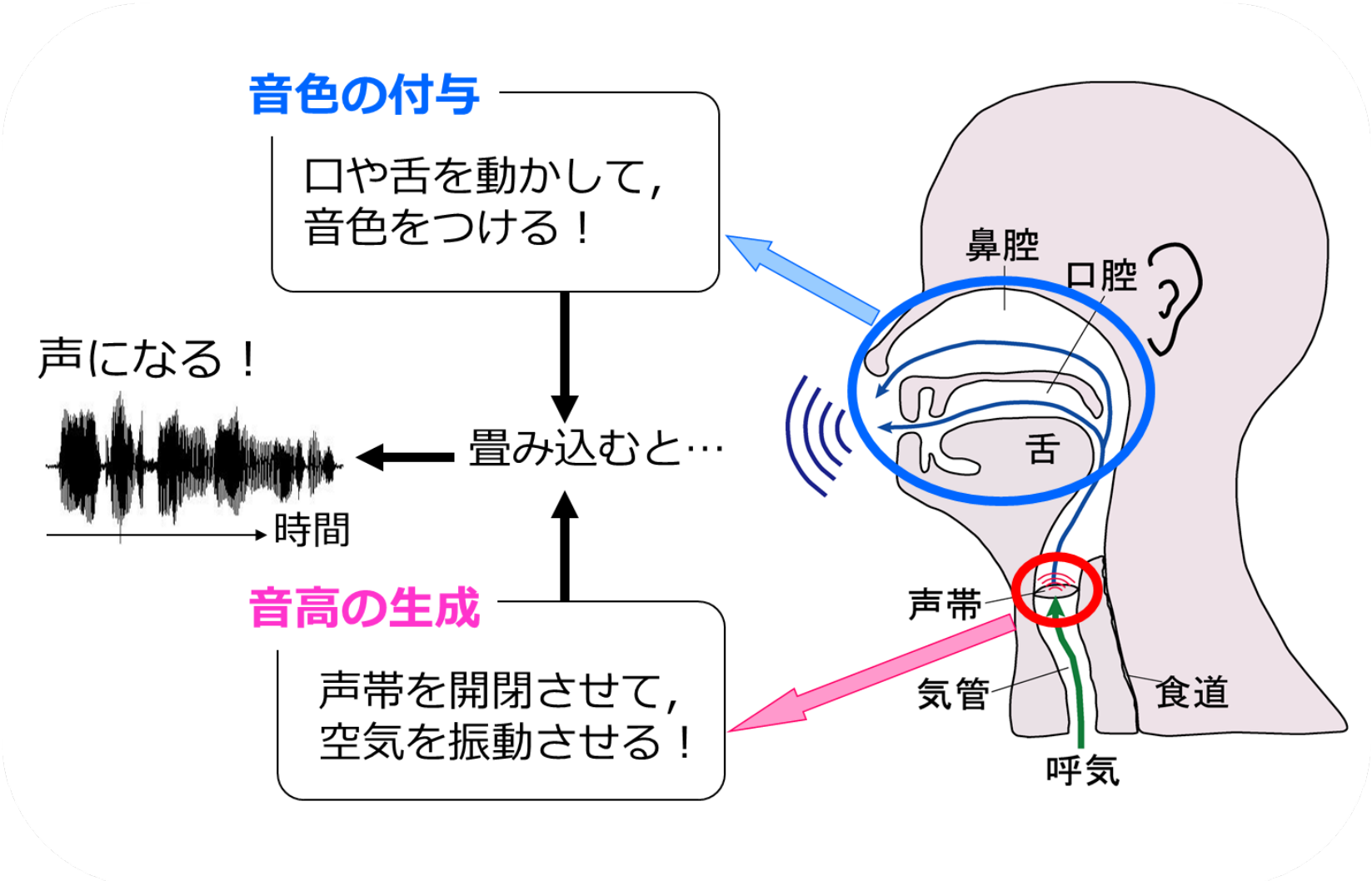


テキスト解析

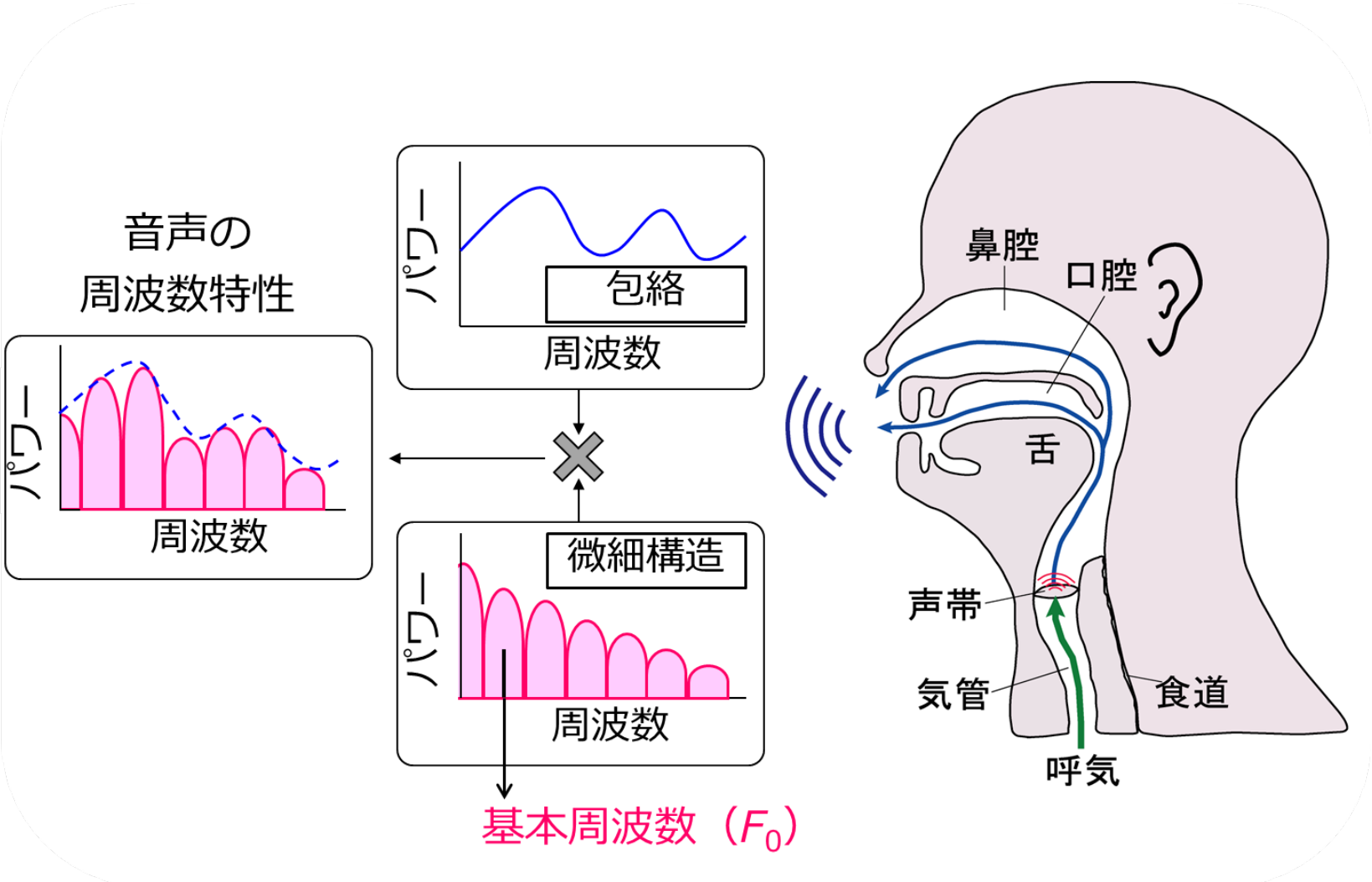
- 自然言語処理寄りの処理
  - 言語識別 (Language identification)
  - テキスト正規化 (Text normalization)
  - 形態素解析、Part-Of-Speech (POS) tagging
  - 構文解析、係り受け解析
- 音声合成独自の処理
  - 発音 (音素)、発音変形 (音韻交替)
    - 音韻交替の例：二本 (にほん) → 三本 (さんぼん)
  - アクセント、ストレス、アクセント結合、声調
    - アクセント結合の例：

      にひゃく + メートル → にひゃくメートル
  - ポーズ位置・長さなど

発声機構

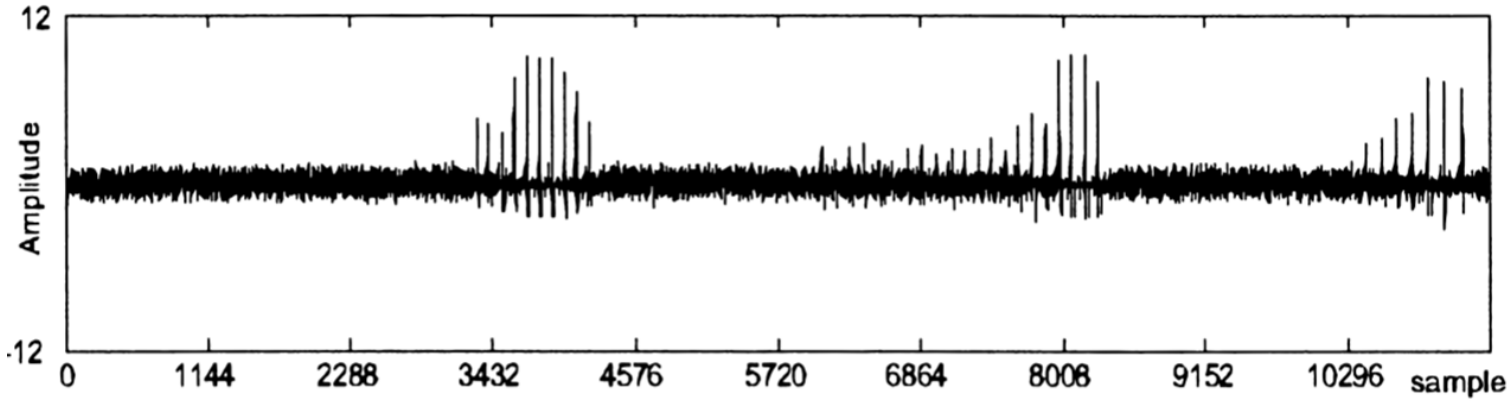

音色の付与
口や舌を動かして，
音色をつける！
声になる！
時間
畳み込むと…
音高の生成
声帯を開閉させて，
空気を振動させる！
鼻腔
口腔
舌
声帯
気管
食道
呼気

発声機構

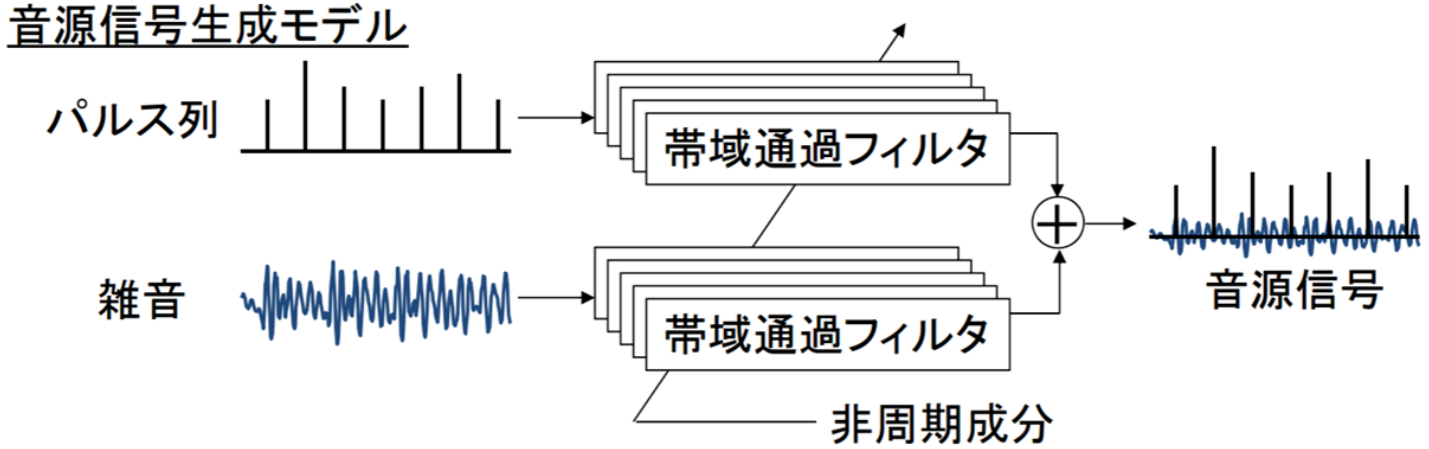

音声の
周波数特性
パワー
周波数
パワー
包絡
周波数
微細構造
パワー
周波数
基本周波数（F_0）
鼻腔
口腔
舌
声帯
気管
食道
呼気

音源信号のモデル化 (混合励振源) [6]

- パルス列と白色ノイズの重み付き和で、より高精度な表現
- STRAIGHT [3, 4] と呼ばれる分析再合成法が有名

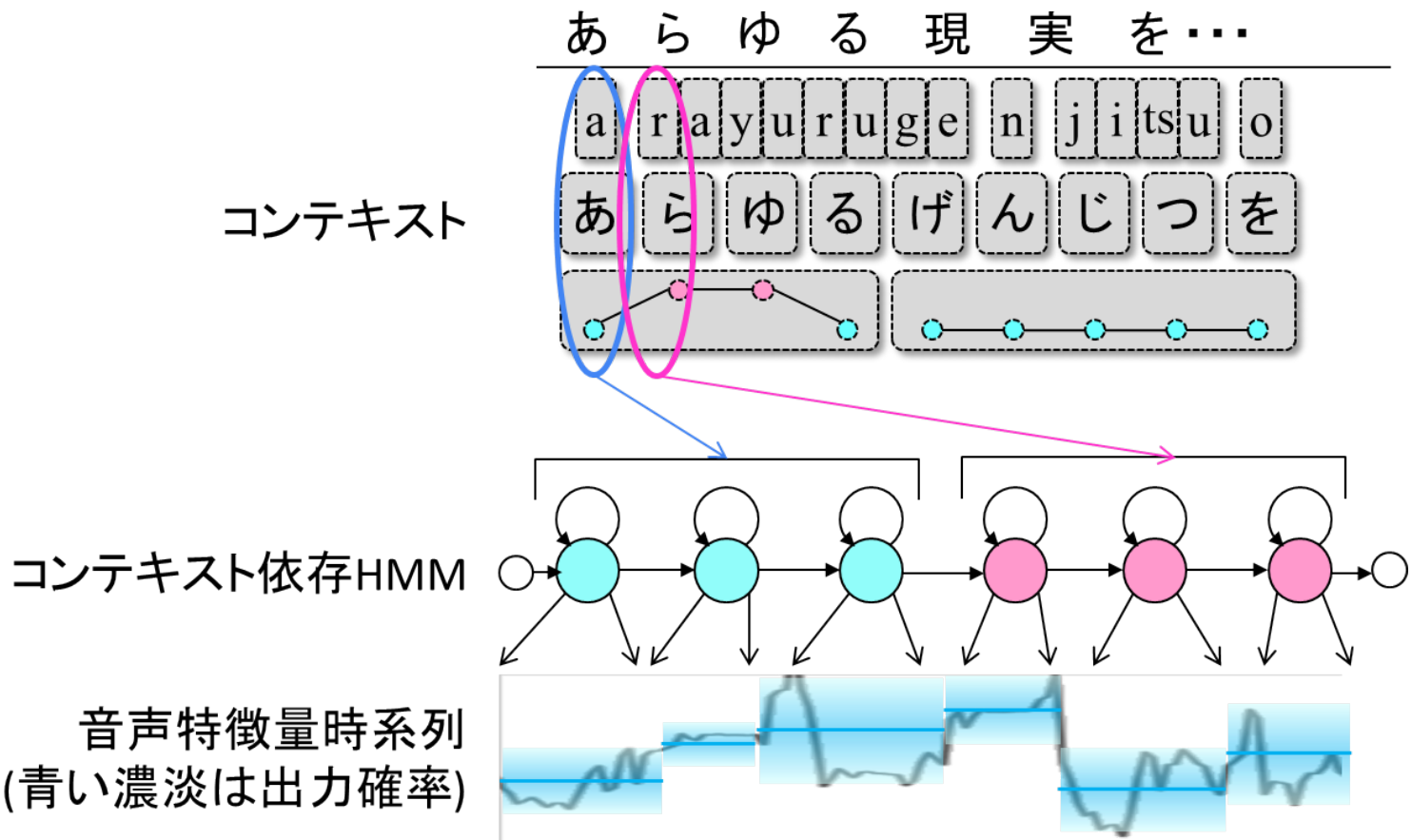


- 周波数ごとの重み付き和で生成 (図は [5] から引用)



コンテキスト依存 HMM の学習

- 各コンテキスト毎に HMM を学習。各 HMM 状態でセグメントの最初・真ん中・最後あたりをモデル化する

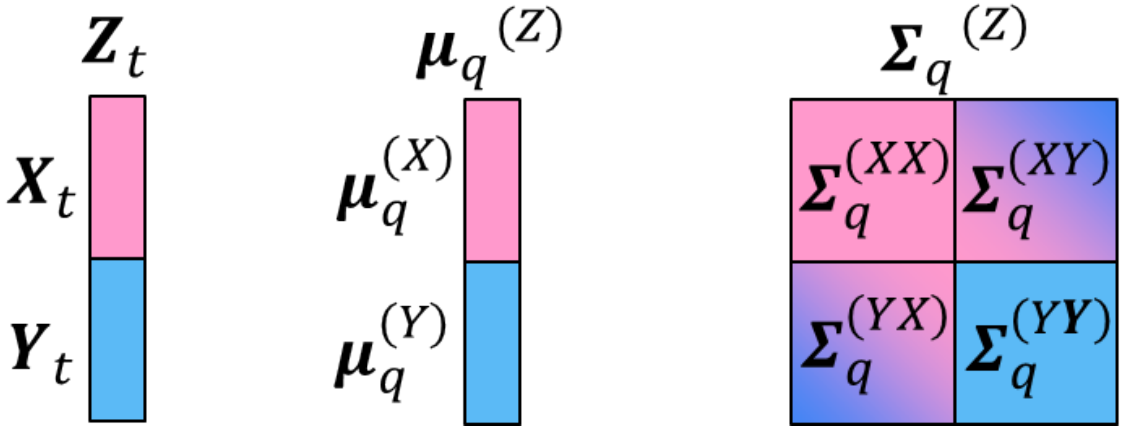


GMM による同時確率のモデル化

- 入力特徴量 $\boldsymbol{X}_t$、出力特徴量 $\boldsymbol{Y}_t$
- $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ の p.d.f. を GMM でモデル化
- HMM と同様に EM アルゴリズムで学習可能

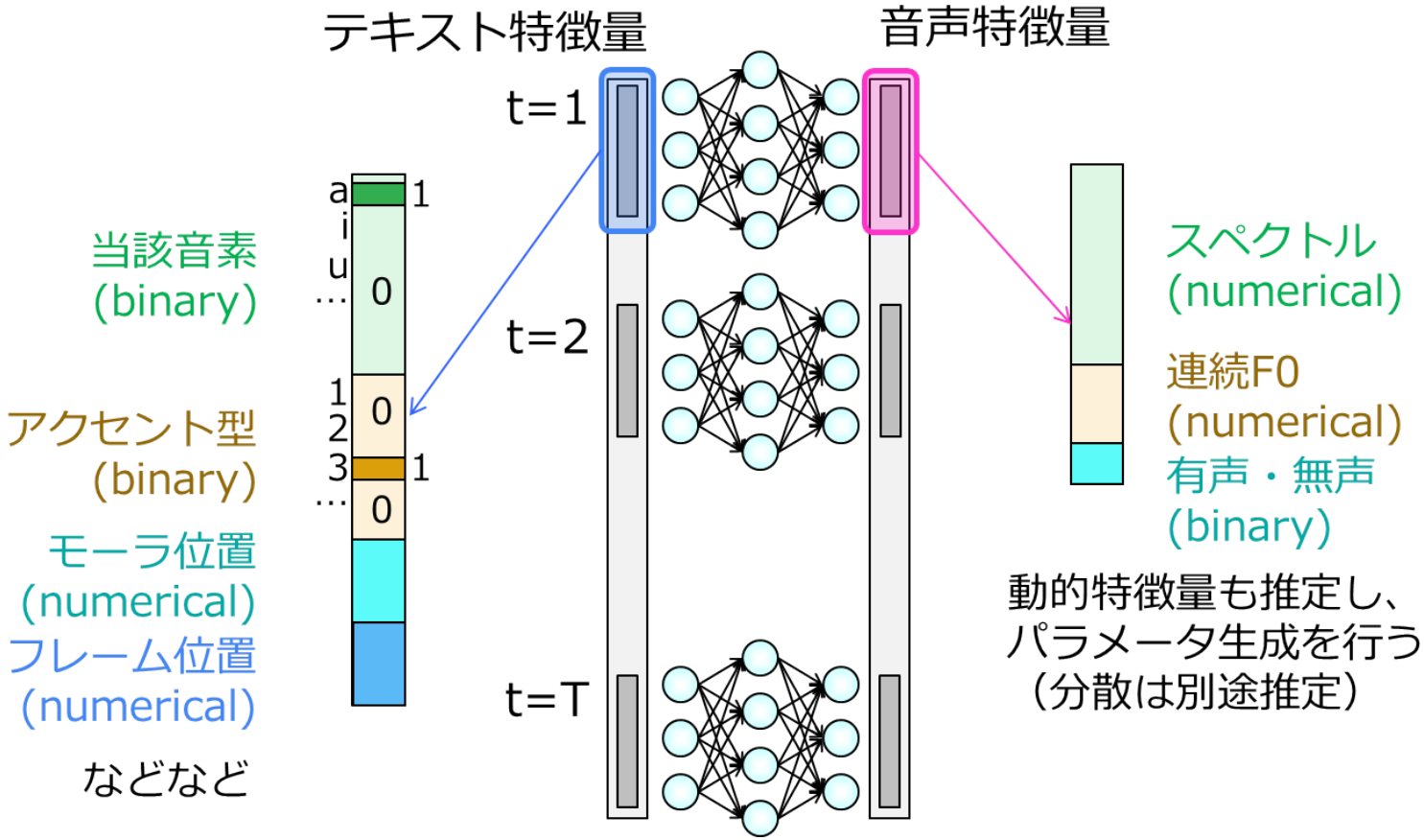
$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}) = \sum_{q=1}^{Q} w_q^{(Z)} \mathcal{N}\left(\boldsymbol{Z}_t|\, \boldsymbol{\mu}_q^{(Z)}, \boldsymbol{\Sigma}_q^{(Z)}\right), \tag{1}$$

$$\boldsymbol{\mu}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_q^{(X)} \\ \boldsymbol{\mu}_q^{(Y)} \end{bmatrix}, \ \boldsymbol{\Sigma}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)} \\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)} \end{bmatrix}, \tag{2}$$



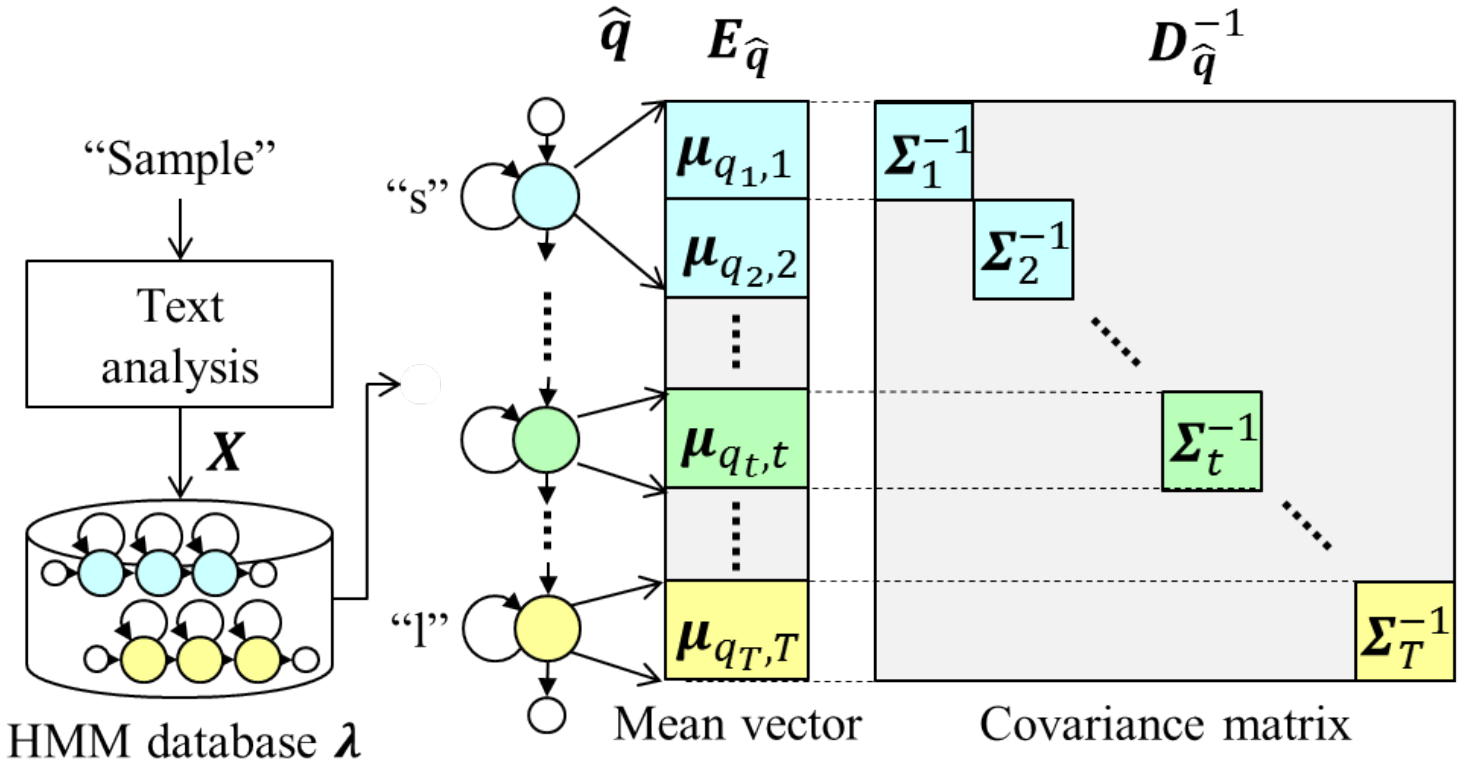
DNN による特徴量予測 [7]

- l 番目の隠れ層の出力 $\boldsymbol{h}_l = \boldsymbol{f}_l(\boldsymbol{W}_l \boldsymbol{h}_{l-1} + \boldsymbol{b}_l)$
- ($\boldsymbol{f}_l(\cdot)$ は活性化関数)



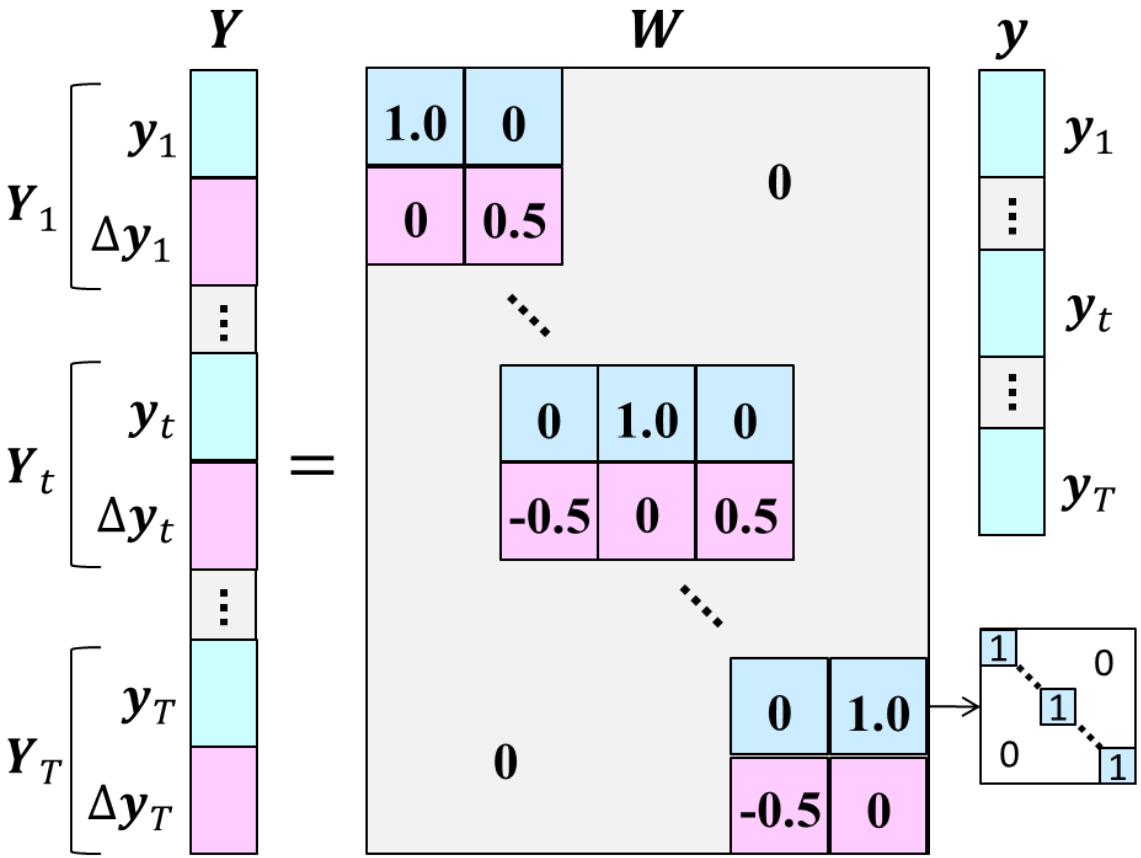
音声パラメータ系列の従う正規分布

- 同一 HMM 状態に対応するフレームでは、同一の統計量
  - HMM 状態が変わるフレームで $\boldsymbol{E}_{\hat{q}}$ は不連続に遷移
- この最尤推定では結局 $\hat{\boldsymbol{y}}_{\hat{q}} = \boldsymbol{E}_{\hat{q}}$ となり不連続な時系列
  - 動的特徴量の導入による時間変化のモデル化



動的特徴量の導入

- 時間変化を表す動的特徴量を導入し、HMM を学習
- 1 次の動的特徴量 $\Delta \boldsymbol{y}_t = 0.5\boldsymbol{y}_{t+1} - 0.5\boldsymbol{y}_{t-1}$

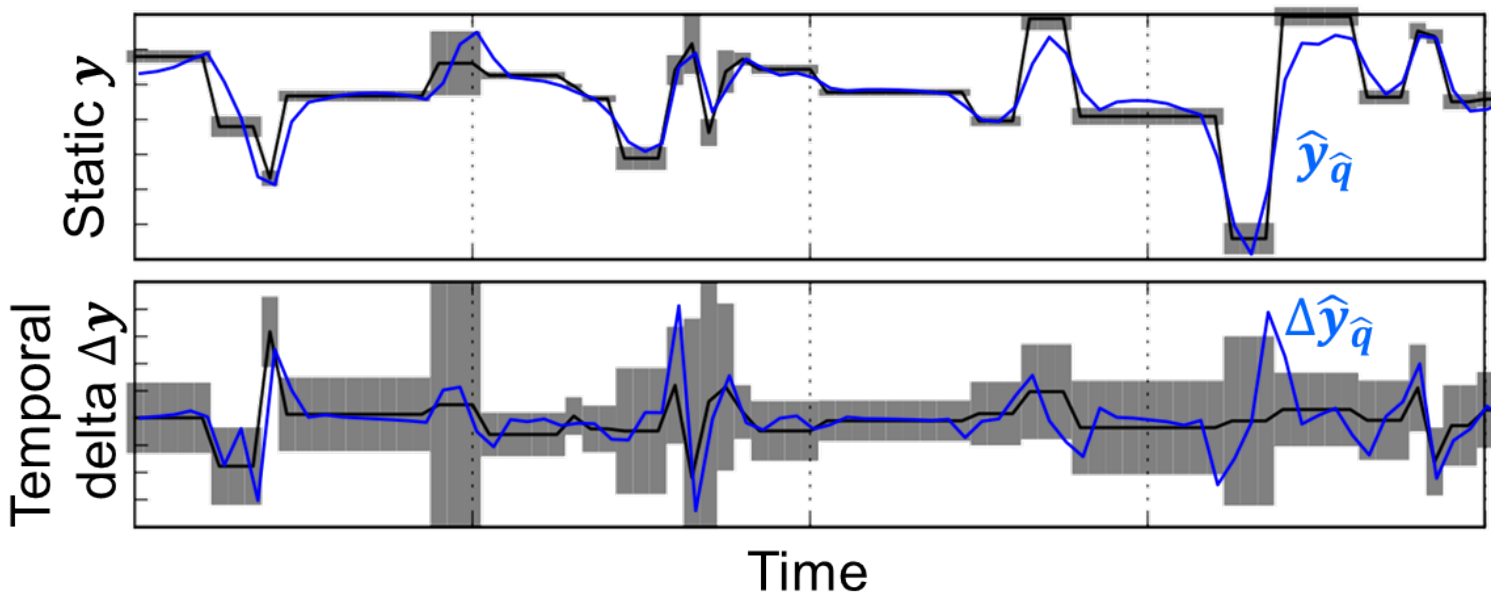


最尤パラメータ生成

- 動的特徴量の制約の下で最尤推定
- $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \mathrm{argmax}\, P\left(\boldsymbol{Y}|\boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \mathrm{argmax}\, \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right)$
- 上式の対数の微分を 0 とおき、次式が得られる

動的特徴量を考慮した音声パラメータ生成

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{E}_{\hat{\boldsymbol{q}}}$$



Section 3

どうやって高品質化する？

高品質化手法

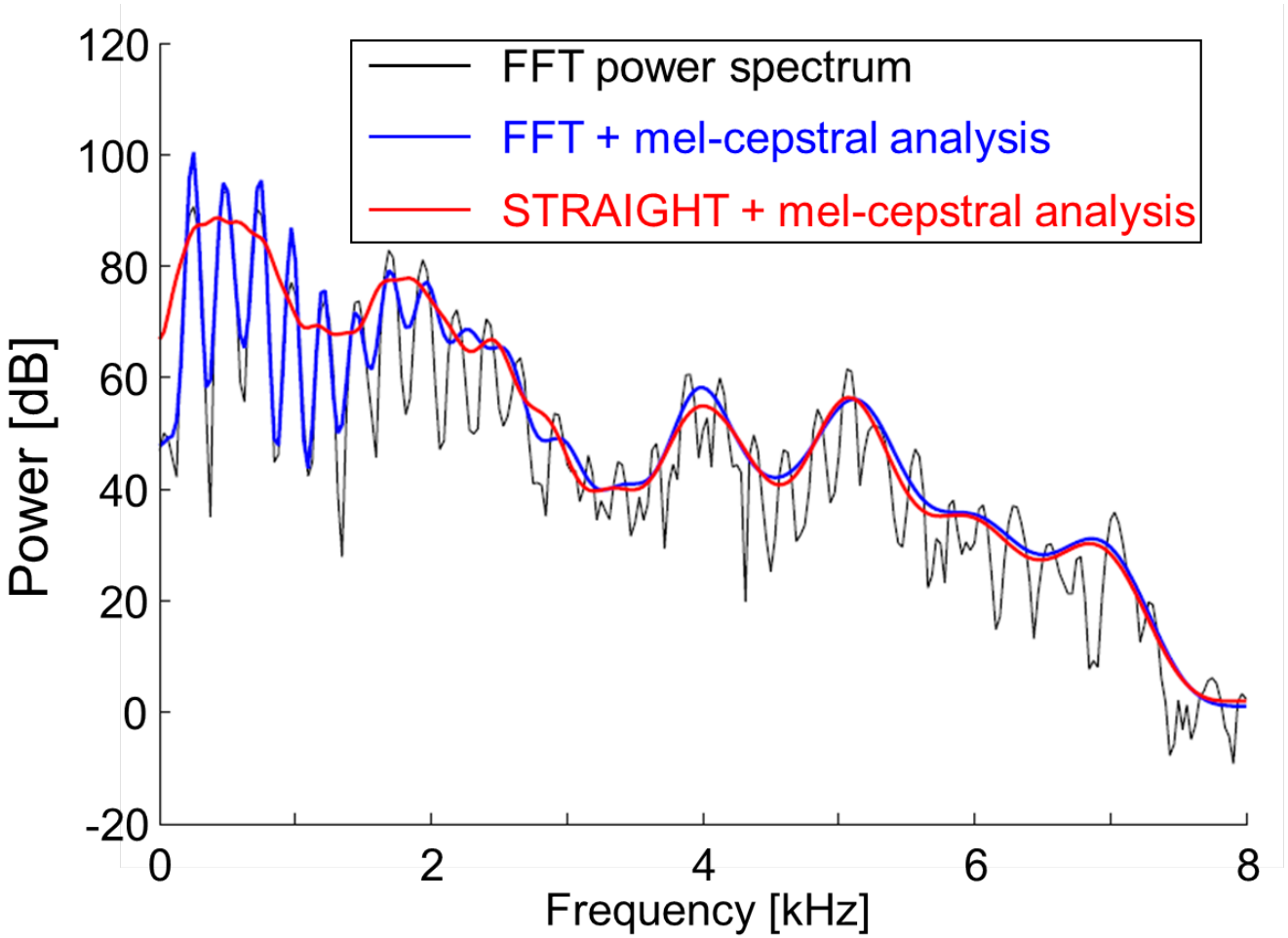
- 分析合成法を改善
  - STRAIGHT [3, 4]
  - WORLD [8, 9, 10]
  - Auto-encoder [11, 12]
- 音響モデルを改善
  - Trajectory model [13]
  - Additive model [14] / hierarchical model [15]
  - LSTM [16]
- 音声パラメータ生成を改善
  - Cepstrum emphasis [17]
  - Global Variance (GV) [18]
  - Modulation Spectrum (MS) [19]
  - Adversarial speech synthesis [20]
- 複数モジュールを同時に改善
  - Direct waveform modeling [21, 22]

分析合成法

- どういう分析合成器が必要？
  - 声道特徴量（スペクトル包絡など）と音源特徴量（F_0 など）を独立に分解
  - あらゆる音声に対して頑健に動作
- 統計的音声合成は、STRAIGHT [3, 4] のおかげで躍進
  - TANDEM-STRAIGHT では F_0 に同期した 2 つの窓関数を利用
  - `http://www.wakayama-u.ac.jp/~kawahara/HowTANDEMSTRAIGHTworks/` (要 Quicktime)

STRAIGHT によるスペクトル包絡 [23]

- ケプストラム法では，F_0 がスペクトル包絡に影響
- 一方，STRAIGHT では，F_0 の影響を限りなく除去



WORLD の登場

- STRAIGHT は特許が絡むので、産業に応用しづらい
- WORLD と呼ばれる BSD ライセンスのシステムが登場
  - **HMM 音声合成において、STRAIGHT と同品質** [24]
  - `http://ml.cs.yamanashi.ac.jp/world/index.html`で C++ 版・ MATLAB 版を入手可能 (2016 年時点)
- 紹介に留めるが，他にも多様な分析合成法がある
  - AhoCoder [25] (University of Basque Country)
  - Vocaine [26] (Google)

機械学習ベースの特徴量抽出

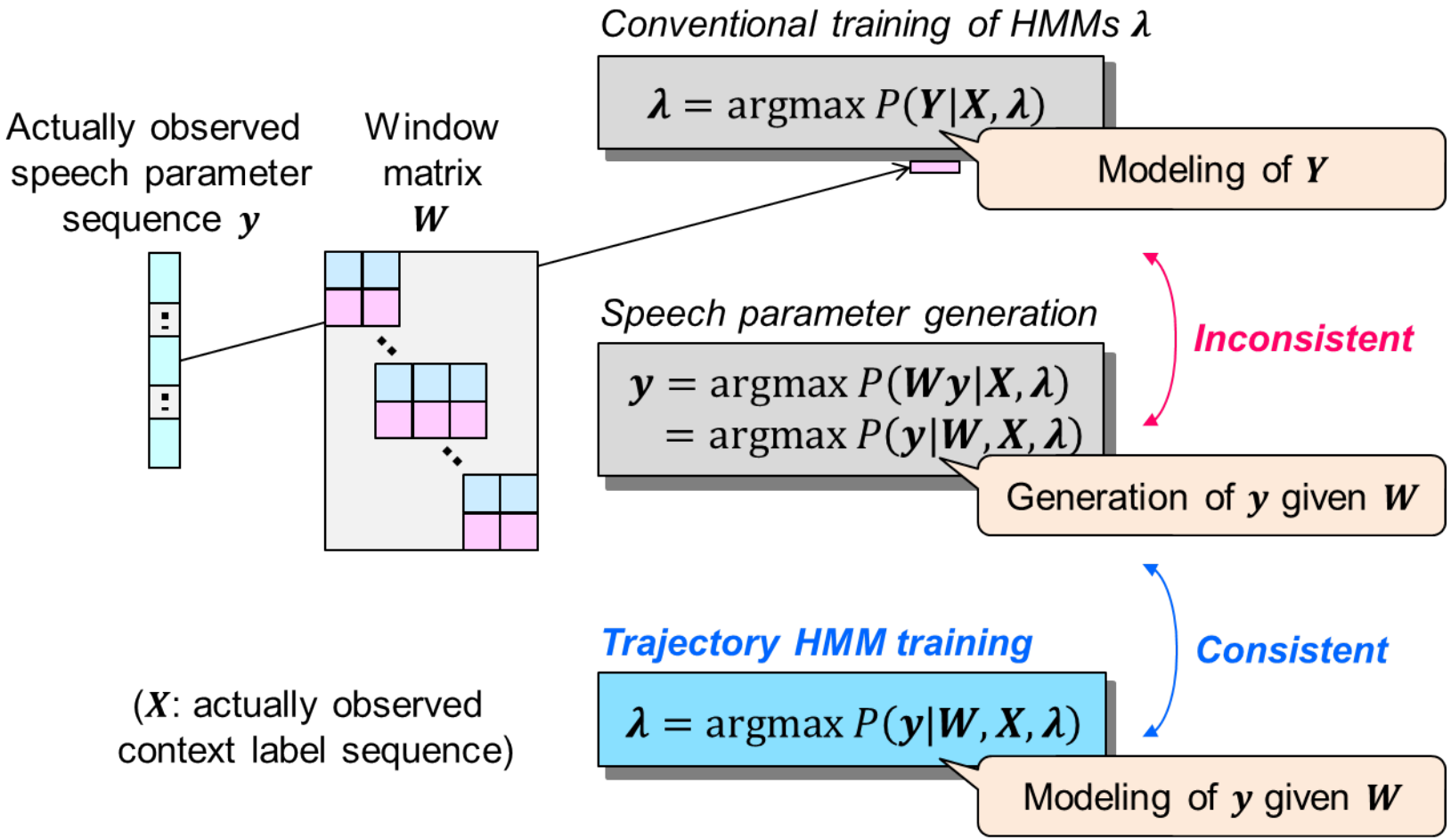
- 信号処理 (STRAIGHT,WORLD など) による特徴抽出
  - 音声波形 -> 信号処理によるスペクトル包絡 & 次元削減
- 機械学習による特徴抽出
  - 音声波形 -> 機械学習による次元削減，F_0 抽出
  - 深層学習の発達により活発化
  - PCA, auto-encoder [11, 12]

音響モデル

- 入力特徴量と出力特徴量を如何に効果的に対応付ける？
  - Trajectory model [13] ... 静的・動的特徴量を考慮した系列モデリング
  - Additive/hierarchical model [14, 15] ... 複数モデルによる加算構造
  - LSTM [27] ... 動的特徴量の自動学習

学習・生成における矛盾

- HMM 音声合成では，生成時に静的・動的特徴量を考慮するのに対し，学習時に無視する．



HMM から trajectory HMM へ

- Trajectory HMM: 静的・動的特徴量の制約 (行列 $\boldsymbol{W}$) の下，系列をモデル化
- $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ から $P(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\lambda})$ は解析的に導出可能
  - Viterbi 状態系列 $\hat{\boldsymbol{q}}$ で近似

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \sim P(\boldsymbol{Y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{Y}|\boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}) \tag{3}$$

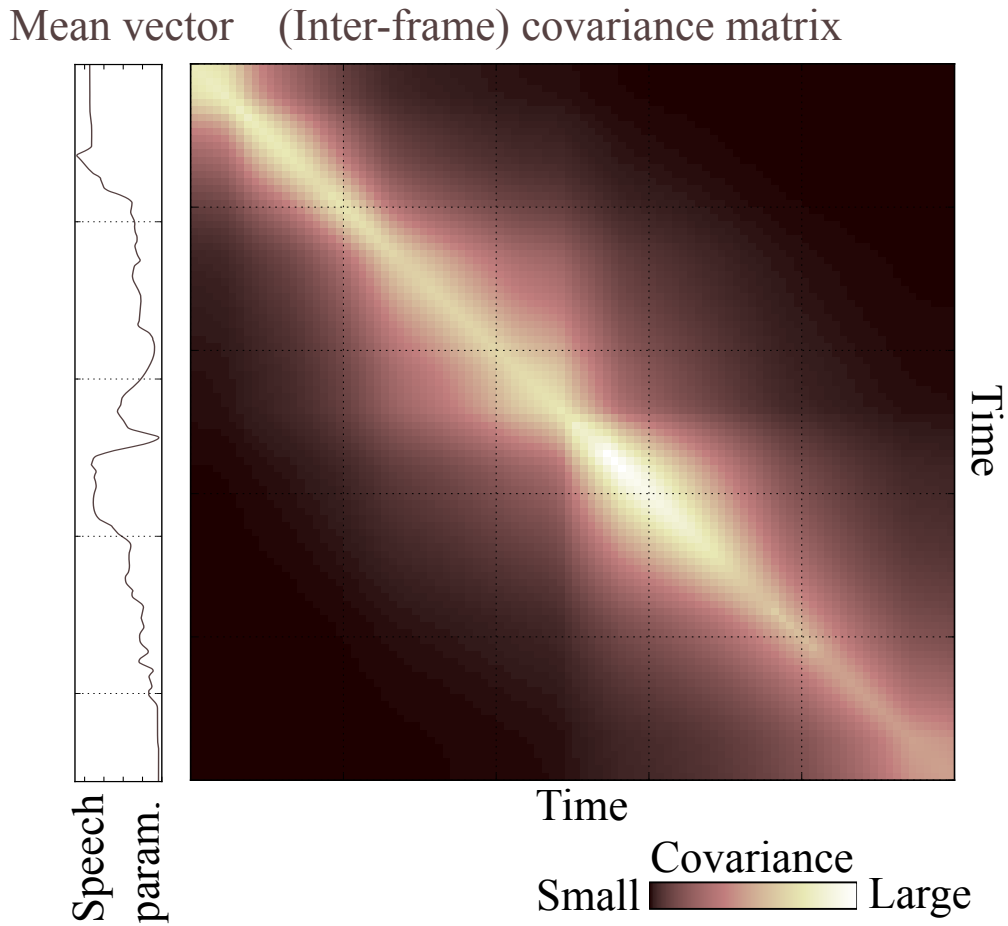
  - $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ を利用して導出

$$\mathcal{N}(\boldsymbol{Y}|\boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}) = \frac{1}{Z}\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}) = \frac{1}{Z}P(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\lambda}) \tag{4}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W}\right)^{-1}, \boldsymbol{\mu}_{\hat{\boldsymbol{q}}} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{E}_{\hat{\boldsymbol{q}}} \tag{5}$$

トラジェクトリモデルのパラメータ

- 平均 $\boldsymbol{\mu}_{\hat{q}}$ は最尤パラメータ生成で得られるパラメータに一致
- 共分散行列 $\boldsymbol{\Sigma}_{\hat{q}}$ は動的特徴量を考慮した時間依存性を表現

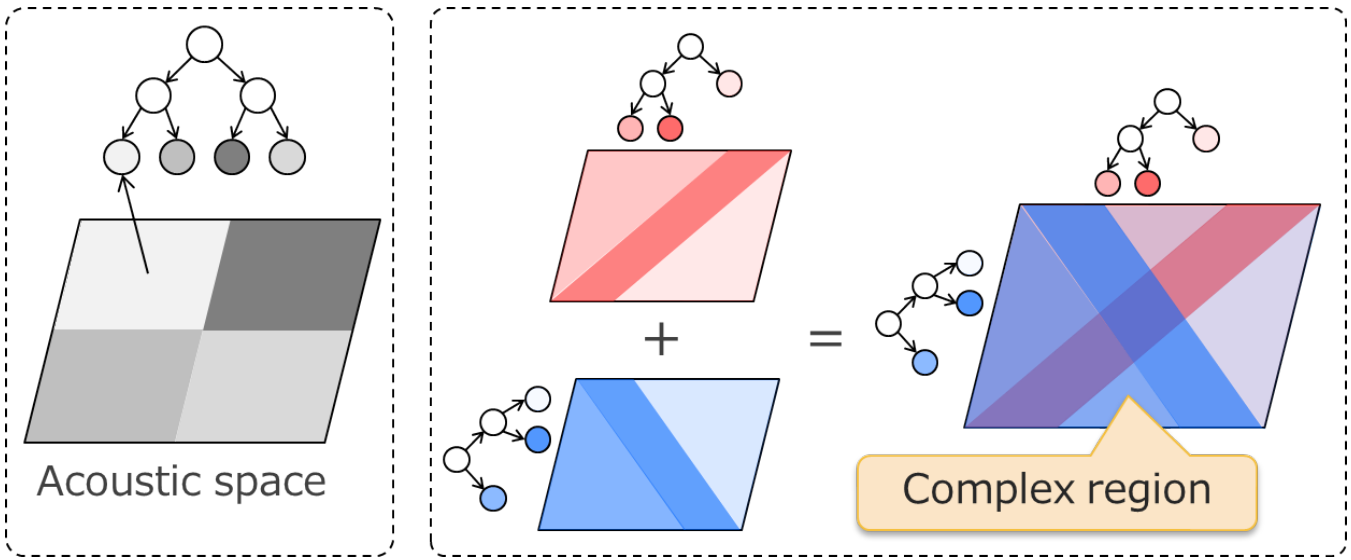


トラジェクトリモデルの改良

- トラジェクトリ HMM を始まりとして、多様なトラジェクトリモデルが提案されている
  - Trajectory HMM [13]
  - Trajectory GMM (同時 [2] ・条件付 [28] ・周辺)
  - Trajectory DNN [29]
  - Latent trajectory HMM [30] / GMM [31]
  - Factor-analyzed trajectory HMM [32]
- 関連研究として、フィルタ構造を仮定するモデルもある
  - Hidden trajectory model [33] ... 隠れ変数系列の moving average process を仮定
  - Autoregressive HMM [34] ... 音声特徴量系列の autoregressive process を仮定

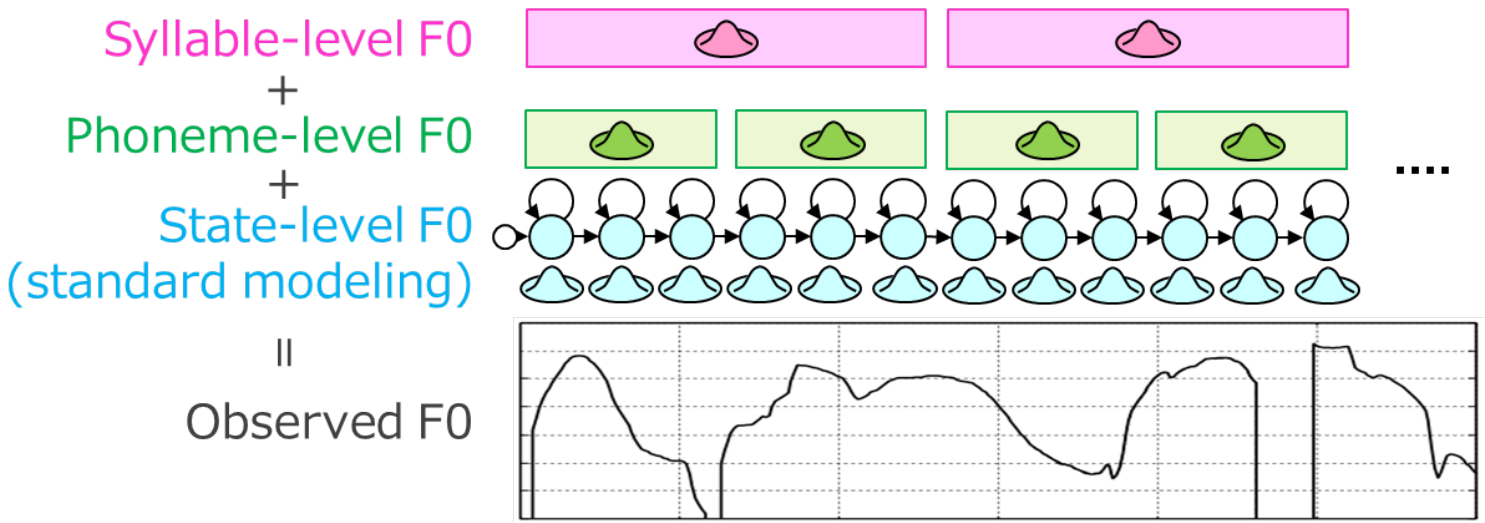
加算モデル [14]

- テキスト音声合成では多様なコンテキストが必要
- 通常、コンテキスト全体に対して影響の強い要素で音響空間を分割 → 影響の小さい要素（弱いコンテキスト）が無視される
- コンテキストの部分的な加算構造により、あらゆるコンテキストを考慮 → 音声特徴量 $\boldsymbol{y} = \boldsymbol{y}_{\text{factor1}} + \boldsymbol{y}_{\text{factor2}} + \cdots$



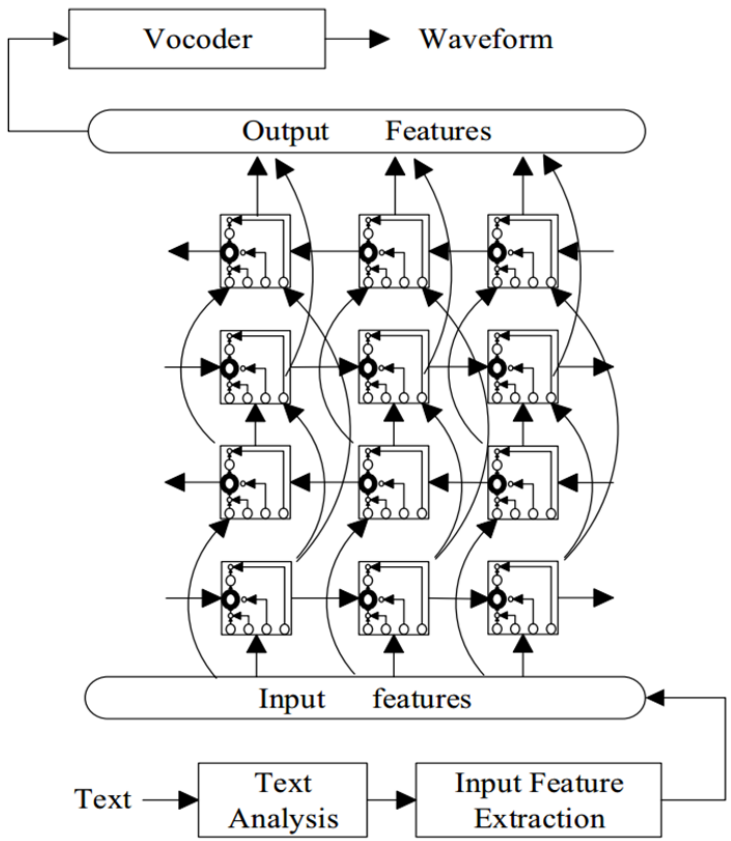
階層モデル [15, 35]

- F_0 のような超分節的特徴は、長い時間区間に影響する
  (例：日本語は単語より長いフレーズ単位で F_0 を決定)
  $\rightarrow$ 通常の音素単位の予測より長い時間単位での予測が有効
- 音素・ワード・シラブル・単語等の階層モデルで F_0 を表現



リカレント構造を持った音響モデル

- リカレント構造を持ったニューラルネットワークにより，動的特徴量計算をネットワークに内包
- 例：Recurrent neural network, Long short-term memory



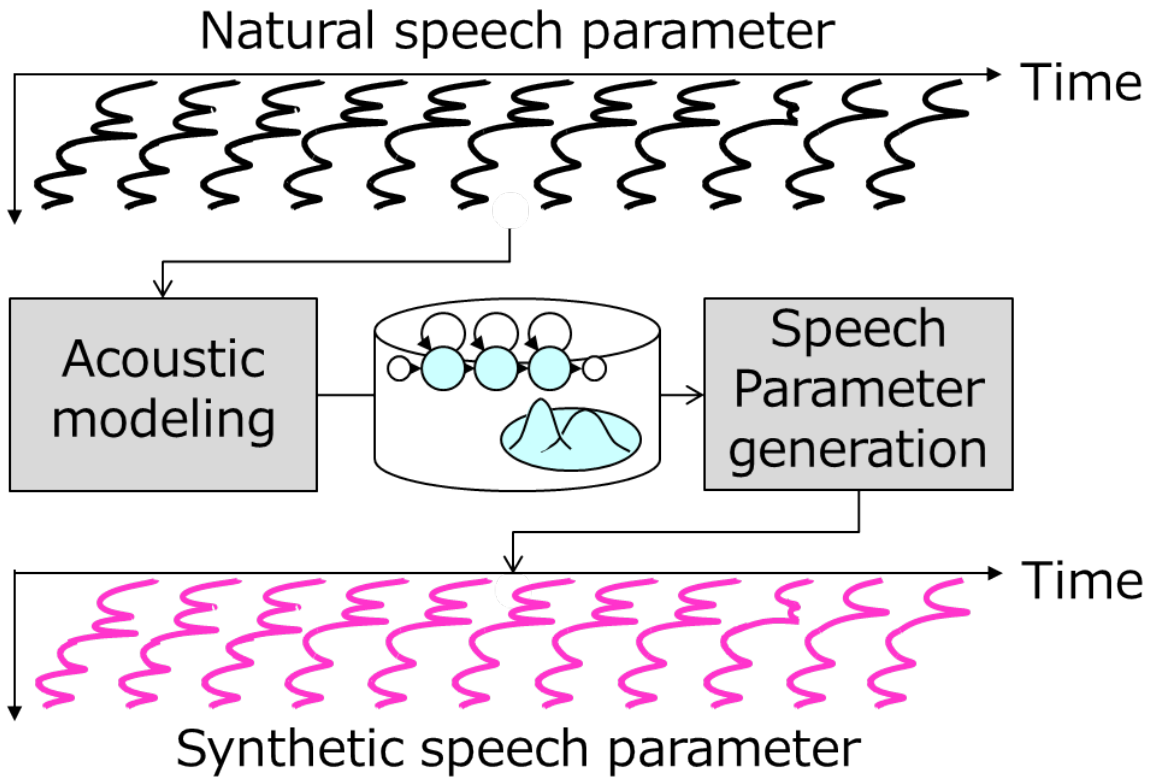
音声合成・変換における多様な DNN 構造

- 性能向上のため、多様な DNN 構造が検討されている。

- Uni-directional LSTM [27] ... 低遅延の合成処理
- Simplified-LSTM [36] ... 低フットプリント・高速演算
- Three-way RBM [37] ... 音響・音韻・話者情報の同時モデリング
- Attention model [38] ... HMM フリーなテキスト－音声 alignment
- Language/speaker LSTMs [39] ... 言語と話者の factorization
- Highway net [40] ... より deep なモデルへ
- Complex-valued net [41] ... 複素スペクトルのモデル化
- Dilated CNN [42] ... 後述

生成パラメータの過剰平滑化

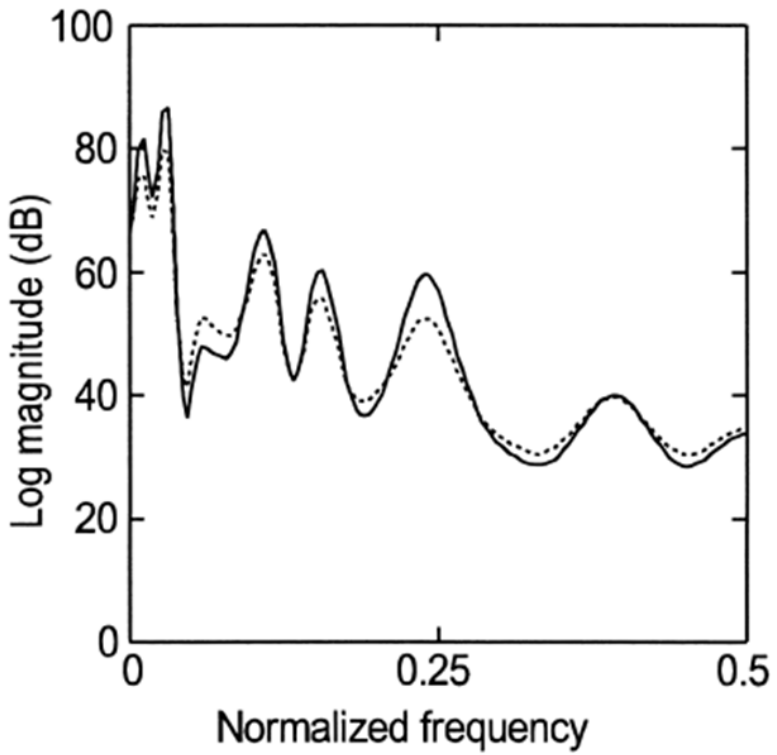
- 統計的生成処理における平均化は、合成音声パラメータを過剰に平滑化させ、音質を劣化させる。
- 平滑化により消失した何かを復元することで音質が改善



Cepstrum emphasis [6]

- ルールベースの強調処理
- ハイパーパラメータ β を用いて、ケプストラムを変形

$$c'_t(m) = \beta c_t(m) \ (\beta >= 1, m >= 2) \tag{6}$$



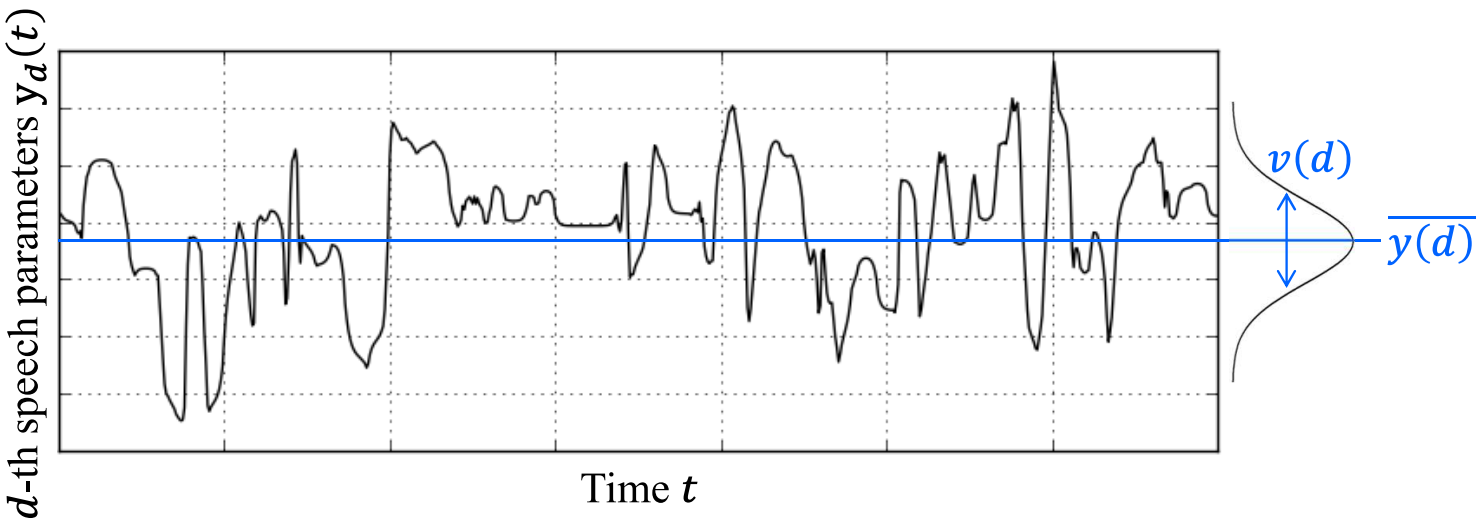
3 ポストフィルタリングの効果（点線：ポストフィルタリング前 $D(z)$，実線：ポストフィルタリング後 $D(z) \cdot \bar{D}^{\beta}(z), \beta = 0.5$）

系列内変動（Global Variance: GV）[43, 18]

- GV: 音声パラメータ時系列の分散 (= central 2nd moment)

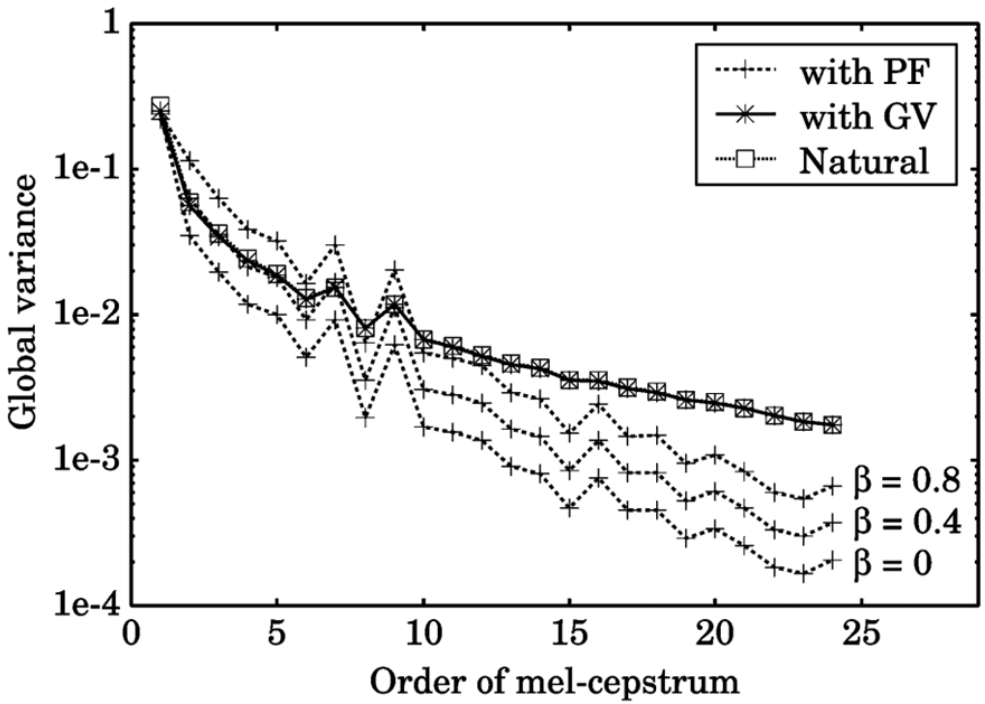
$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(y_t(d) - \bar{y}(d))^2,\ \bar{y}(d) = \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d),$$

- 直感的に言えば「系列の広がり具合」を表す



Cepstrum emphasis (CepEm) vs. GV

- CepEm は、GV ドメインで見るとバイアスの効果
  → 低次のケプストラムにおいて過強調
- GV 補償 [43] は、ケプストラムの次元毎に強調度が異なる
  → 自然音声と合成音声の GV が一致

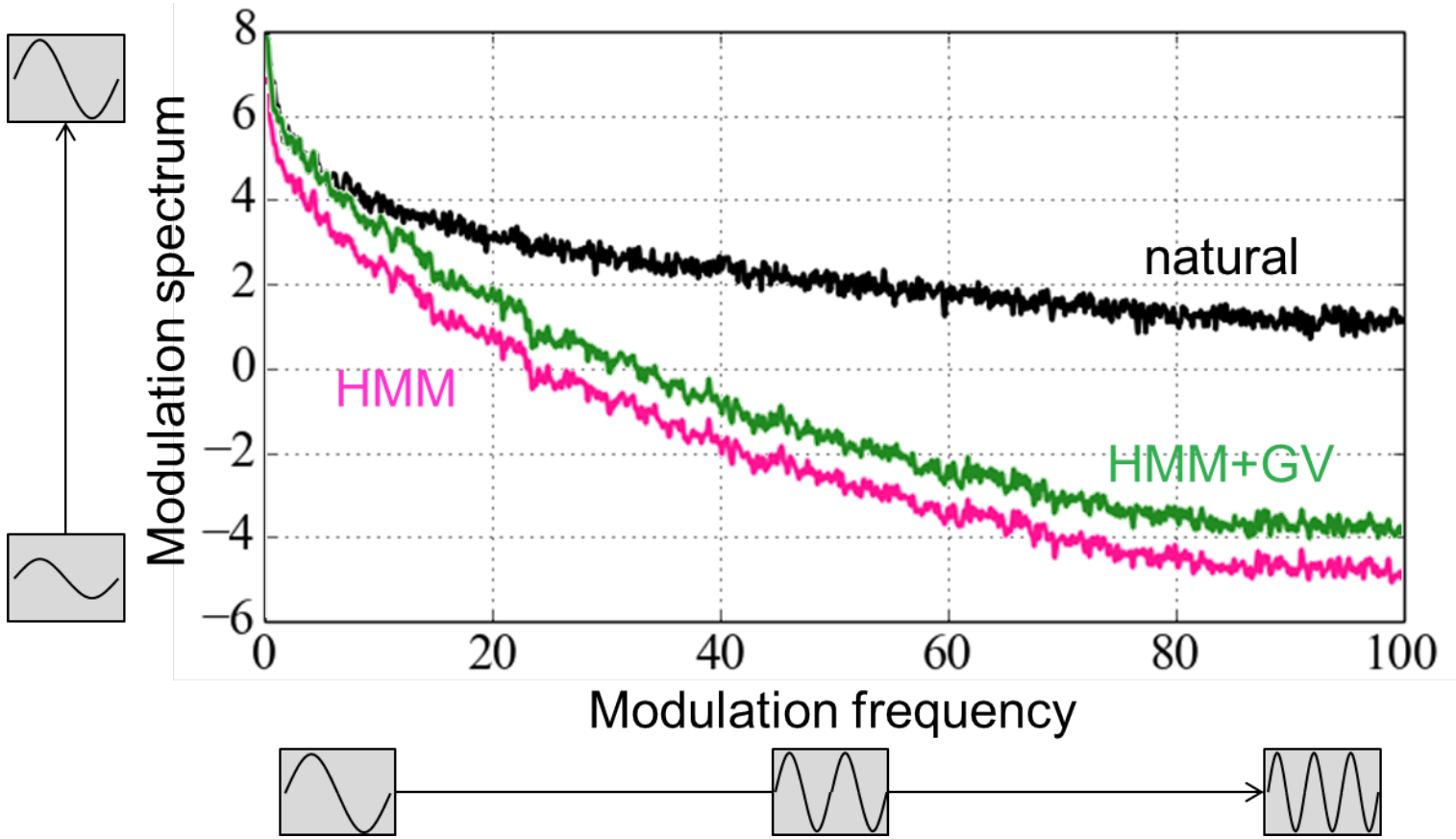


変調スペクトル (Modulation Spectrum: MS) [19]

- 定義：パラメータ時系列の (対数) パワースペクトル
  $\rightarrow$ 時系列の変動（ゆらぎ）を定量化
- 期待される効果：
  - 最尤パラメータ系列は、コンテキスト内で定常（ゆらがない）
  - 自然音声は、同一コンテキストでもゆらぐ

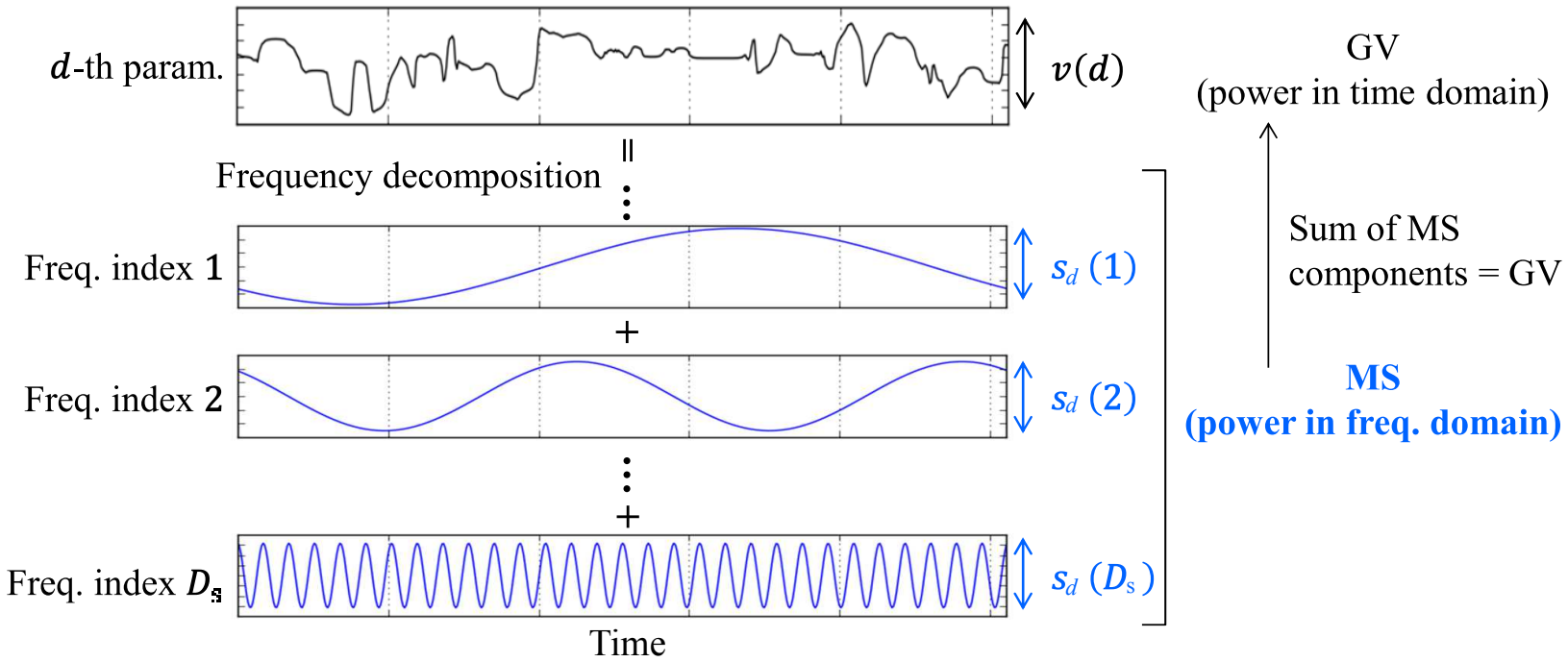
変調スペクトルの例

- GV 強調は、MS ドメインで見るとバイアスの効果
  → 低域変調周波数において過強調



GV vs. MS

- GV も MS も、系列のパワーを計算する
- ただし、MS は周波数要素に分解後に計算され、MS の和＝GV になる



CepEm vs. GV vs. MS

- Cepstrum emphasis, GV, MS は下表の関係性を持つ
- (Mod. freq. = Modulation frequency)

	Cep. emphasis	GV	MS
Variable	Scalar	Vector	Matrix
Feature-dependent emphasis?	No	**Yes**	**Yes**
Mod. freq.-dependent emphasis?	No	No	**Yes**

GV/MS を考慮した音声パラメータ生成 [43, 45]

- どうやって GV/MS を補償する？
- 音響モデルと GV/MS モデルの Product-of-Experts [44].
  - 複数モデルの AND オペレーションに相当
  - 各モデル $\boldsymbol{\lambda}_*$ は独立に学習．ハイパーパラメータは，重み ω

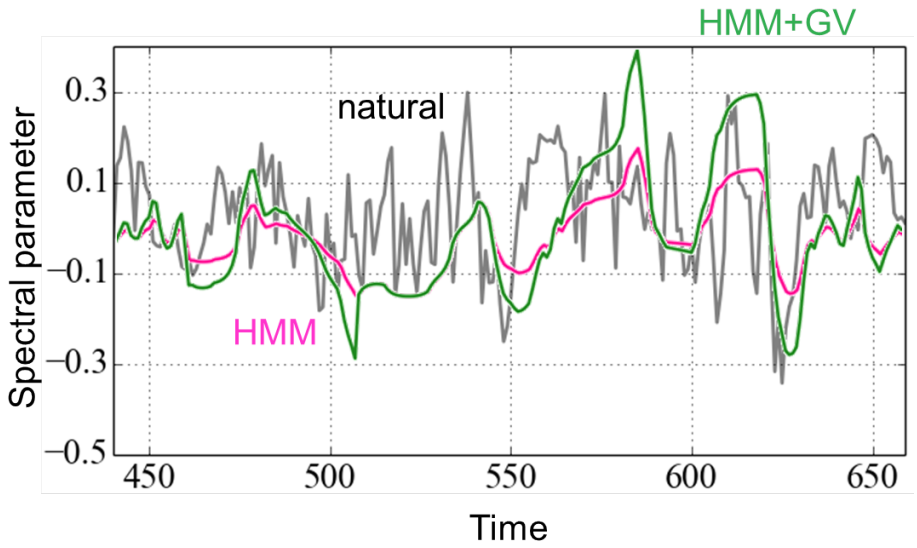
MS(GV) を考慮した音声パラメータ生成

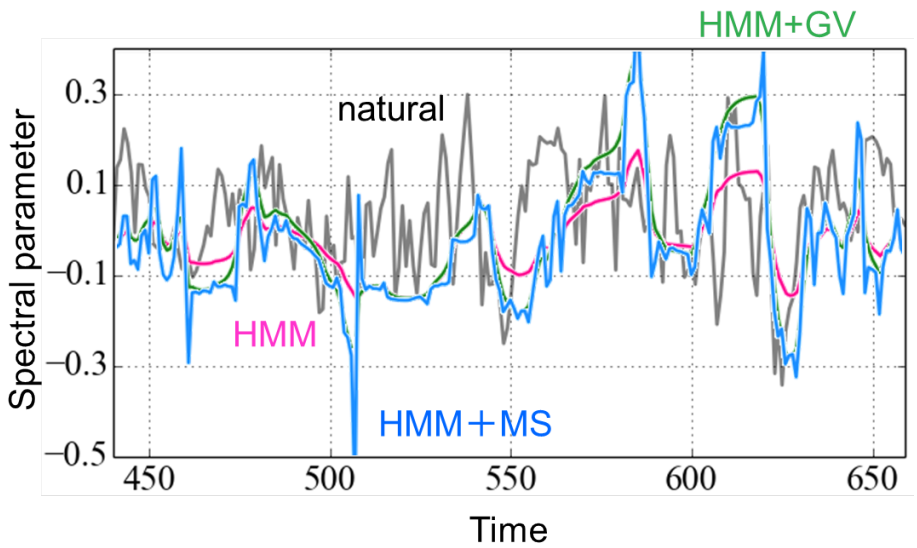
$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \mathrm{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\lambda}_{\mathrm{HMM}}) P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \boldsymbol{\lambda}_{\mathrm{MS}})^{\omega}$$

- MS $\boldsymbol{s}(\boldsymbol{y})$ は，$\boldsymbol{y}$ に関する 2 次式
- 上式の対数は 4 次式 $\rightarrow$ 反復法で推定

Example

- GV は広がり、MS は振動を復元する。音質は MS $>$ GV。





GV/MS を考慮したトラジェクトリモデル学習

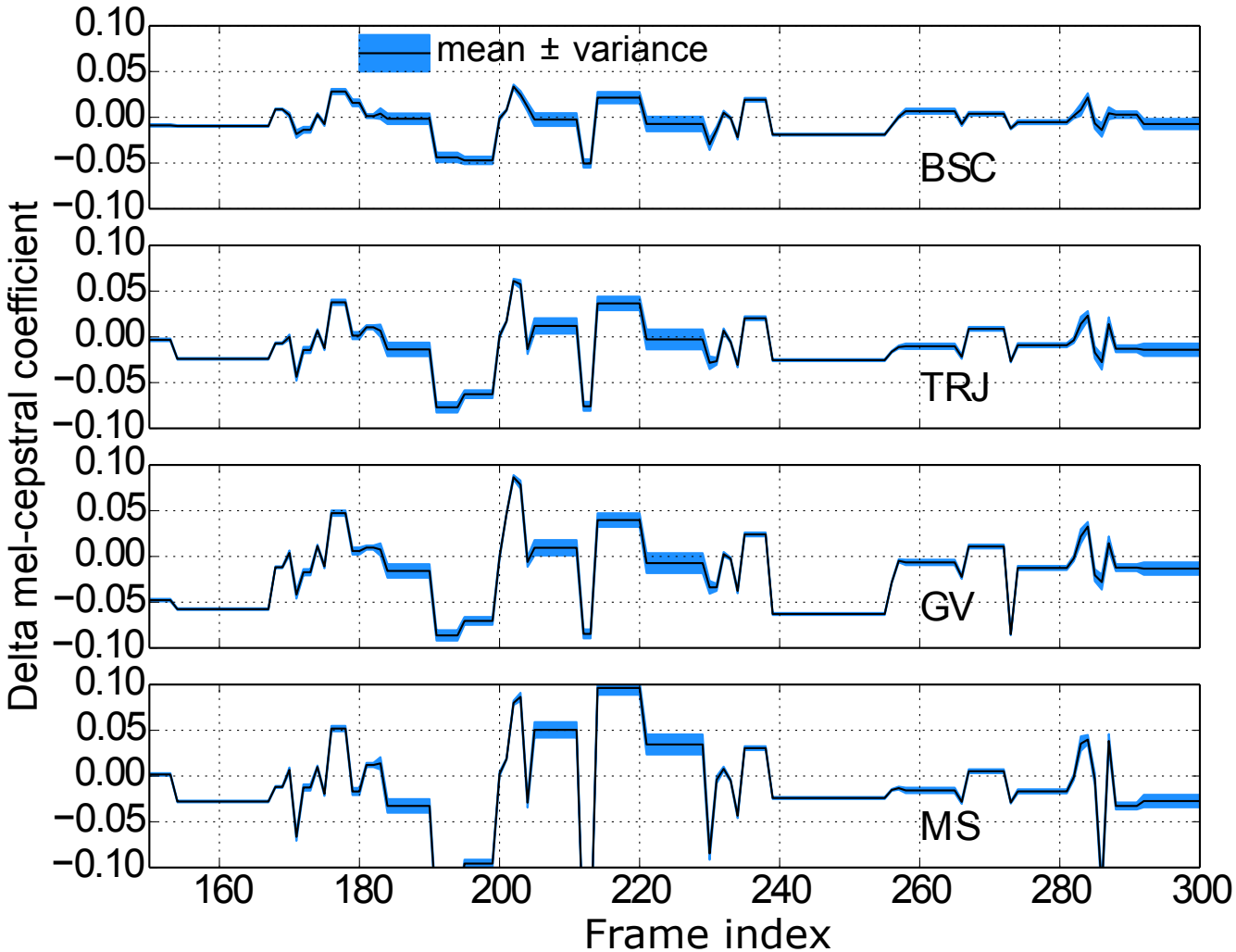
- MS(GV) を考慮したパラメータ生成は、高品質だが，反復法を必要とするため生成に時間がかかる
  - 生成時間を増やすことなく、MS(GV) の音質改善効果を得られないか？
- **MS(GV) 制約付きのトラジェクトリモデル学習** [46, 28, 47]
  - 生成ではなくモデル学習に MS(GV) を組み込む
  - 生成される音声パラメータ $\hat{\boldsymbol{y}}$ が MS(GV) を復元するように音響モデル $\boldsymbol{\lambda}_{\mathrm{HMM}}$ を学習

MS(GV) を考慮したトラジェクトリモデル学習

$$\boldsymbol{\lambda}_{\mathrm{HMM}} = \operatorname{argmax} P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\lambda}_{\mathrm{HMM}})\, P(\boldsymbol{s}(\boldsymbol{y})\,|\hat{\boldsymbol{y}}, \boldsymbol{X}, \boldsymbol{\lambda}_{\mathrm{MS}})^{\omega}$$

HMM の出力確率の系列の例

- MS(GV) によって平均ベクトル系列が振動する（広がる）
  → そこから生成される音声パラメータも振動する（広がる）



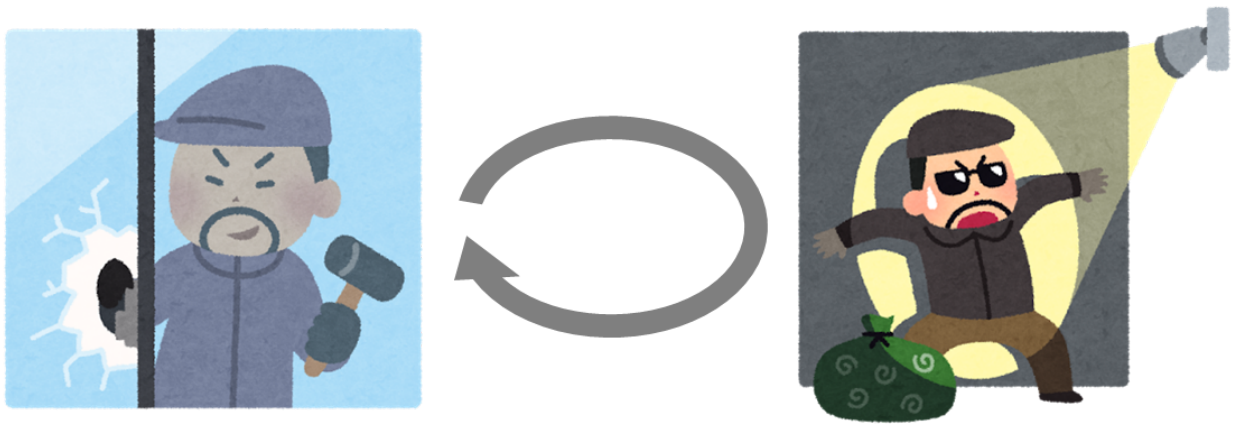
BSC: 通常の HMM 学習，TRJ: トラジェクトリ学習

変調スペクトルの利用

- 音声特徴量の前処理 [48]
  - 音声知覚に寄与しない MS を除去
- ポストフィルタ [19]
  - 合成・変換処理に依らず適用可能かつ低遅延 (125msec)
  - スペクトル、F0、継続長に適用可能
- パラメータ生成 [45]
  - 処理時間は長いが超高品質
  - **TTS コンペで世界最高品質 (6 言語中 3 言語の自然性)** [48]
- 音響モデル学習 [28, 47]
  - 高品質かつ高速合成・変換
- 音響モデル適応
  - 少量の音声データでも高品質化
- 波形加工に基づく音声パラメータ変換 [49]
  - **VC コンペで世界最高品質 (話者性)** [49]

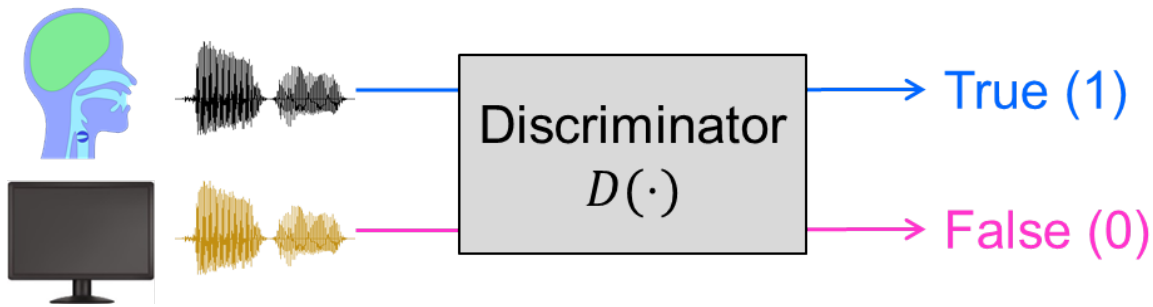
特徴量の解析的設計から自動設計へ

- 解析的特徴量の補償は確かに有効
  → 特徴量の設計が非常に面倒 (だるい)
  → 従う確率分布の設計が非常に面倒 (だるい)
- 発見と設計を機械学習に任せられないか？
  → **Anti-spoofing** に敵対する音声合成 [20]



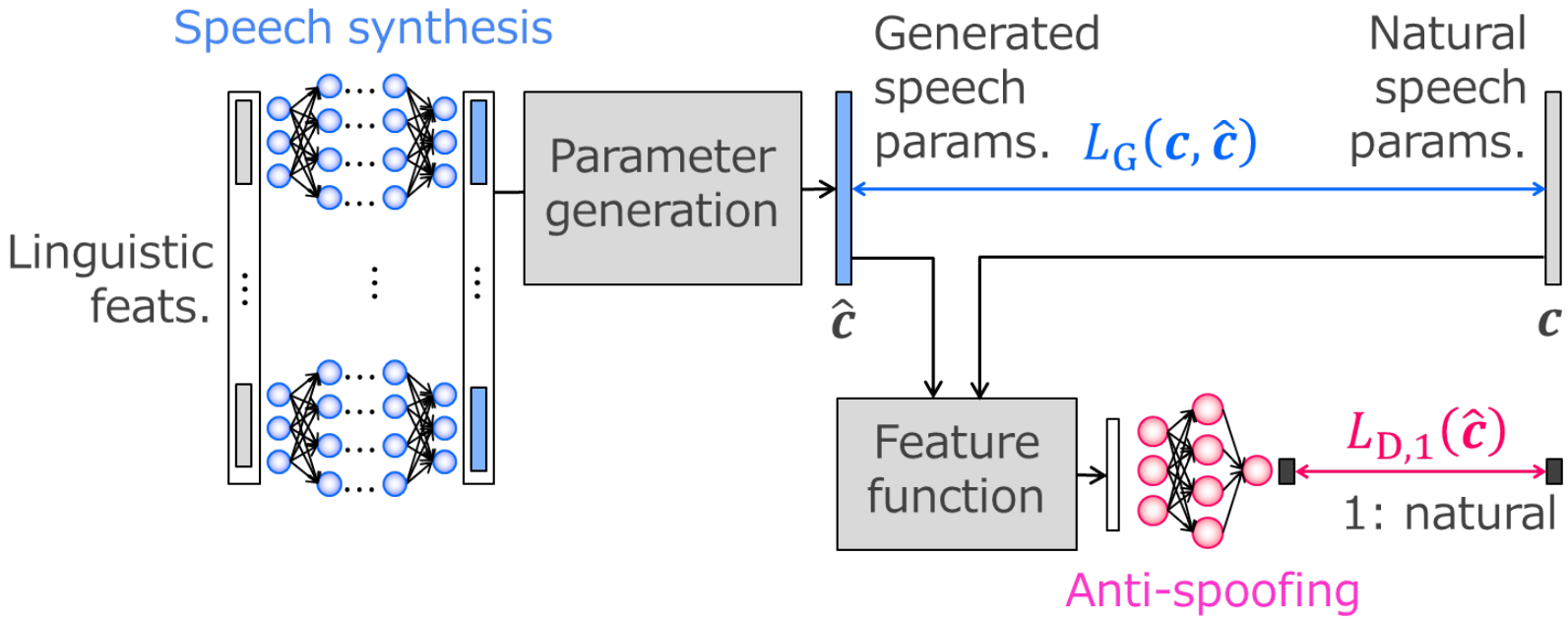
Anti-spoofing verification (ASV) [50]

- 合成音声による「声のなりすまし」を防ぐ技術
  → 音声合成・変換の高品質化に伴い発達
- 自然音声と合成音声を学習データとして識別器を学習
  → 自然音声と合成音声の違いを見つけて識別



- ASV を騙すように音声合成をアップデートすれば良い？
  → 自然音声と合成音声の違いを見つけて補償

ASV に敵対する音声合成 [20]



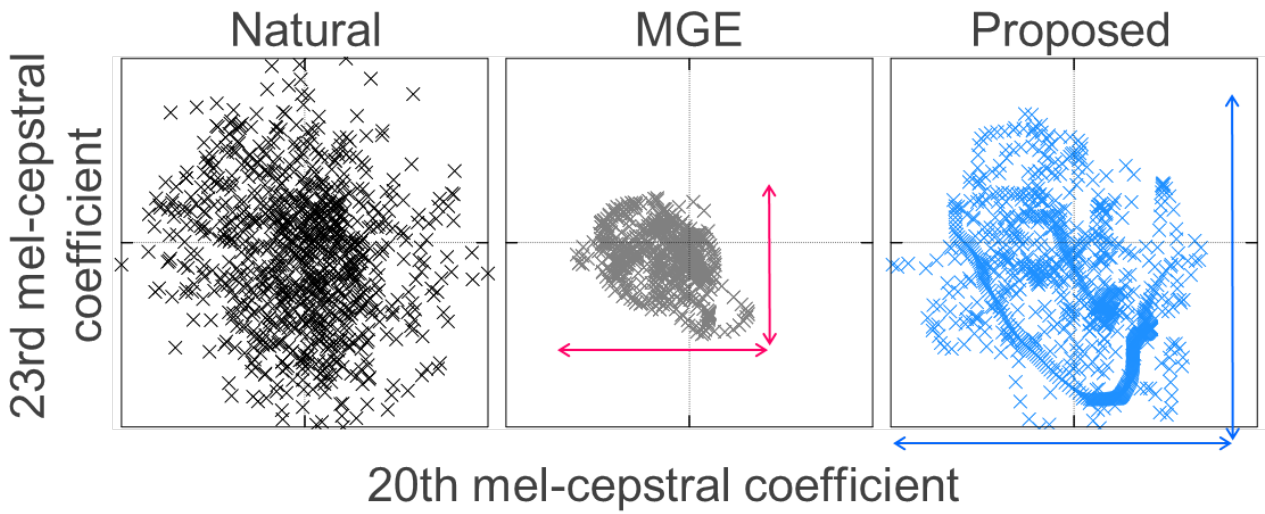
ASV に敵対する音声合成のコスト関数

$$L = L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega L_{\mathrm{D},1}(\hat{\boldsymbol{y}}) \rightarrow \text{最小化}$$

- 1 項目は生成誤差を最小化
- 2 項目は生成パラメータを自然音声と識別させる

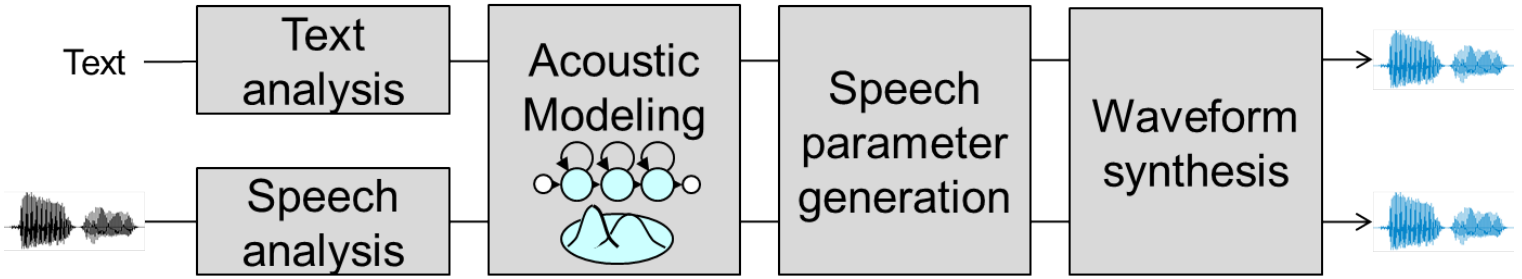
特徴

- 学習アルゴリズム：マルチタスク学習＆敵対的学習
  - マルチタスク学習：複数タスクの共通要因を学習
  - 敵対的学習：2 つのデータセットの分布をそろえる



- 特徴量設計
  - 解析的特徴量 (GV や MS) を直接利用可能
  - 設計そのものを DNN に任せることも可能
- 正則化としての敵対学習

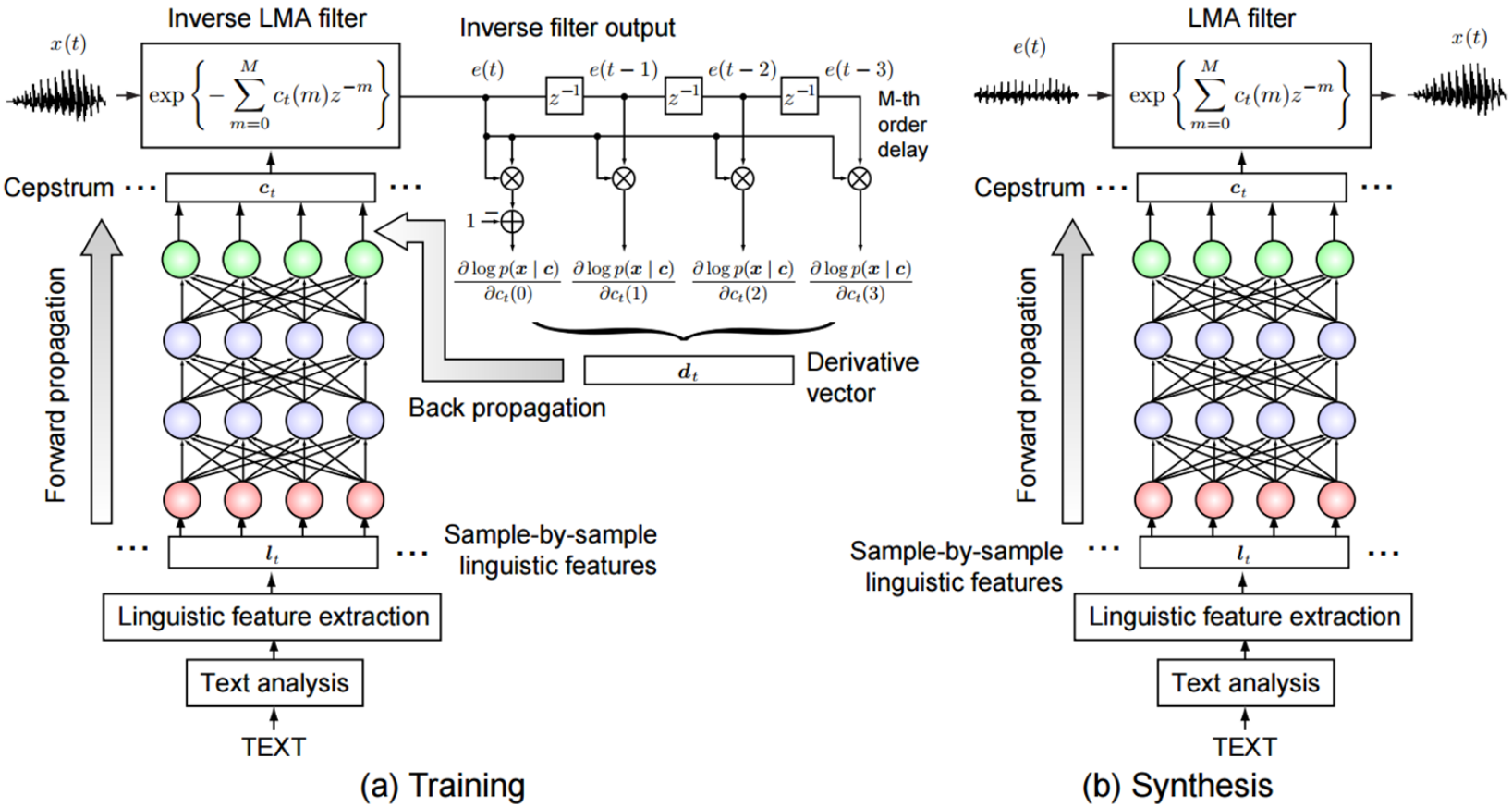
同時最適化



- 各モジュールの最適化が全体の最適化とは限らない
- 複数モジュールを同時最適化
- ここでは，acoustic modeling, speech parameter generation, waveform generation を同時最適化する手法を紹介

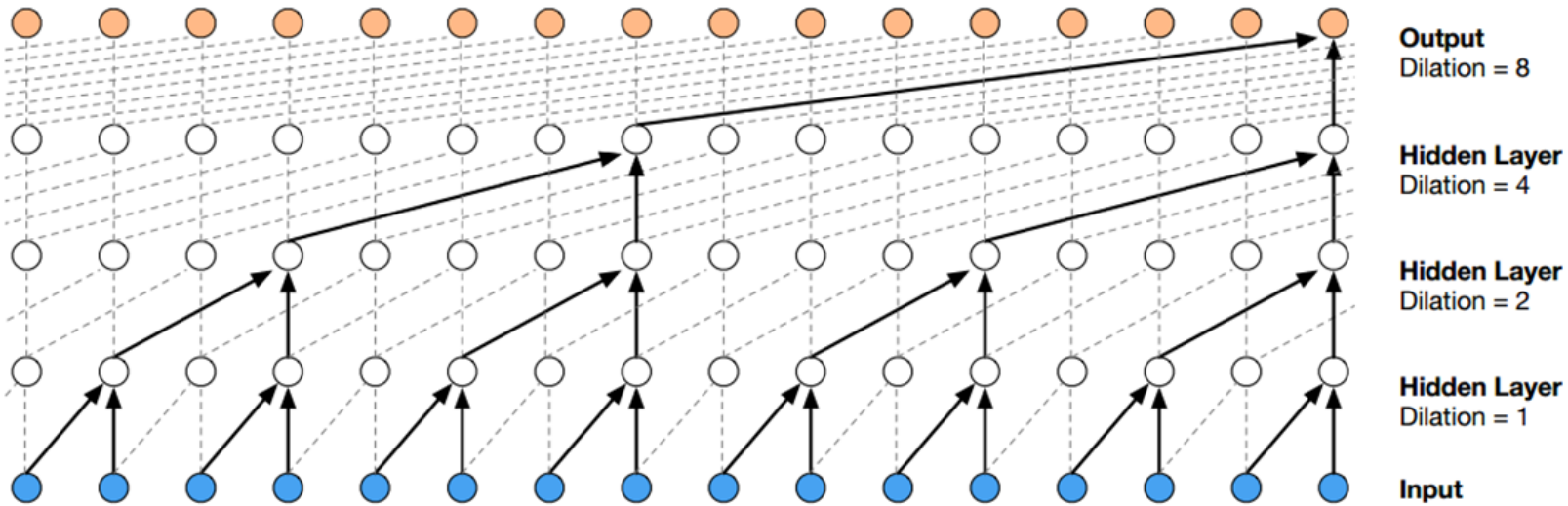
Direct waveform modeling [21]

- source-filter モデルを仮定して，生成パラメータ $\boldsymbol{y}$ から得られる音声波形 $\boldsymbol{x}$ の尤度 $P(\boldsymbol{x}|\boldsymbol{y})$ を最大化



WaveNet [42]

- 波形生成を量子化波形に対する多クラス分類問題と捉える
- フレーム分析と source-filter モデルを仮定せず，1 サンプル毎に波形を生成
- Dilated CNN を用いて，Autoregressive process を表現 ($1 \cdots T-1$ サンプルの音声波形から T フレーム目の音声波形を推定)



Section 4

どう応用する？

応用例

- 実際には複数分野で使用される技術だが、本講義では以下の様に分類する

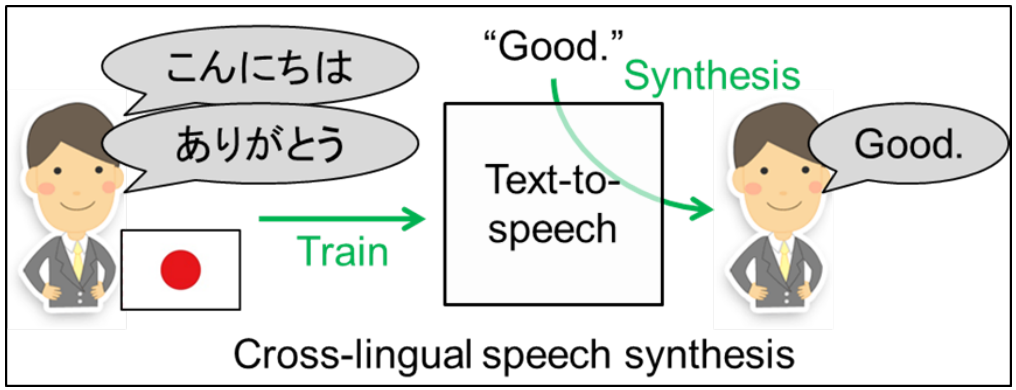
- 音声翻訳のための音声合成・変換
  - クロスリンガル音声合成
  - パラ言語翻訳・感情音声合成
- 多言語化・多様な個人性のための音声合成・変換
  - 教師なし音声合成・転移学習
  - 因子モデル
- 言語教育のための音声合成・変換
  - ノンネイティブ音声合成・変換

音声翻訳で何を翻訳すべき？

- 話し言葉に含まれる情報
  - 言語情報…文字により標記される要素
  - パラ言語情報…文字で標記できないが、発話者が意図的に含ませた要素（例：態度、発話様式）
  - 非言語情報…文字で標記できず、発話者が意図せず含ませた要素（例：声の個人性、身体状態）
- 言語を超えた情報の翻訳
  - 言語情報…ステレオタイプな機械翻訳
  - パラ言語情報 …パラ言語情報の翻訳
  - 非言語情報 …クロスリンガル音声合成・変換

非言語情報の翻訳

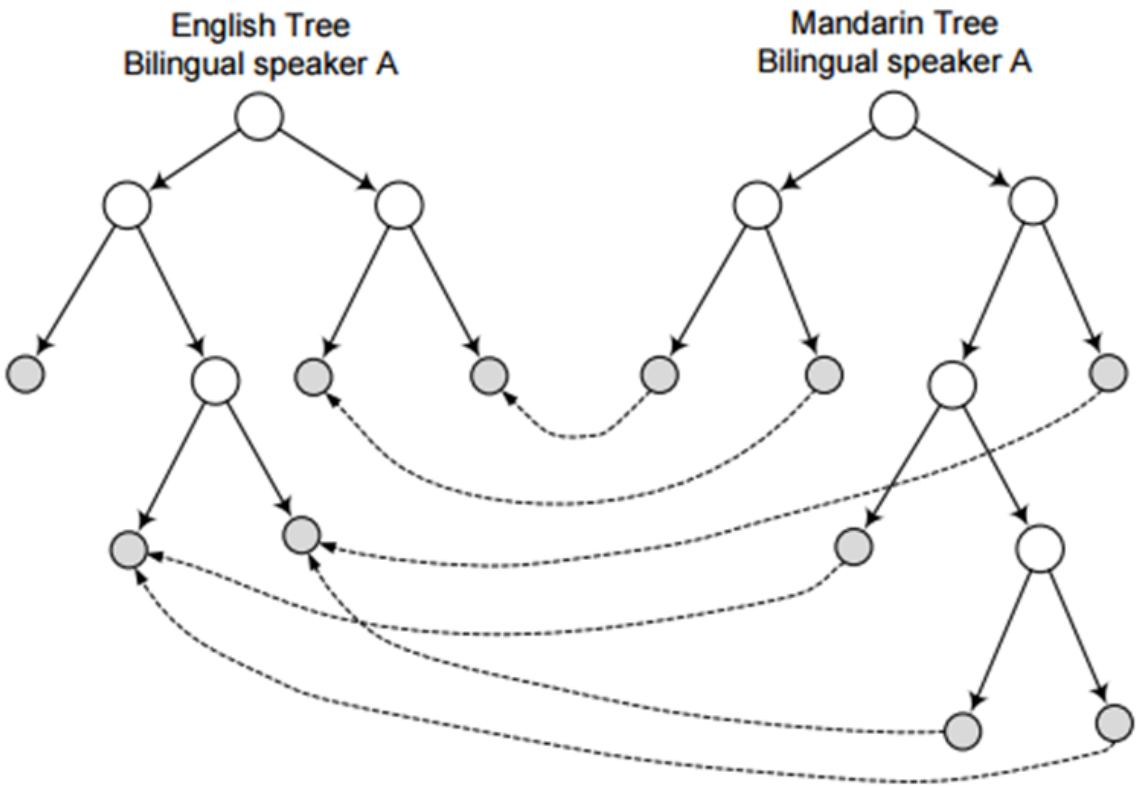
- 発話者の母語音声と異なる言語の音声を合成したい
  → クロスリンガル音声合成・変換技術



- 何が問題？
  → 当該話者によるパラレルデータが存在しない！
  (例：日本語話者の日本語音声はあるが、同話者による英語音声がない)
- ある言語の音声特徴量 or 音響モデルを別言語の音声特徴量 or 音響モデルに変形
  - モデルマッピングに基づく方法
  - 特徴量マッピングに基づく方法

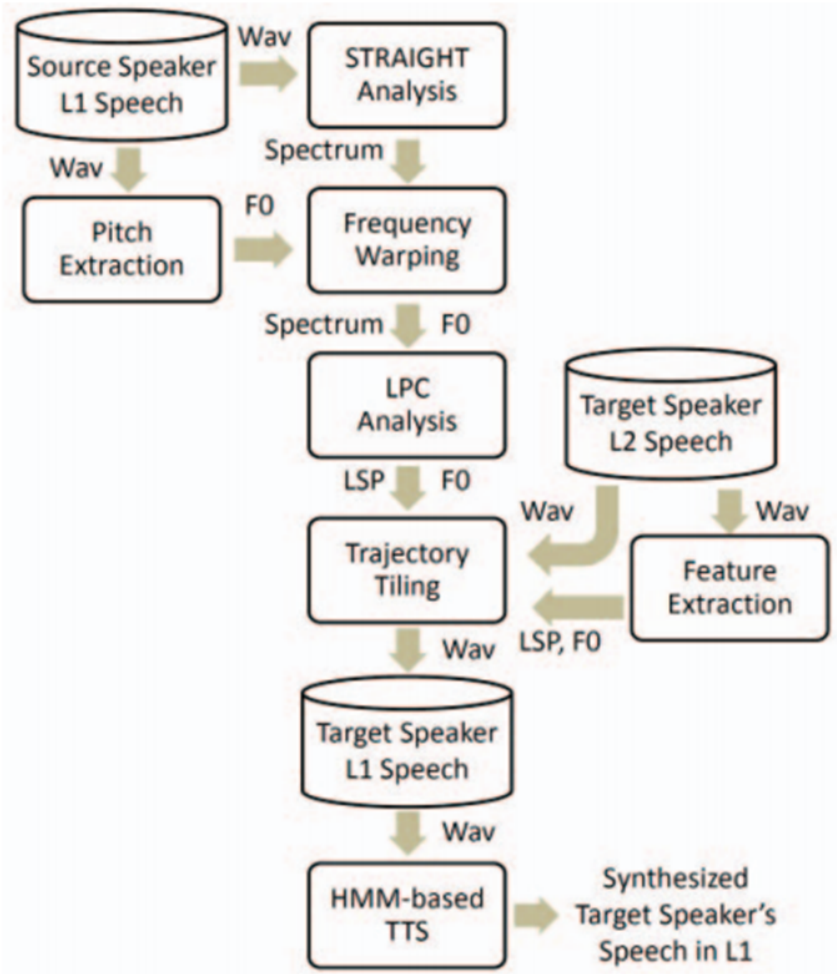
HMM マッピングに基づく方法 [52]

- バイリンガルデータを用いた HMM 状態マッピング
  - 1. 言語 A、B を話すバイリンガル話者のデータから、各言語の HMM を学習
  - 2. HMM 状態間 KL 距離を測り、距離最小の状態ペアを決定 (DNN 版 [51])
  - 3. 言語 A を話すモノリンガル話者の HMM を学習し、2 のマッピングルールで言語 B の HMM に変換



特徴量マッピングに基づく方法 [53]

- 特徴量間の距離関数を定義 (例: スペクトル・ F0 歪み)
- 距離関数を最小化するように、言語 A の音声特徴量を言語 B の音声特徴量に並び替え

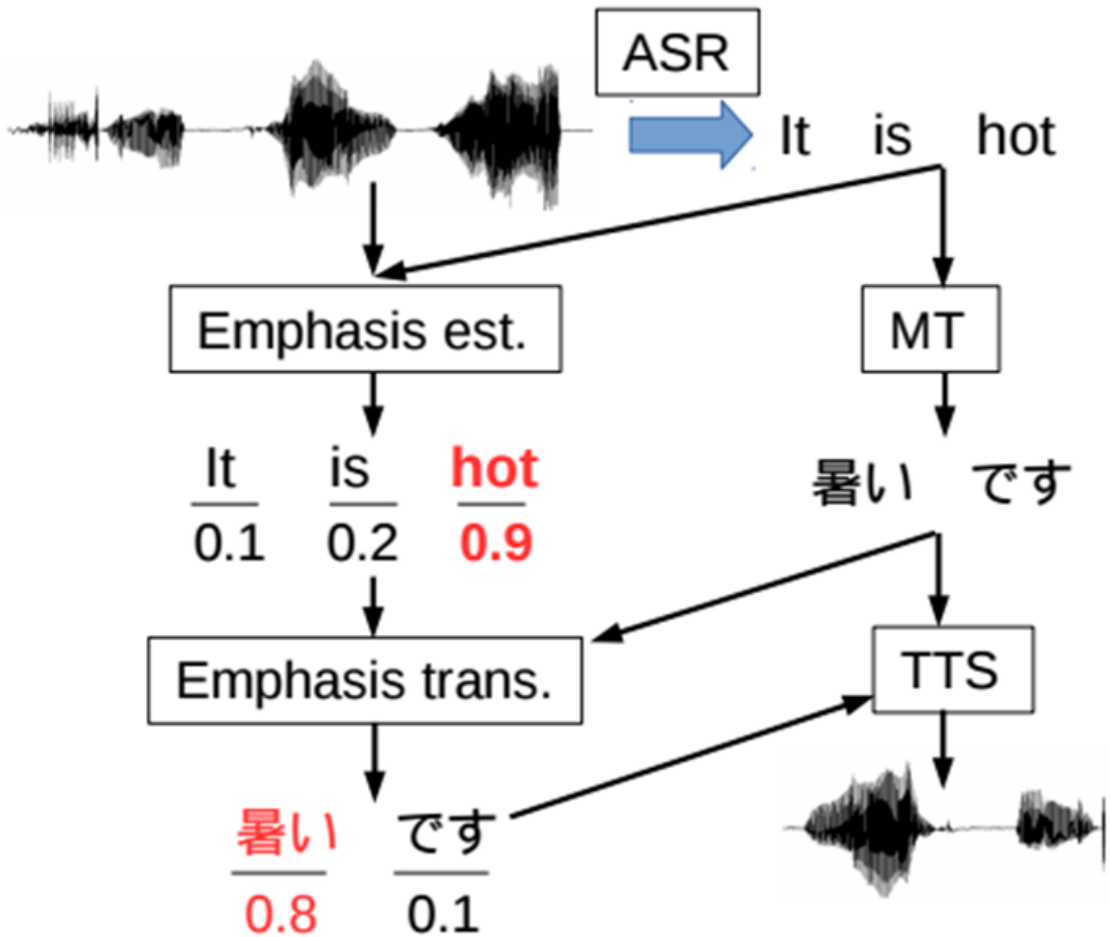


パラ言語の翻訳

- パラ言語情報を認識して、翻訳・合成
- どうやって認識 (・翻訳) する？

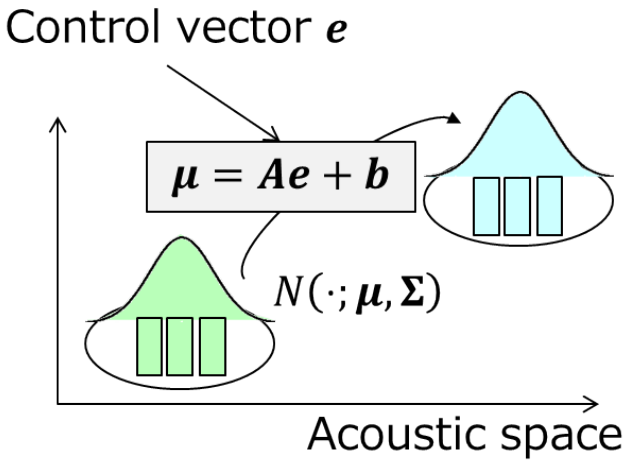
  → 音声認識・機械翻訳と非同期 (文単位など粗い時間単位)

  → 音声認識・機械翻訳と同期 (より細かい時間単位へ) [54]



MR-HMM [56]

- 制御ベクトル (例：感情の 1-of-K vector) で統計量を制御
- コンテキスト加算モデル [14]、感情加算モデル [55] に拡張



- 音声翻訳での利用 [54]
  - 音声yからパラ言語eを認識 $e = \mathrm{argmax}\, P(y|e, \lambda)$
  - パラ言語eから音声yを生成 $y = \mathrm{argmax}\, P(y|e, \lambda)$
  - 異言語間のパラ言語を翻訳（例：CRF、att. model）

多様な言語・話者性を実現する音声合成

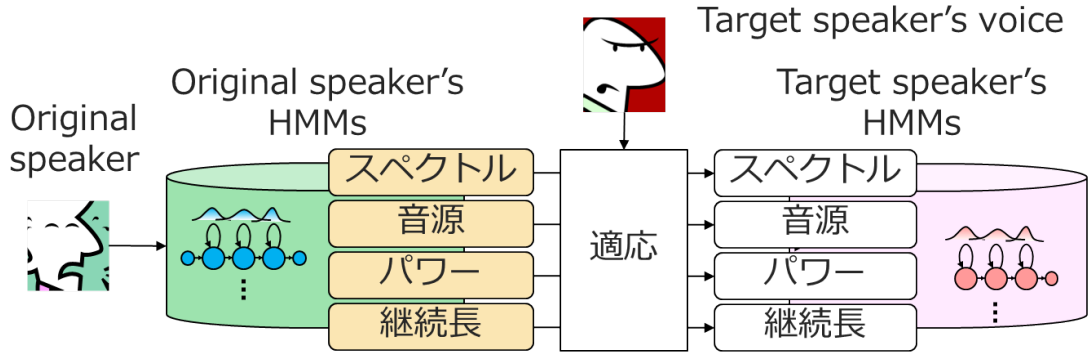
- 存在しうる、あらゆる言語・音声の音声合成を実現するには？
- あらゆる言語
  - 言語知識 (音素セットなど) はテキスト音声合成に必要
    → 方言・希少言語等の音声合成は困難
  - 言語知識なしで音声合成 → 教師なし学習
  - 言語知識の豊富な言語を利用して音声合成 → 転移学習
- あらゆる音声
  - 少量の音声から音声合成を実現できないか？→ モデル適応
  - 音声情報を効率よく表現できないか？→ 因子モデル

言語知識なしの音声合成・転移学習

- 音素セットやテキスト一音素対応なし
  - 声道特徴量の決定木クラスタリングで音素セット推定 [57]
  - 言語知識豊富な別言語の音声認識を利用 [58]
- 音声に対応するテキストなし (音声認識もなし)
  - para-speech のある，言語知識豊富な言語の単語に対応 [59]
- 言語に依存しない言語解析
  - Unicode から統一的な音素セットを推定 [60]
- 低コスト化
  - Crowd sourcing を利用した大量データ収集 [61]

モデル適応

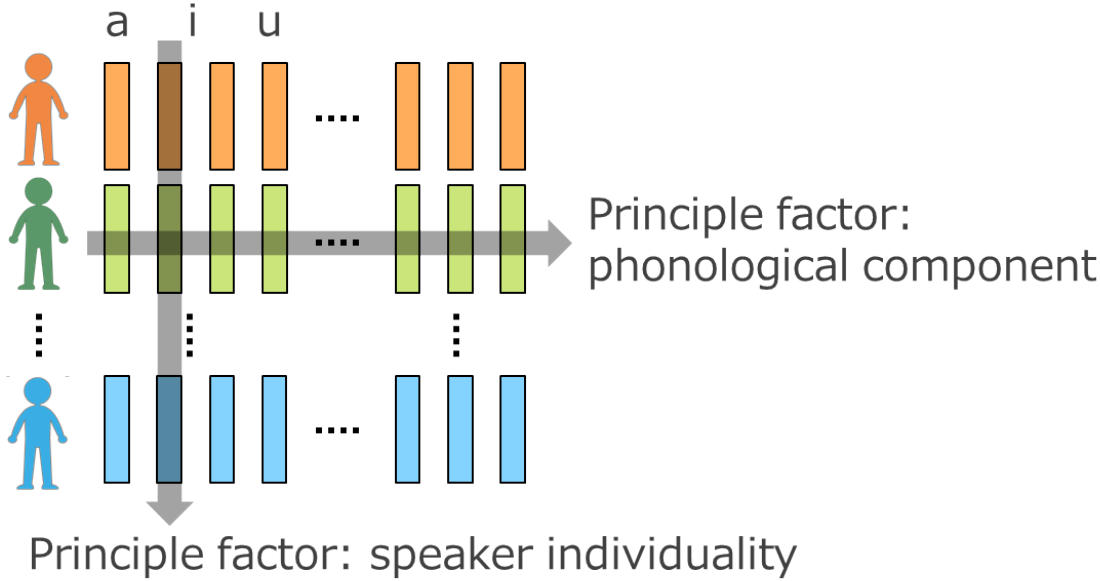
- 学習済みの音響モデルを元に、目標話者の少量データから、目標話者の音声合成を作る技術
  → 学習 (数百文〜) より少量のデータ (数文〜) で **OK**



- HMM 音声合成・ GMM 音声変換・ GPR 音声合成
  - 部分空間のアフィン変換 (空間全体では非線形変換)
  - 例: MLLR $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$. ($\boldsymbol{y}, \boldsymbol{y}'$ が音声特徴量、$\boldsymbol{A}, \boldsymbol{b}$ が適応パラメータ)
- DNN 音声合成・変換
  - 話者識別技術 (i-vector)、話者コード、部分的なモデル更新

因子モデル

- 単一話者データのスペクトル特徴量の主成分 → 音韻性
- 複数話者データのスペクトル特徴量の主成分 → 話者性

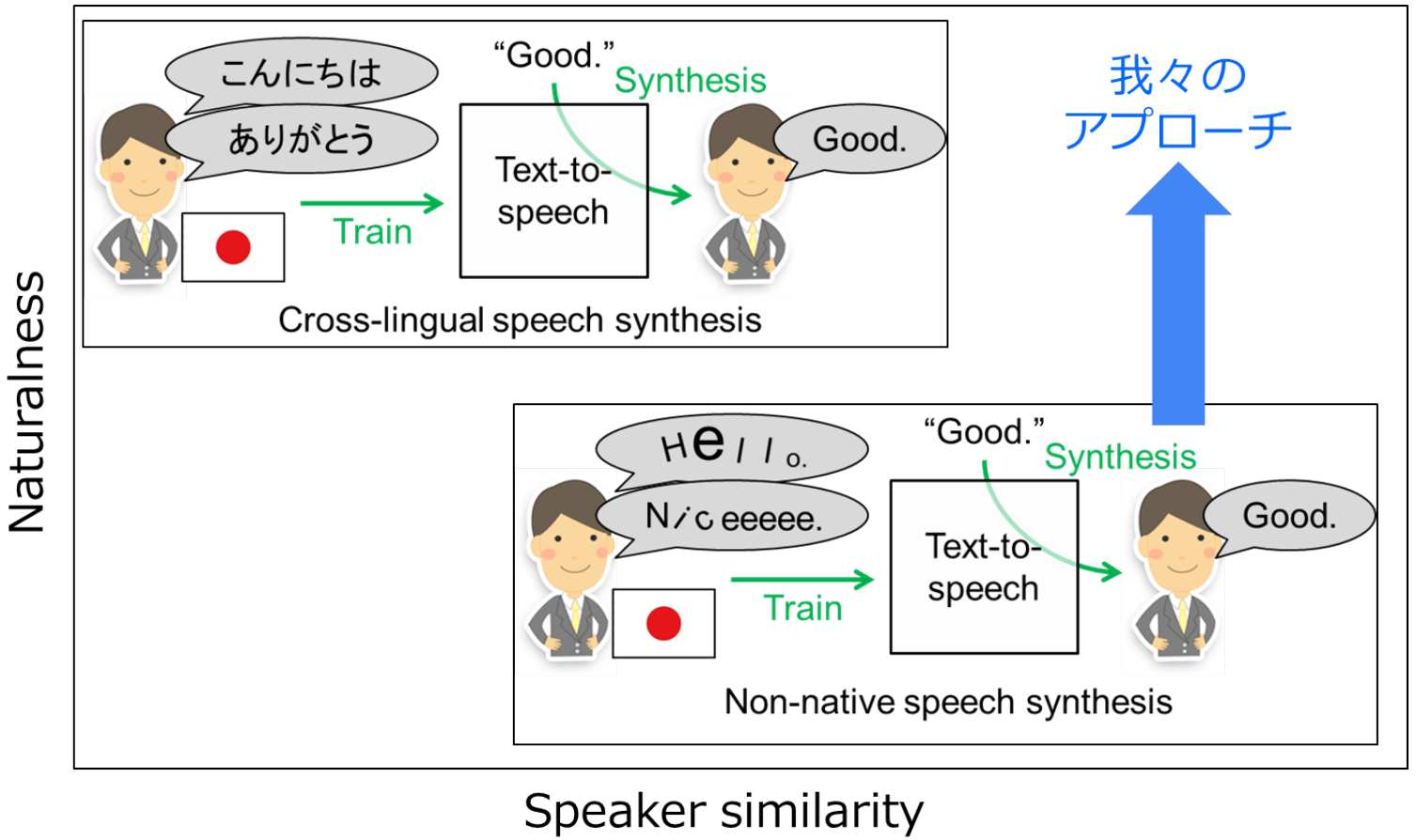


- データ空間 or モデル空間において効率良く話者性を表現
  例：主成分分析 (データ、モデル [62])
  例：Tucker 分解 (モデル [63])
  例：因子分析 (データ [64])、auto-encoder(データ [65])
- 言語と音声の因子推定 [39]

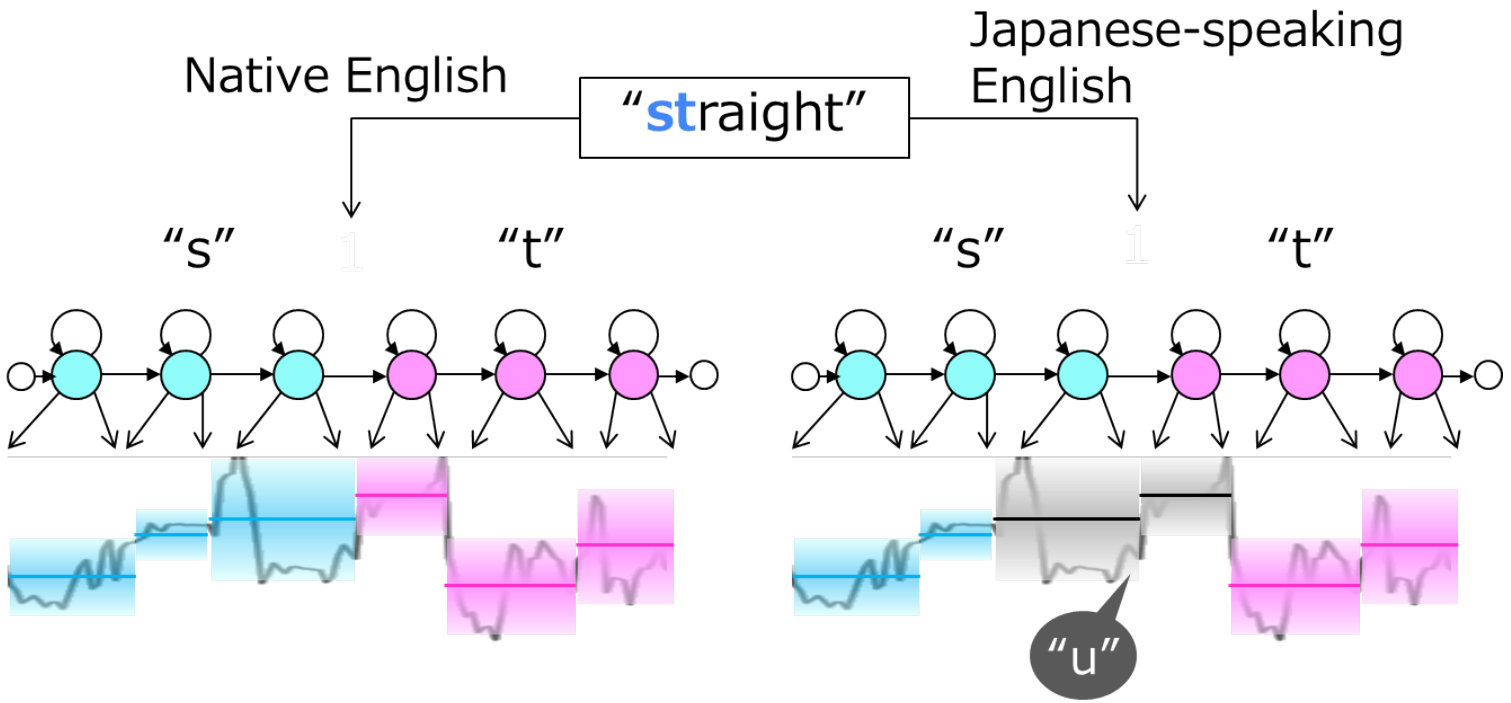
言語教育のための音声合成技術

- CALL システムにおける収録音声を合成音声へ
- 韻律の可視化
- 学習者に依存した教師音声の制御
  - Attractiveness
  - Familiarity
- 学習者の声色で教師音声を合成
  - 日本語母語話者の声で、見本となる英語音声を合成 [66, 67]

クロスリンガル音声合成 vs. ノンネイティブ音声合成

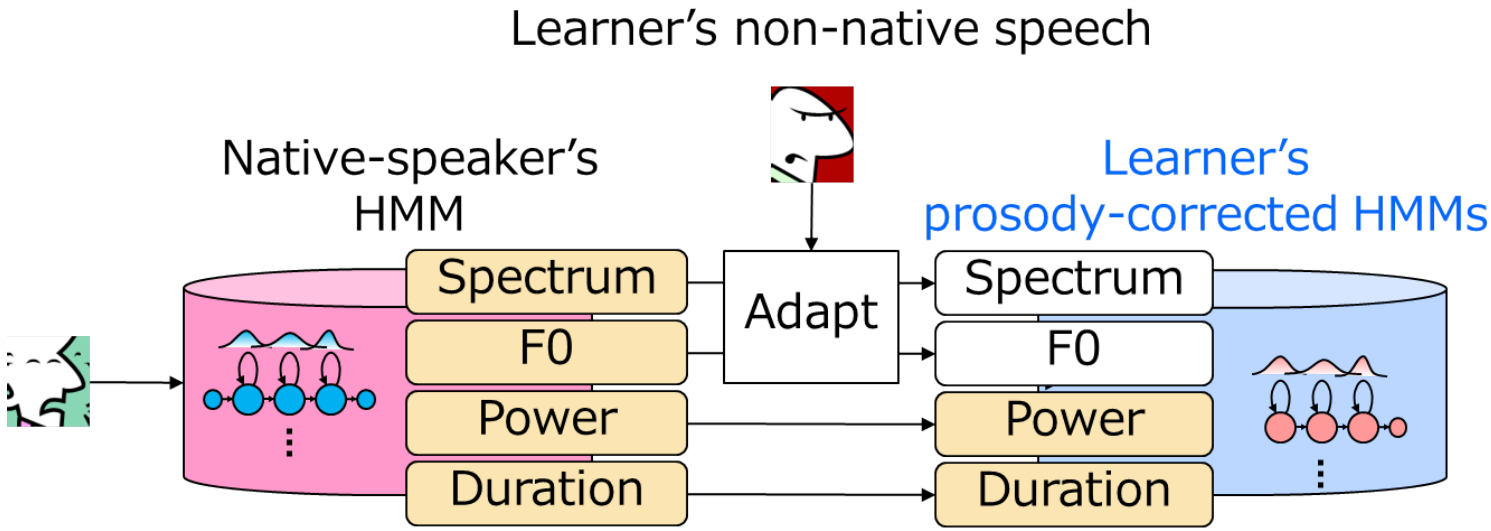


日本人英語音声合成の品質が劣化する理由



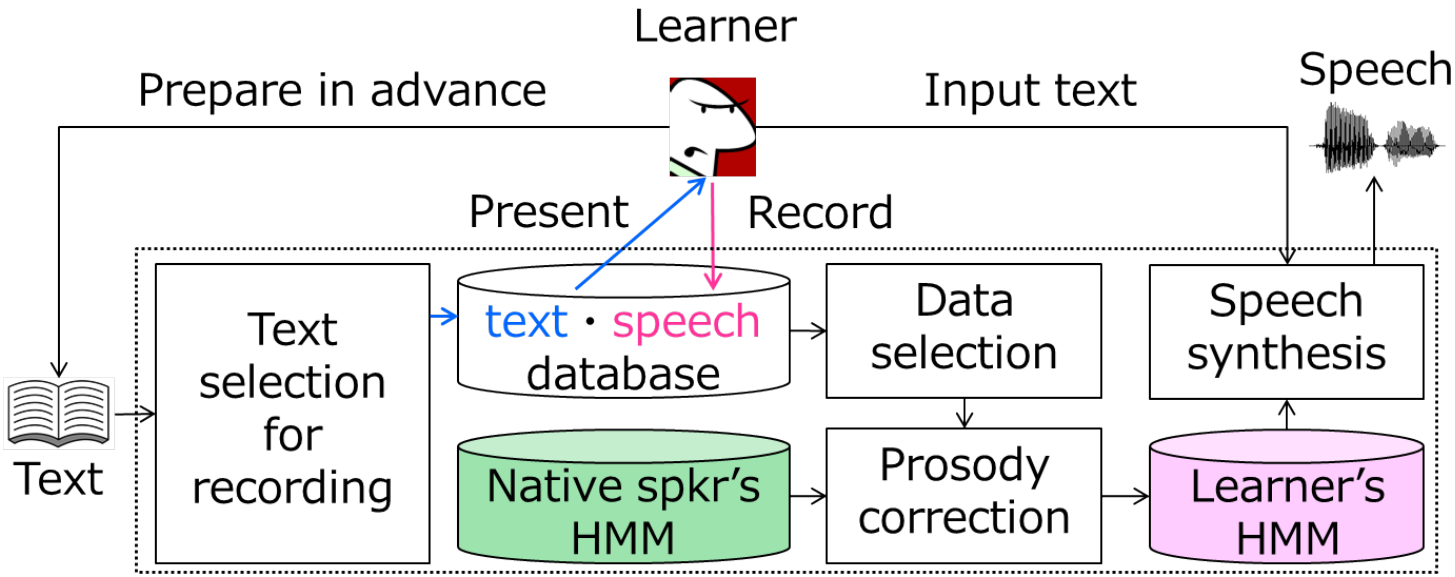
- 音韻誤り…テキストと音声が対応しない
  - 音素挿入の場合，音素の混じった音を学習してしまう
- 韻律誤り…ストレス・継続長
  - 等時性が異なる場合，継続長予測が困難

英語学習者の合成音声の韻律を補正 [66]



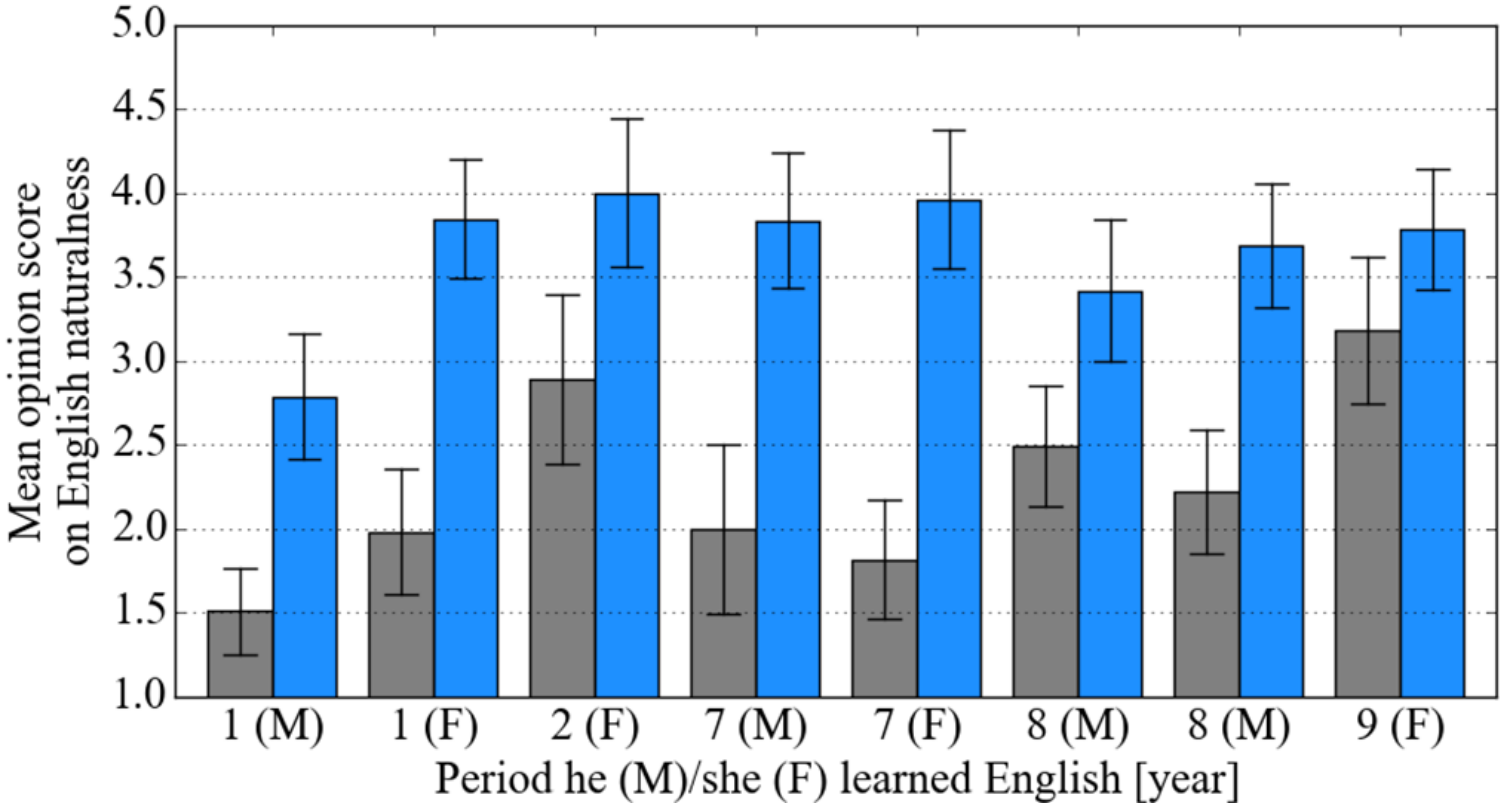
- 英語母語話者のパワーと継続長を利用
  → 日本語母語話者の個人性を反映しつつ，自然性を改善
- 英語母語話者の HMM をベースにした学習
  → **HMM** クラスタリングにおける音韻誤りの影響を緩和

英語学習者に向けたシステム構築 [67]



- 学習者の発話しやすい文章から，合成器学習に適した文を選択・収録
  - 例：学校の英語授業のテキストから選択
- 発話誤りの自動検出
  - 音響モデルの学習精度により自動判別
- 完全自動化された合成器構築・音声合成

評価結果 [67]



- 中学 1 年生を学習者と想定して，英語合成音声を補正できるかどうかを評価
- 学習年数にほとんど依存せず，補正効果がある

Section 5

参考文献

[1] Y. Sagisaka,
"Speech synthesis by rule using an optimal selection of non-uniform synthesis units,"
in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[2] H. Zen, K. Tokuda, and A. Black,
"Statistical parametric speech synthesis,"
Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] H. Kawahara, Jo Estill, and O. Fujimura,
"Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,"
in *MAVEBA 2001*, Firentze, Italy, Sep. 2001, pp. 1–6.

[4] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne,
"Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,"
Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[5] 戸田智基,
"統計的手法による音声変換,"
東京大学音声音響信号処理, 2013.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,
"Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis,"
IEICE Transactions on Information and Systems, vol. J87-D-II, no. 8, pp. 1563–1571, 2004.

[7] H. Zen, A. Senior, and M. Schuster,
"Statistical parametric speech synthesis using deep neural networks,"
in *Proc. ICASSP*. Vancouver, Canada, May 2013.

[8] M. Morise,
"CheapTrick, a spectral envelope estimator for high quality speech synthesis,"
Speech Communication, vol. 67, pp. 1–7, 2015.

[9] M. Morise, H. Kawahara, and H. Katayose,
"Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,"
in *Proc. AES 35th International Conference*, London, United Kingdom, Feb. 2009.

[10] M. Morise,

"An attempt to develop a singing synthesizer by collaborative creation,"
in *Proc. SMAC*. Stockholm, Aug. 2013.

[11] S. Takaki and J. Yamagishi,
"A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis,"
in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5535–5539.

[12] P. K. Muthukumar and A. W. Black,
"A deep learning approach to data-driven parameterizations for statistical parametric speech synthesis,"
vol. abs/1409.8558, 2014.

[13] H. Zen, K. Tokuda, and T. Kitamura,
"Refomulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,"
Computer Speech and Language, vol. 21, no. 1, pp. 153–173, Jan. 2007.

[14] Y. Nankaku, K. Nakamura, H. Zen, and K. Tokuda,
"Acoustic modeling with contextual additive structure for HMM-based speech recognition,"
in *Proc. ICASSP*, Las Vegas, U. S. A., Apr. 2008, pp. 4469–4472.

[15] Y. Wu and F. Soong,
"Modeling pitch trajectory by hierarchical HMM with minimum generation error training,"
in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4017–4020.

[16] S. Fan, Y. Qian, and F. Soong,
"TTS synthesis with bidirectional LSTM based recurrent neural networks,"
in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 1964–1968.

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,
"Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,"
in *Proc. EUROSPEECH*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

[18] T. Toda, A. W. Black, and K. Tokuda,
"Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,"
IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[19] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura,
"Postfilters to modify the modulation spectrum for statistical parametric speech synthesis,"
IEEE Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 755–767, 2016.

[20] 齋藤佑樹, 高道慎之介, and 猿渡洋,
"DNN 音声合成のための anti-spoofing を考慮した学習アルゴリズム,"
in 日本音響学会 *2016* 年秋季研究発表会講演論文集, Sep. 2016, vol. 3-5-1.

[21] K. Tokuda and H. Zen,
"Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis,"
in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4215–4219.

[22] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda,
"HMM-based singing voice synthesis and its application to Japanese and English,"
in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 265–269.

[23] "HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/,"
.

[24] 高道慎之介, 戸田智基, 森勢将雅, and 中村哲,
"HMM 音声合成における音声分析合成器 STRAIGHT と WORLD の比較,"
in 日本音響学会 *2015* 年秋季研究発表会講演論文集, Sep. 2015, vol. 1-Q-27.

[25] D. Erro, I. Sainz, E. Navas, and I. Hernaez,
"Harmonics plus noise model based vocoder for statistical parametric speech synthesis,"
IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 2, pp. 184–194, 2014.

[26] Y. Agiomyrgiannakis,
"VOCAINE the vocoder and applications in speech synthesis,"
in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4230–4234.

[27] H. Zen and H. Sak,
"Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,"
in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4470–4474.

[28] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura,
"Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion,"
in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4859–4863.

[29] Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda,
"Trajectory model training considering global variance for speech synthesis based on neural network,"
in *Proc. autumn meeting of ASJ 2015*, Fukushima, Japan, Sep. 2015 (In Japanese), pp. 237–238.

[30] H. Kameoka,
"Modeling speech parameter sequences with latent trajectory hidden Markov moddel,"
in *Proc. MLSP*, San Francisco, U.S.A., Sep. 2015.

[31] P. L. Tobing, T. Toda, H. Kameoka, and S. Nakamura,
"Acoustic-to-articulatory inversion mapping based on latent trajectory Gaussian mixture model,"
in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 953–957.

[32] T. Toda and K. Tokuda,
"Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM,"
in *Proc. ICASSP*, Las Vegas, U. S. A., Apr. 2008, pp. 3925–3928.

[33] M.-Q. Cai, Z.-H. Ling, and L.-R. Dai,
"Statistical parametric speech synthesis using a hidden trajectory model,"
Speech Communication, vol. 72, pp. 149â 159, 2015.

[34] M. Shannon, H. Zen, and W. Byrne,
"Autoregressive models for statistical parametric speech synthesis,"
IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 3, pp. 587–597, 2013.

[35] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai,
"Modeling F0 trajectories in hierarchically structured deep neural networks,"
Speech Communication, vol. 76, pp. 149â 159, 2016.

[36] Z. Wu and S. King,
"Investigating gated recurrent networks for speech synthesis,"
in *Proc. ICASSP*, Shanghai, China, Mar. 2016.

[37] T. Nakashika and Y. Minami,
"Generative acoustic-phonemic-speaker model based on three-way restricted Boltzmann machine,"
in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 1487–1491.

[38] W. Wang, S. Xu, and B. Xu,
"First step towards end-toend parametric TTS synthesis: Generating spectral parameters with neural attention,"
in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2243–2247.

[39] B. Li and H. Zen,
"Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis,"

in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2468–2472.

[40] X. Wang, S. Takaki, and J. Yamagishi,
"Investigating very deep highway networks for parametric speech synthesis,"
in *Proc. SSW9*, Sunnyvale, CA, USA, Sep. 2016, pp. 181–186.

[41] Q. Hu, J. Yamagishi, K. Richmond, K. Subramanian, and Y. Stylianou,
"Initial investigation of speech synthesis based on complex-valued neural networks,"
in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5630–5634.

[42] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu,
"WaveNet: A generative model for raw audio,"
vol. abs/1609.03499, 2016.

[43] T. Toda and K. Tokuda,
"A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,"
IEICE Transactions on Information and Systems, vol. E90-D, no. 5, pp. 816–824, 2007.

[44] G. Hinton,
"Product of experts,"
in *Proc. ICANN*, 1999, pp. 1–6.

[45] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura,
"Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis,"
in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4210–4214.

[46] T. Toda and S. Young,
"Trajectory training considering global variance for HMM-based speech synthesis,"
in *Proc. ICASSP*, Taipei, Taiwan, Aug. 2009, pp. 4025–4028.

[47] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura,
"Modulation spectrum-constrained trajectory training algorithm for HMM-based speech synthesis,"
in *Proc. ICASSP*, Dresden, Germany, Sep. 2015, pp. 1206–1210.

[48] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura,
"The NAIST text-to-speech system for the Blizzard Challenge 2015,"
in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[49] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda,

"The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016,"
in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 1667–1671.

[50] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md Sahidullah, and Aleksandr Sizov,
"ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,"
in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2037–2041.

[51] F.-L. Xie, F. K. Soong, and H. Li,
"A KL divergence and DNN approach to cross-lingual TTS,"
in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5515–5519.

[52] H. Liang, Y. Qian, F. K. Soong, and L. Gongshen,
"A cross-language state mapping approach to bilingual (Mandarin-English) TTS,"
in *Proc. ICASSP*, Las Vegas, U. S. A., Apr. 2008, pp. 4641–4644.

[53] Y. Qian, J. Xu, and F. K. Soong,
"A frame mapping based HMM approach to cross-lingual voice transformation,"
in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5120–5123.

[54] Q. T. Do, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura,
"Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs,"
in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3665–3669.

[55] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine,
"Emotional transplant in statistical speech synthesis based on emotion additive model,"
in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 274–278.

[56] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi,
"A style control technique for HMM-based expressive speech synthesis,"
IEICE Transactions on Information and Systems, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[57] P. K. Muthukumar and A. W. Black,
"Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis,"
in *Proc. ICASSP*, pp. 2594–2598. Florence, Italy, May 2014.

[58] K. Sawada, K. Hashimoto, K. Oura, and K. Tokuda,
"The NITECH HMM-based text-to-speech system for the Blizzard Challenge 20155,"
in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[59] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn,
"An attentional model for speech translation without transcription,"
in *Proc. NAACL-HLT*, San Diego, California, U.S.A., May 2016, pp. 949â 959,.

[60] S. Sitaram, A. Parlikar, G. K. Anumanchipalli, and A. W Black,
"Universal grapheme-based speech synthesis',"
in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3360–3364.

[61] Alexander Gutkin, Linne Ha, Martin Jansche, Knot Pipatsrisawat, and Richard Sproat,
"TTS for low resource languages: A Bangla synthesizer,"
in *Proc. LREC*, Paris, France, 2016, pp. 2005–2010.

[62] T. Toda, O. Ohtani, and K. Shikano,
"One-to-many and many-to-one voice conversion based on Eigenvoices,"
in *Proc. ICASSP*, Hawaii, U.S.A., Apr. 2007, pp. 1249–1252.

[63] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose,
"One-to-many voice conversion based on tensor representation of speaker space,"
in *Proc. INTERSPEECH*, Florence, Italy, Jul. 2011, pp. 653–657.

[64] K. Kazumi, Y. Nankaku, and K. Tokuda,
"Factor analyzed voice models for HMM-based speech synthesis,"
in *Proc. ICASSP*, Dallas, Texas, U.S.A., Apr. 2010, pp. 4234–4237.

[65] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki,
"Voice conversion in high-order Eigen space using deep belief nets,"
in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 369–372.

[66] Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura,
"Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic characteristics,"
IEICE Transactions on Information and Systems, vol. E99-D, no. 12, pp. xxxx–xxxx, 2016.

[67] 高道慎之介, 戸田智基, Graham Neubig, Sakriani Sakti, and 中村哲,
"HMM に基づく日本人英語音声合成における中学生徒の英語音声を用いた評価,"
in 日本音響学会 *2015* 年秋季研究発表会講演論文集, Sep. 2015, vol. 2-5-8.