東京大学 信号処理論特論第7回 (2018/06/12)

音声合成・変換 その2

猿渡 洋・高道 慎之介

06/12/2018©Shinnosuke Takamichi,
The University of Tokyo

講義予定

04/10: 第1回 統計的音声音響信号処理概論
05/01: 第2回 非負値行列因子分解
05/08: 第3回 ブラインド音源分離その１
05/15: 第4回 ブラインド音源分離その２
05/22: 第5回 エンハンスメント・高次統計量解析とその応用
05/29: 第6回 【レポート課題１】
06/05: 第7回 音声合成・変換その１
06/12: 第8回 音声合成・変換その２
06/19: 第9回 音場再現の基礎
06/26: 第10回 学外講師・未定
07/03: 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)

本講義の目的

音声合成・変換の近年の発展は？

復習

テキスト音声合成・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)

- 音声を異なる音声に変換
- ヒトの発声制約をこえたコミュニケーションのため

音声の持つ情報

狭義の音声認識
(speech-to-text)

言語情報

テキスト化できる情報

感情認識など
(emotion recognition)

パラ言語情報

話し手が意図的に付与する,
テキスト化できない情報 (例：感情)

話者認識など
(speaker recognition)

非言語情報

話し手の意図とは無関係に付与される,
テキスト化できない情報（例：話者性）

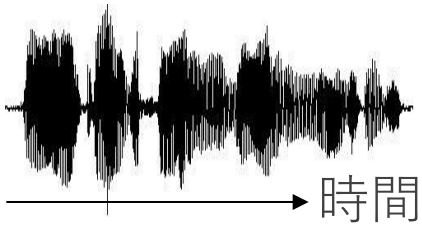
音声の生成過程：ソース・フィルタモデル

音色の付与

口や舌を動かして，
音色をつける！

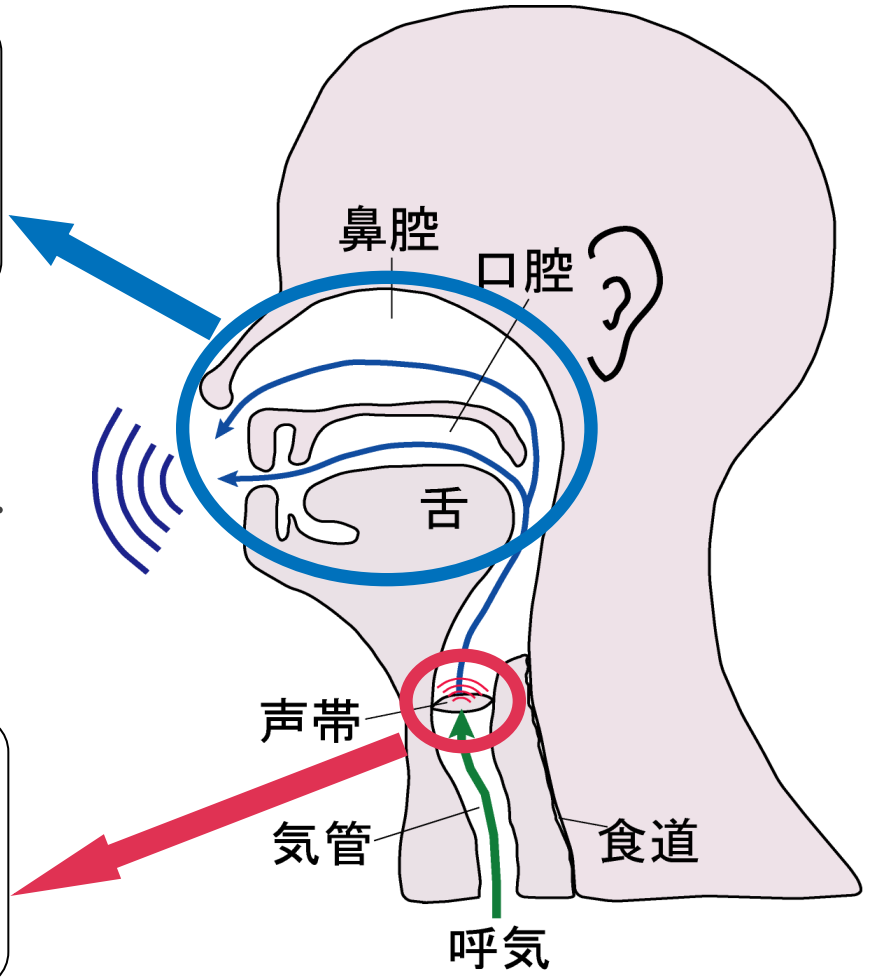
畳み込むと…

声になる！



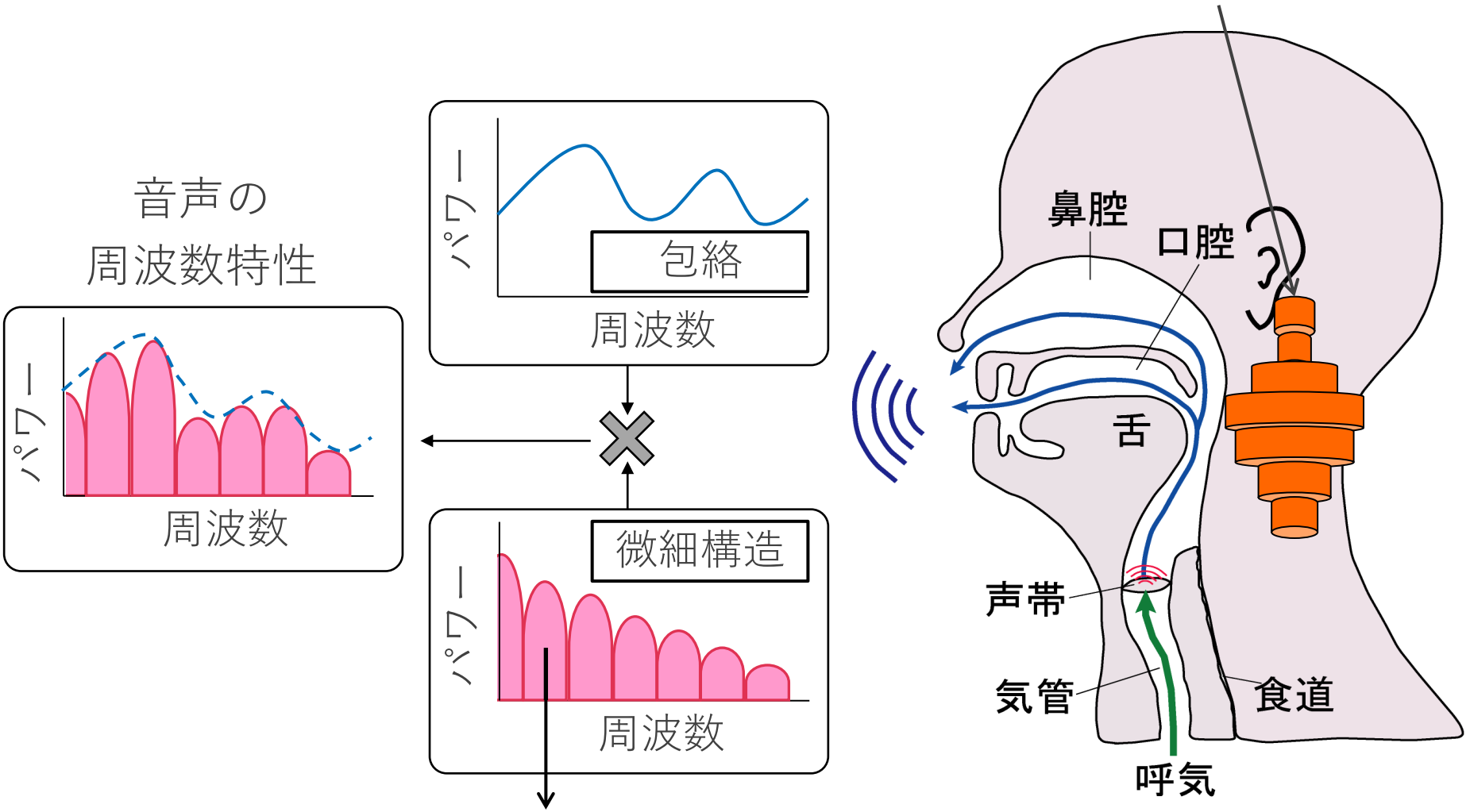
音高の生成

声帯を開閉させて，
空気を振動させる！



音声のスペクトル構造
（音声のスペクトル構造の2要素）

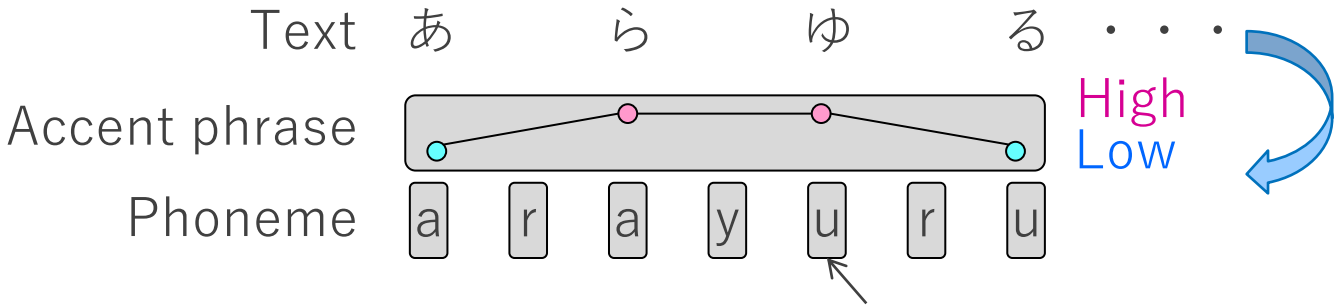
音響管連接でモデル化可能



言語特徴量と音声特徴量

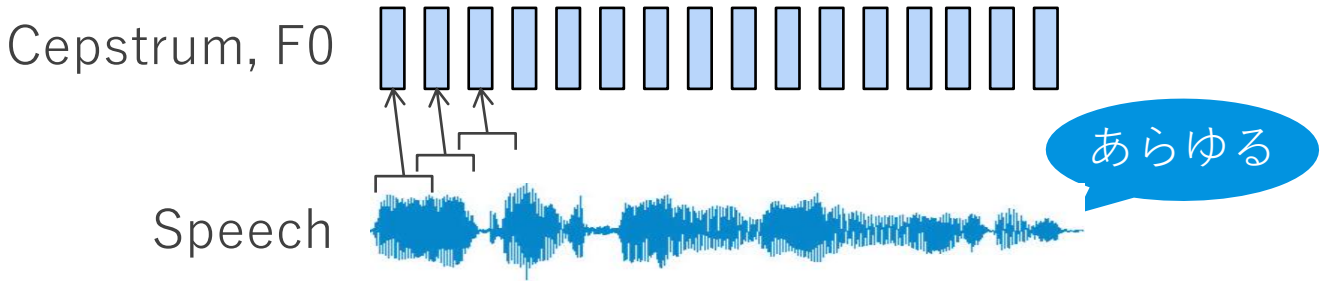
言語特徴量

– テキストから，音素・音節・アクセントなどの特徴量を抽出



音声特徴量

– 音声から，声道・声帯の特徴量を抽出



コーパスベース音声合成の種類

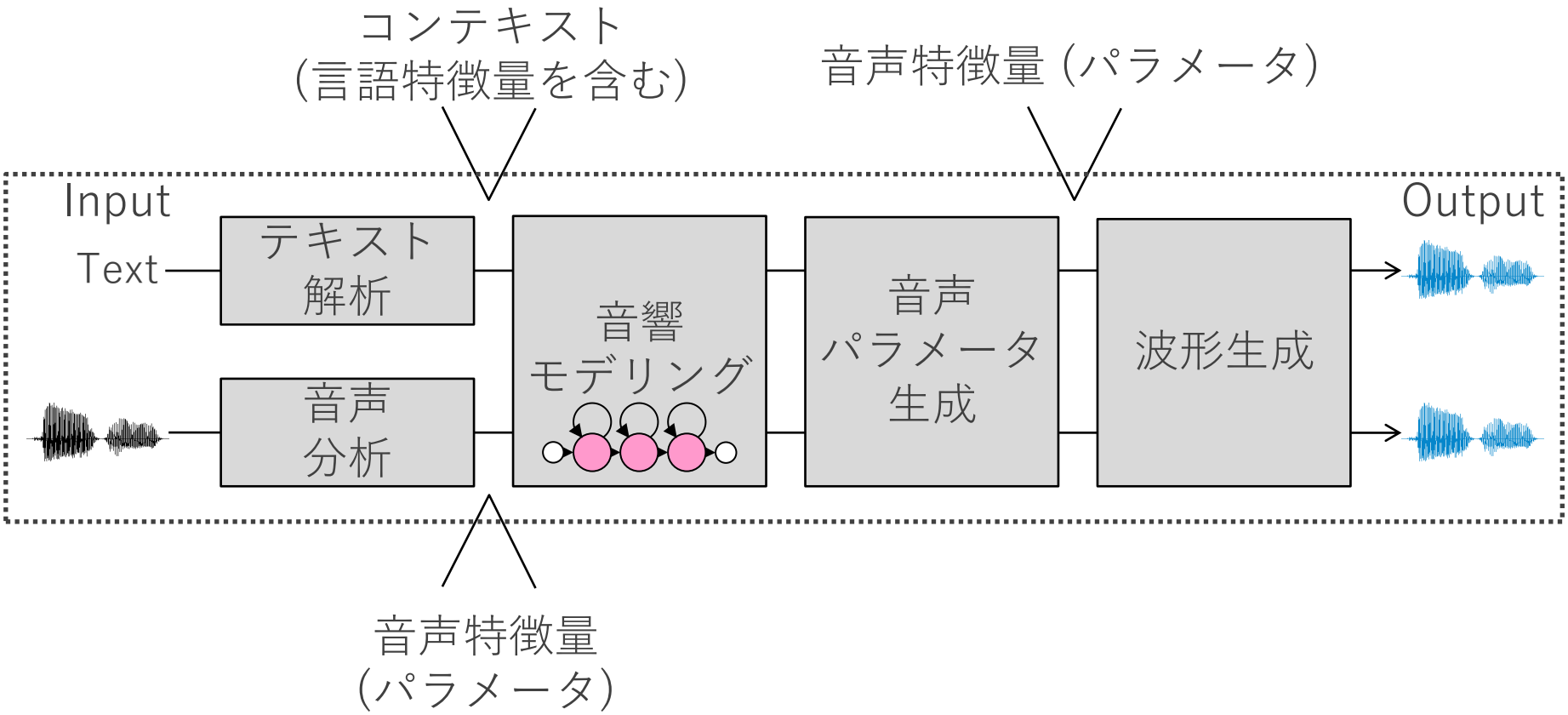
素片選択型合成 (unit selection synthesis)

- 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 (statistical speech synthesis)

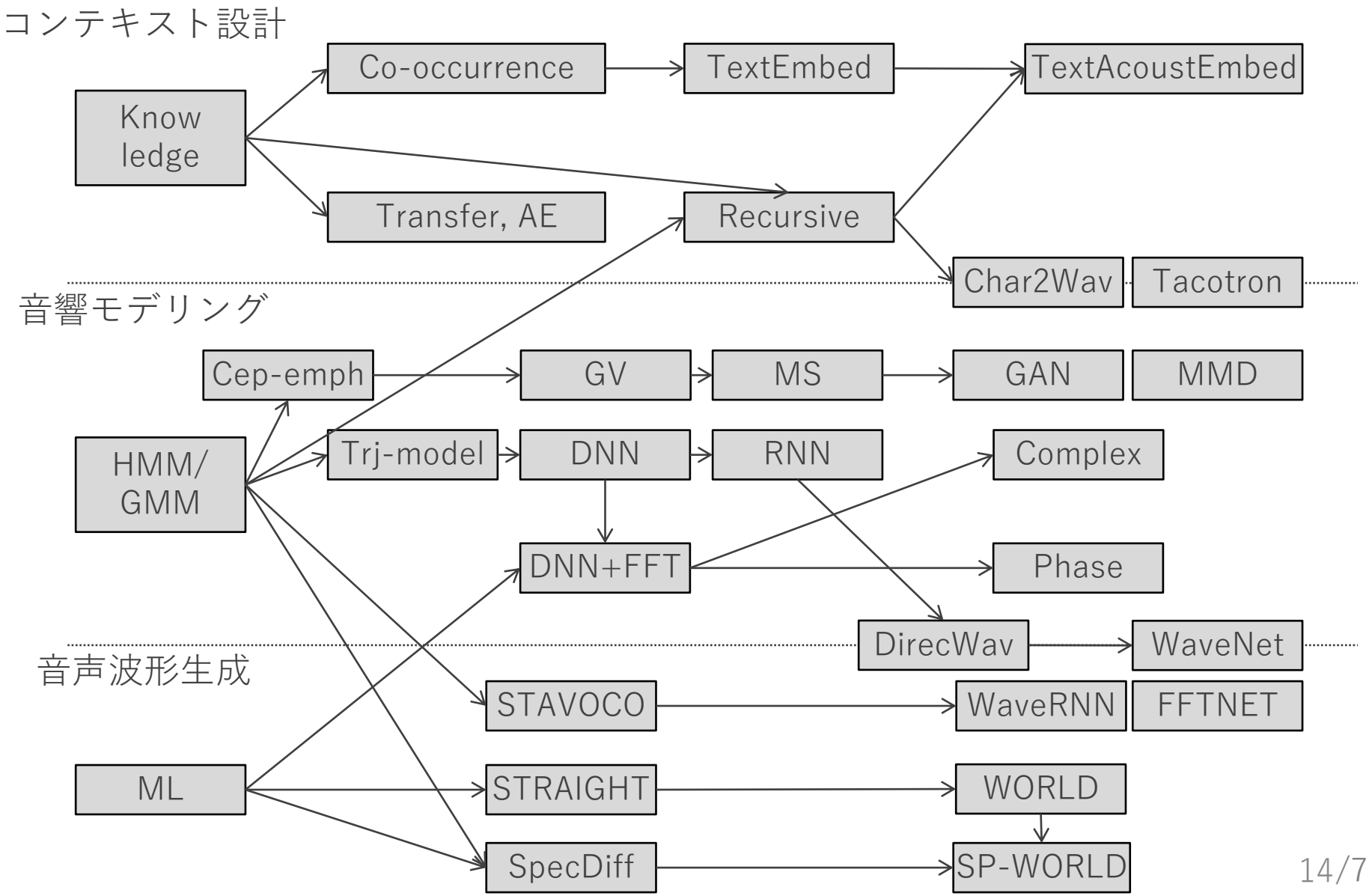
- 音声波形・パラメータを統計モデルでモデル化
- 長所: 声質を制御しやすい、フットプリントが小さい，機械学習の知見を大いに使える
- 短所: 低い音質 (最近は非常に改善されてきた)

統計ベース方式の手順

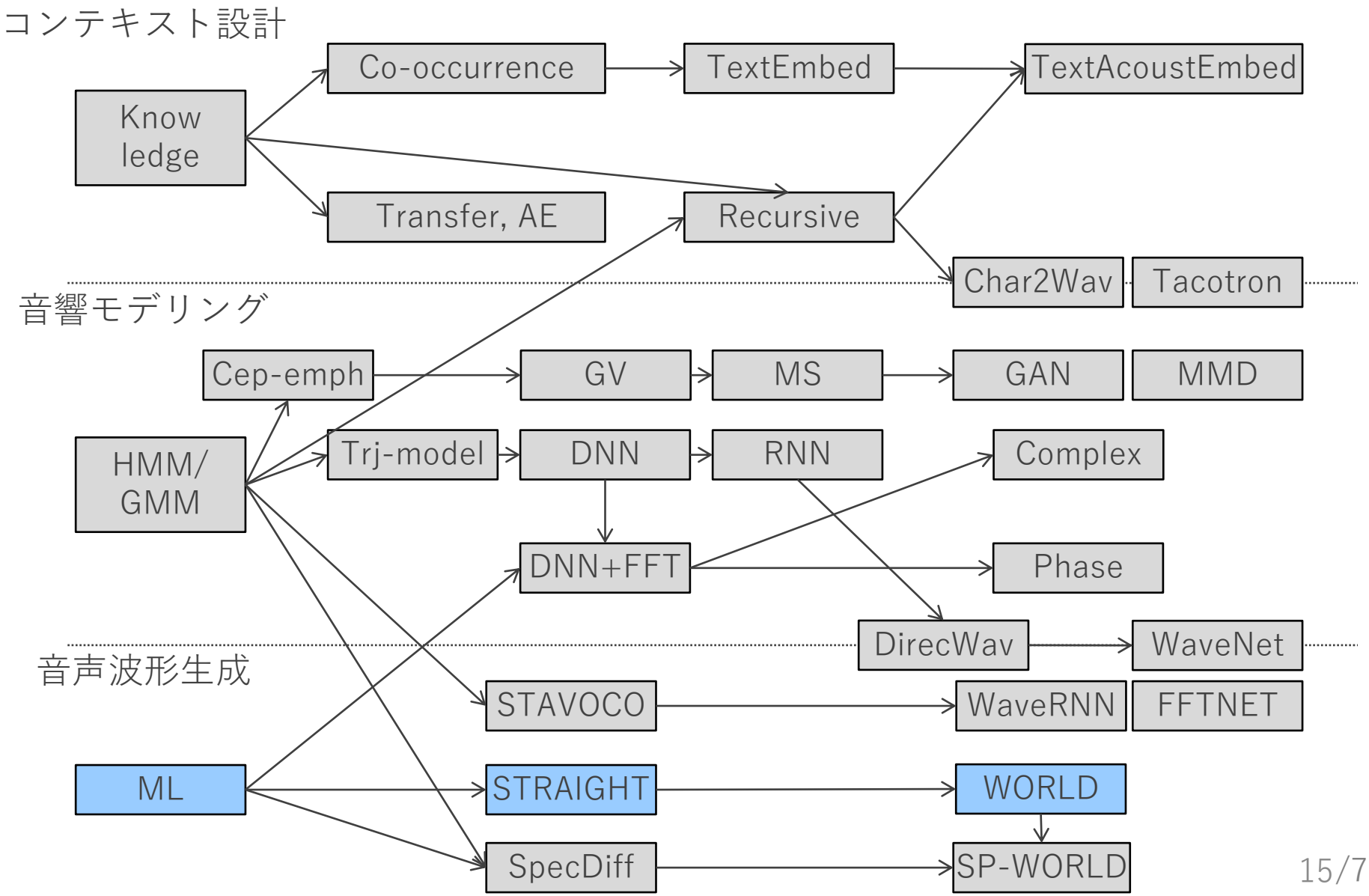

コンテキスト
(言語特徴量を含む)
音声特徴量 (パラメータ)
Input
Text
テキスト
解析
音声
分析
音響
モデリング
音声
パラメータ
生成
波形生成
Output
音声特徴量
(パラメータ)

近年の話題
~DNN音声合成を中心にして~

音声合成変換技術の変遷



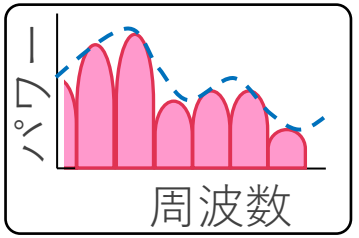
音声合成変換技術の変遷



信号処理ボコーダ

波形生成（ボコーダ）の役割は？

– 音源・声道特性を如何に高精度に抽出・制御し，
そこから自然音声と遜色ない波形を再合成できるか



STRAIGHT [Kawahara99]

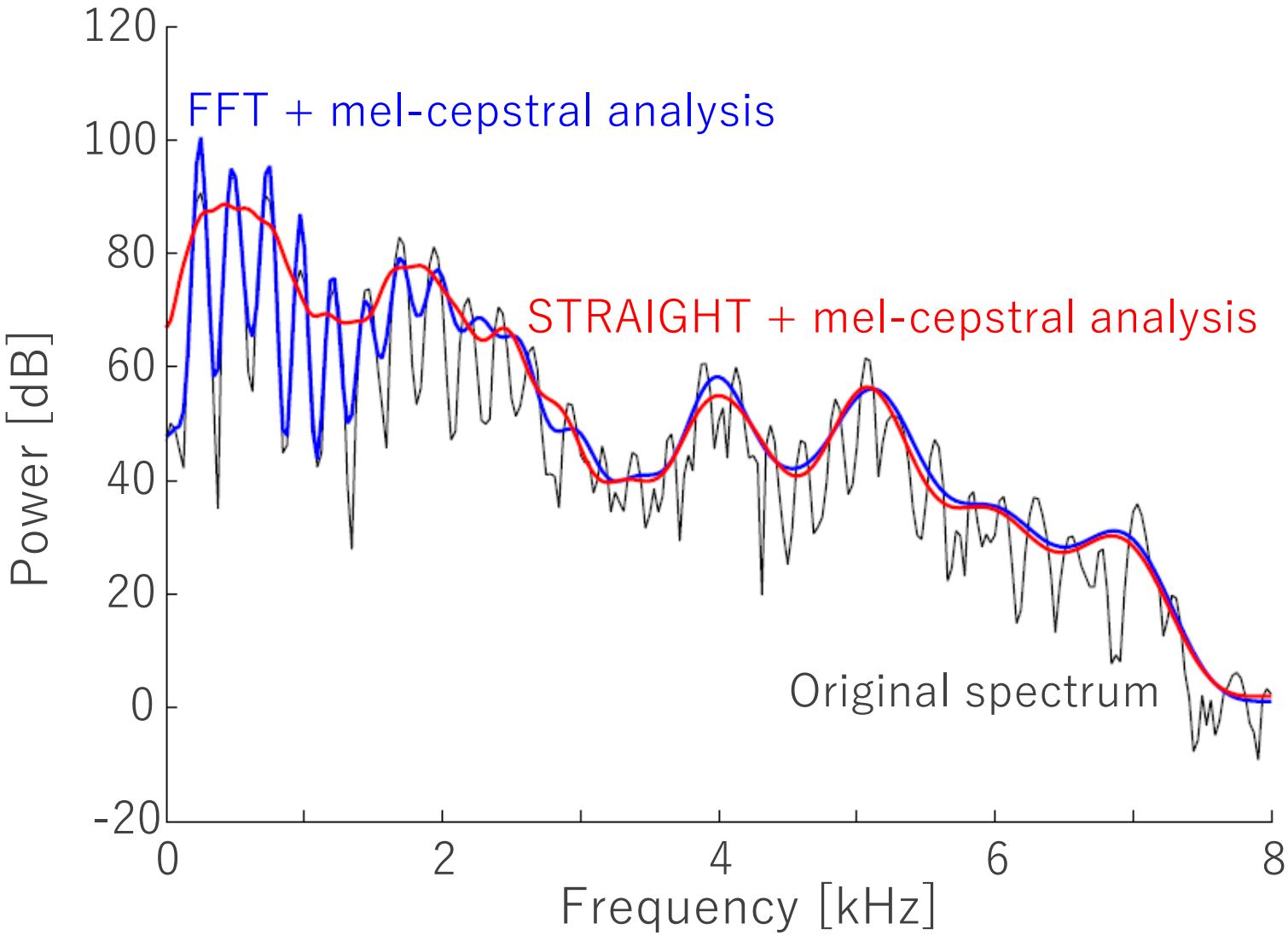
– 音源信号に非周期性指標を導入 [Kawahara01]
– F0-adaptiveな窓関数により，F0の影響を除去 [KawaharaHP]
– HMM音声合成／GMM音声変換の隆盛において重要な役割

WORLD [Morise16]

– STRAIGHTの符号化を継承し，さらに高品質化
– 分析時間位置に依存しない窓関数の設計 [Morise15]
– 修正BSDライセンスでソースが公開され，音声合成の産業応用に

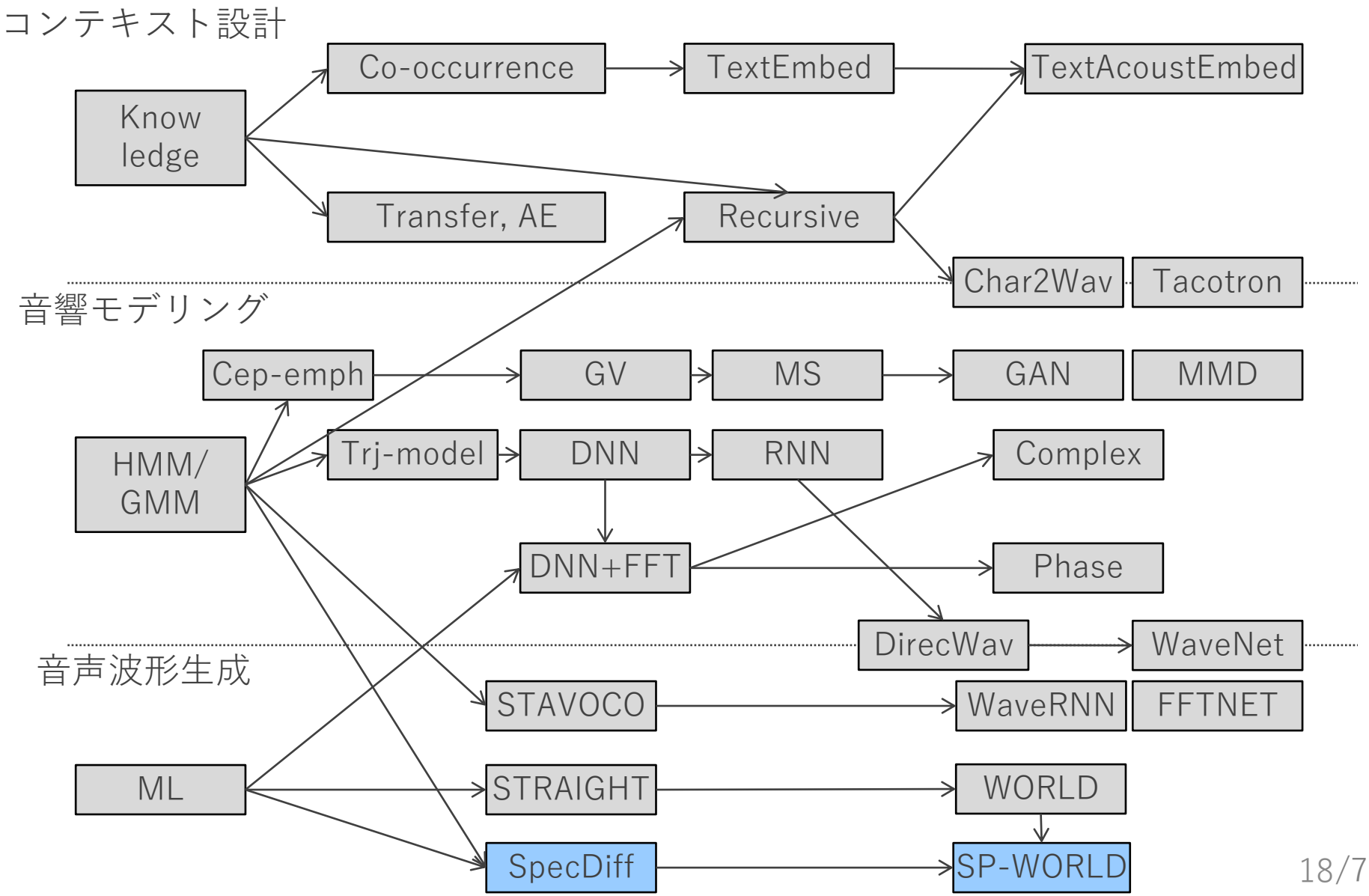
[Kawahara99] Kawahara et al., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.
[Kawahara01] Kawahara et al., “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 2001.
[KawaharaHP] http://www.wakayama-u.ac.jp/~kawahara/HowTANDEMSTRAIGHTworks/
[Morise16] Morise et al., ” WORLD: a vocoder-based high-quality speech synthesis system for real-time applications”, IEICE transactions, 2016.
[Morise15] Morise, “CheapTrick, a spectral envelope estimator for high quality speech synthesis,” Speech Communication, 2015.

STRAIGHTによるスペクトル包絡抽出の例

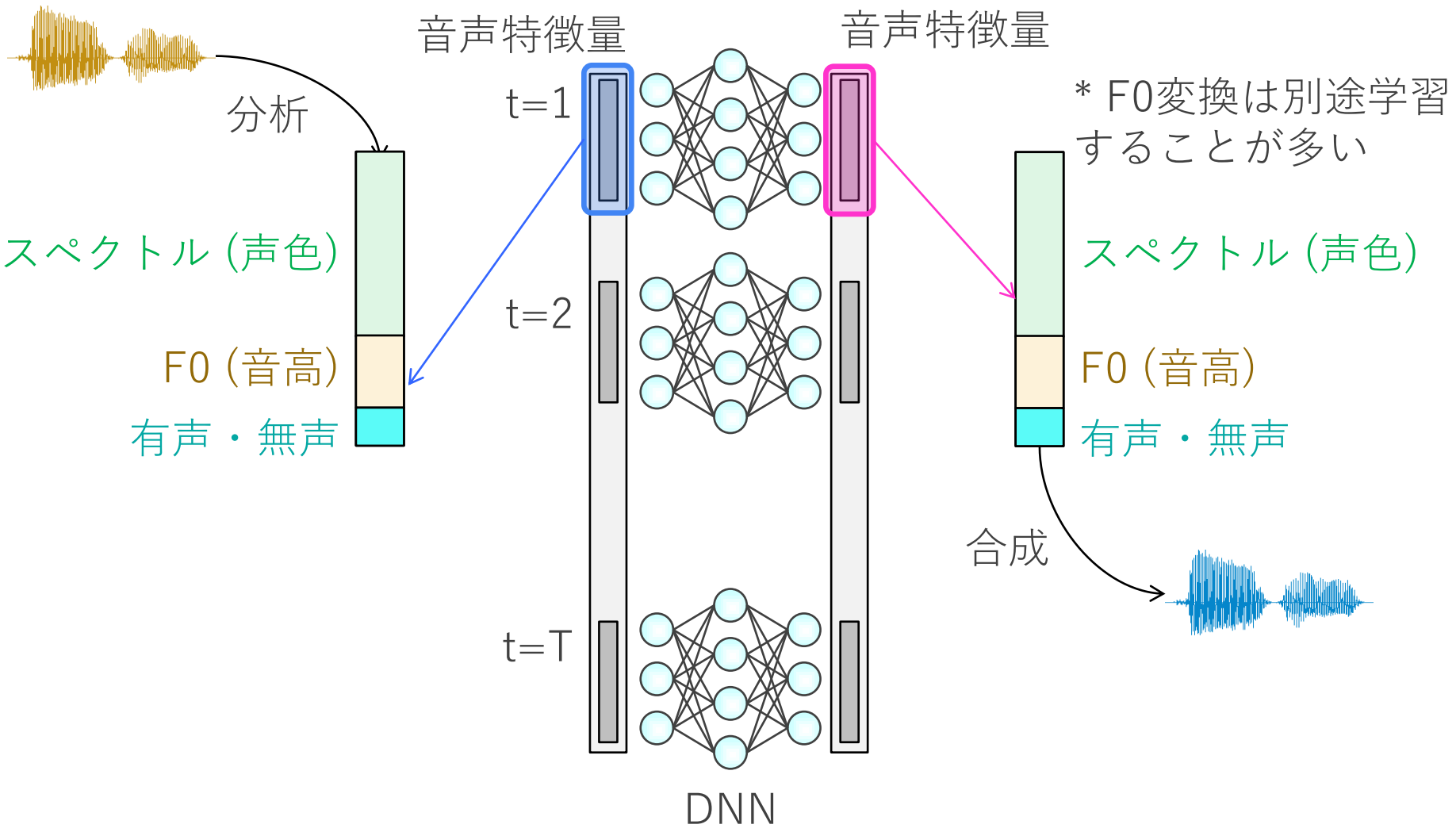


[HTS] HTS 2.3 slide http://hts.sp.nitech.ac.jp/?Download

音声合成変換技術の変遷



通常のDNN音声変換


音声特徴量
音声特徴量
分析
t=1
t=2
t=T
スペクトル (声色)
F0 (音高)
有声・無声
スペクトル (声色)
F0 (音高)
有声・無声
* F0変換は別途学習することが多い
合成
DNN

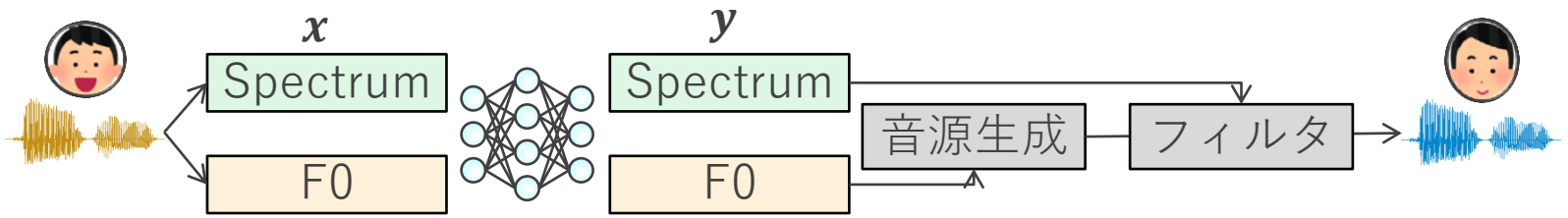
差分スペクトル法による音声変換

F0を変換しないケース：同性の話者変換，歌声変換，音韻変換

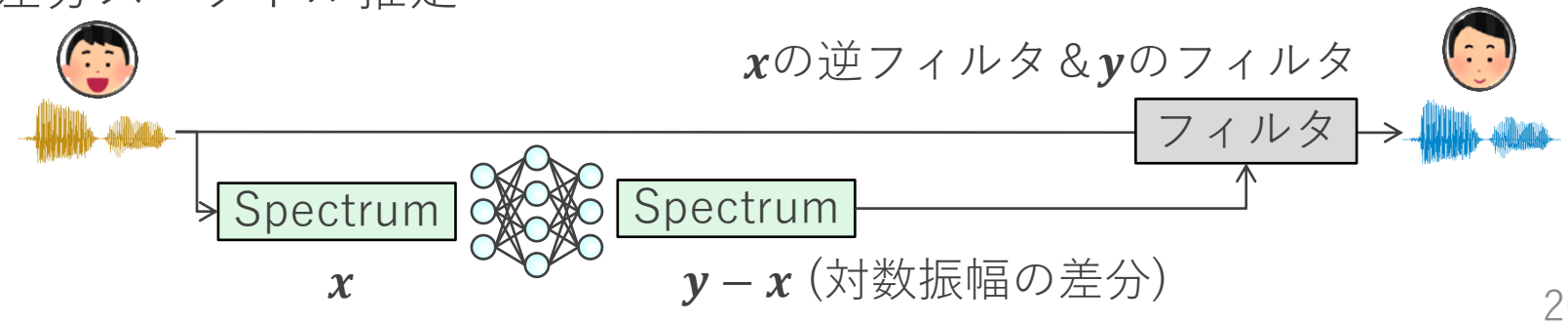
差分スペクトル法

- F0分析なしで，スペクトル包絡の差分を波形にフィルタリング
- F0分析とボコーダによるエラーを回避可能

通常の音声変換



差分スペクトル推定



[Kobayashi18] K. Kobayashi et al., “Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential,” Speech communication, 2018.

差分スペクトル法の発展

スペクトル差分を推定する統計モデル

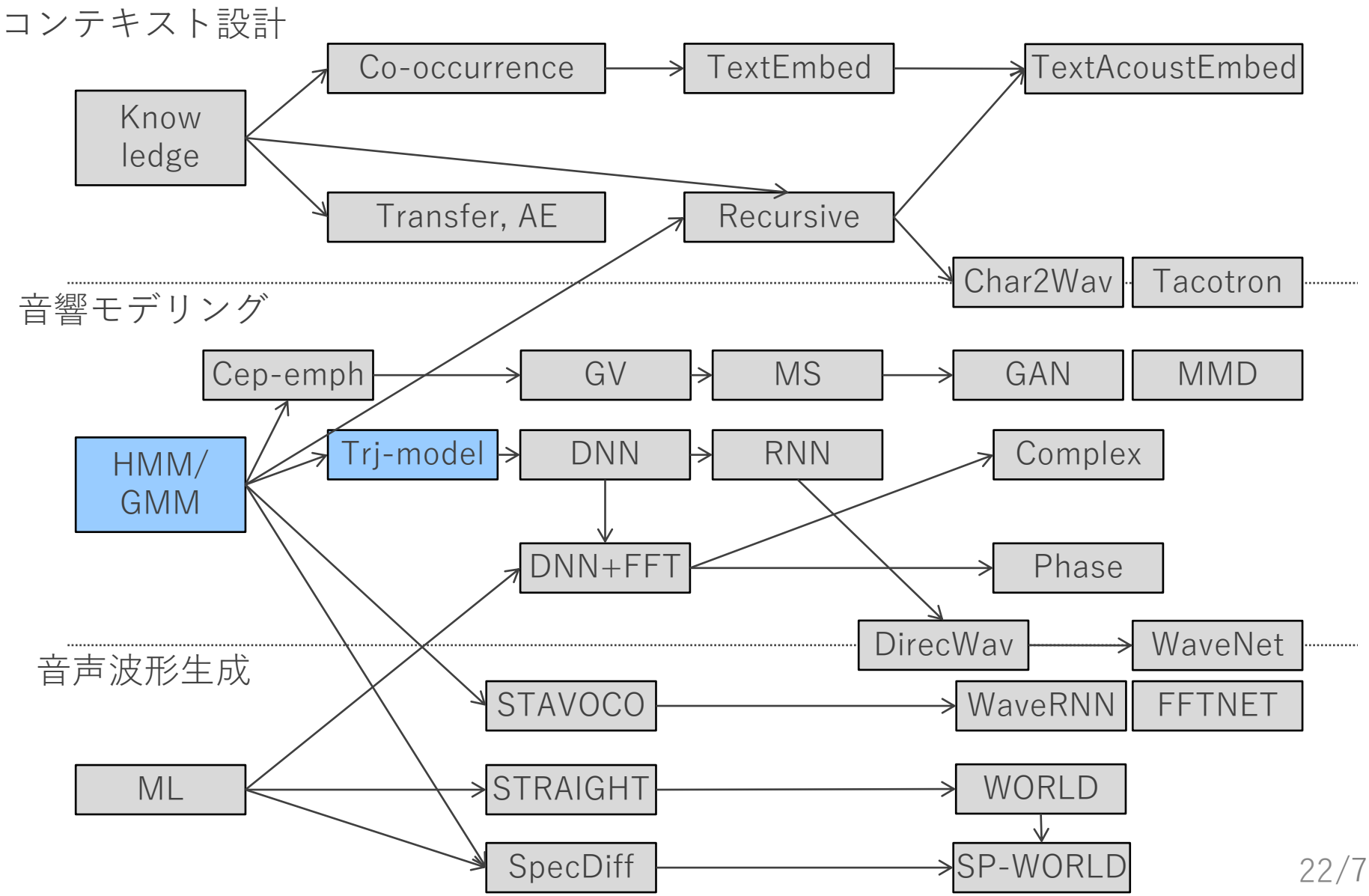
- SpecDiff GMM [Kobayashi18] … 通常のGMMから解析的に導出
- WeightedDiff DNN [Saito17] … 特徴量次元毎に差分重みを推定

フィルタリング法

- MLSAフィルタベース [Kobayashi18]
- WORLDボコーダベース (SP-WORLD) [須田18] … MLSAより高品質

[Kobayashi18] K. Kobayashi et al., “Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential,” Speech communication, 2018.
[Saito17] S. Takamichi et al., “Voice Conversion Using Input-to-Output Highway Networks,” IEICE Transactions, 2017.
[須田18] 須田 他, “高品質声質変換のための特徴量分析再訪,” 日本音響学会2018年春季研究発表会講演論文集, 2018.

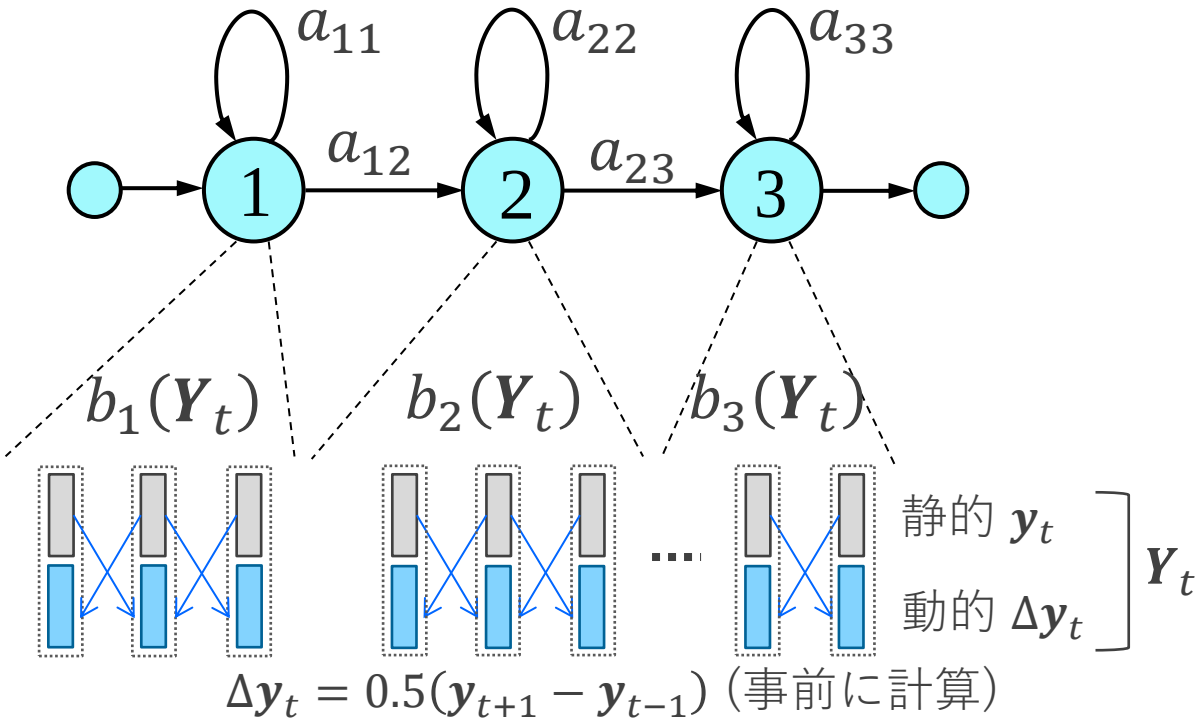
音声合成変換技術の変遷



HMM音声合成の学習部 (復習)

– 別途計算した動的特徴量も学習に利用
– HMM状態内は定常，フレーム間は独立を過程

$$\hat{\boldsymbol{\lambda}} = \text{argmax}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \Sigma_{all\ \boldsymbol{q}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{X})$$

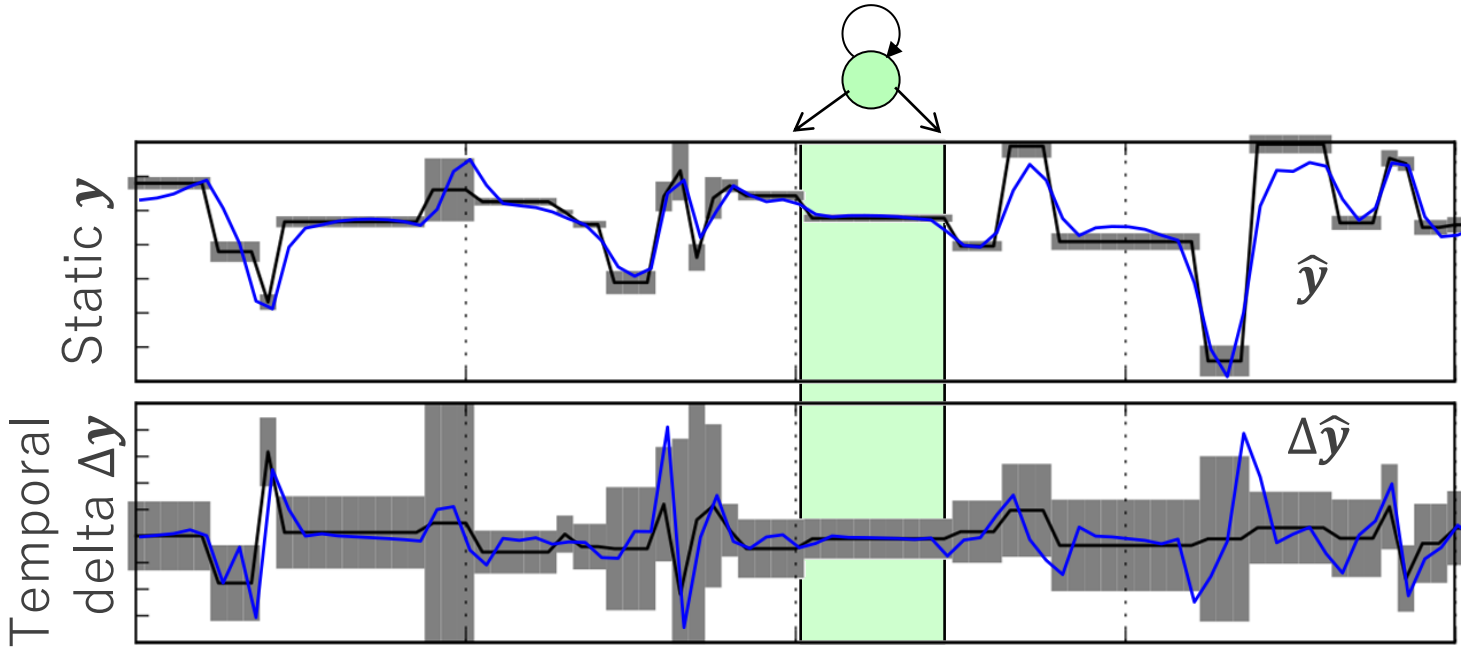


$\Delta\boldsymbol{y}_t = 0.5(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1})$ (事前に計算)

パラメータ生成部 (復習)

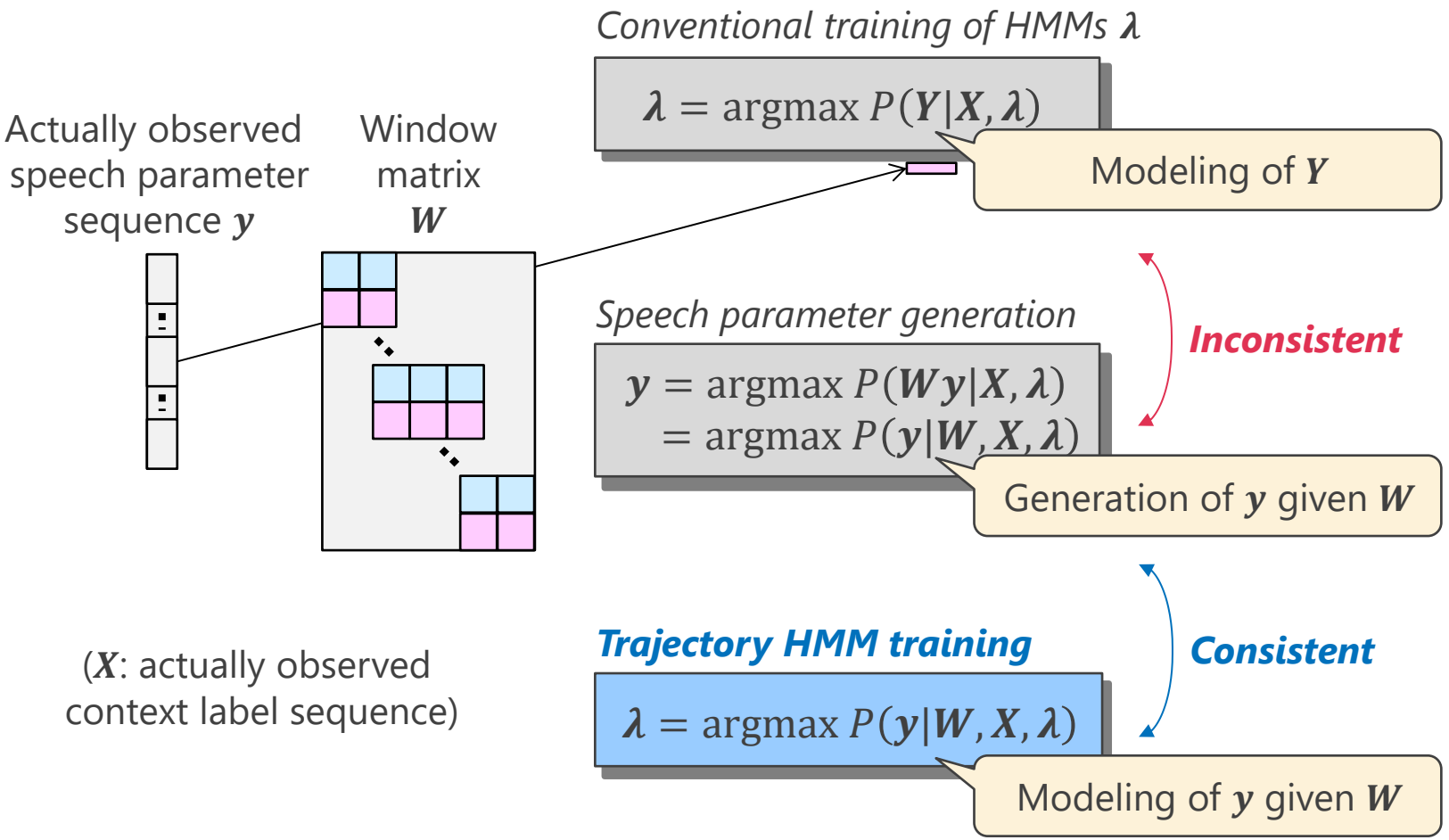
– 最尤状態系列 $\hat{\boldsymbol{q}}$ で近似．動的特徴量を計算する行列 $\boldsymbol{W}$ の制約下
– 音声パラメータの確率分布は正規分布で得られる

$$\hat{\boldsymbol{y}} = \operatorname{argmax} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) = \operatorname{argmax} P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) = \left(\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}$$



[Tokuda00] Tokuda et al., “Speech parameter generation algorithms fro HMM-based speech synthesis”, Proc. ICASSP, 2000.
[Toda07] Toda et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Transactions, 2007.

学習部と生成部の矛盾

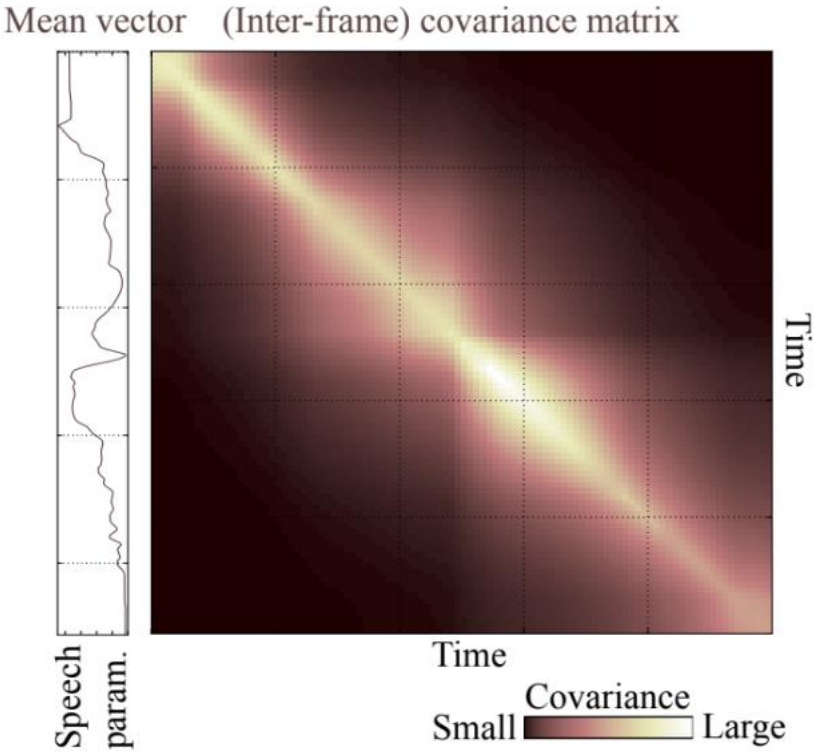

Conventional training of HMMs $\boldsymbol{\lambda}$
$\boldsymbol{\lambda} = \text{argmax}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$
Modeling of $\boldsymbol{Y}$
Actually observed speech parameter sequence $\boldsymbol{y}$
Window matrix $\boldsymbol{W}$
Speech parameter generation
$\boldsymbol{y} = \text{argmax}\, P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\lambda})$
$= \text{argmax}\, P(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\lambda})$
Inconsistent
Generation of $\boldsymbol{y}$ given $\boldsymbol{W}$
($\boldsymbol{X}$: actually observed context label sequence)
Trajectory HMM training
$\boldsymbol{\lambda} = \text{argmax}\, P(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\lambda})$
Consistent
Modeling of $\boldsymbol{y}$ given $\boldsymbol{W}$

トラジェクトリモデル

– 単一状態系列 $\hat{\boldsymbol{q}}$ で近似すると，確率密度関数は正規分布に．
– 平均は最尤パラメータ生成のものと等価，共分散はフレーム間相関

$$\hat{\boldsymbol{\lambda}} = \operatorname{argmax} P(\boldsymbol{y}|\boldsymbol{W}, \hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}) = N(\boldsymbol{y}|\hat{\boldsymbol{y}}, \boldsymbol{\Sigma})$$

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



関連研究

トラジェクトリモデルの発展

- Trajectory GMM [zen09, Takamichi15], DNN [Hashimoto16]
- Latent trajectory HMM [kameoka15]/GMM [Tobing16]
- Factor-analyzed trajectory HMM [Cai15]

AR (auto-regressive) 過程の考慮

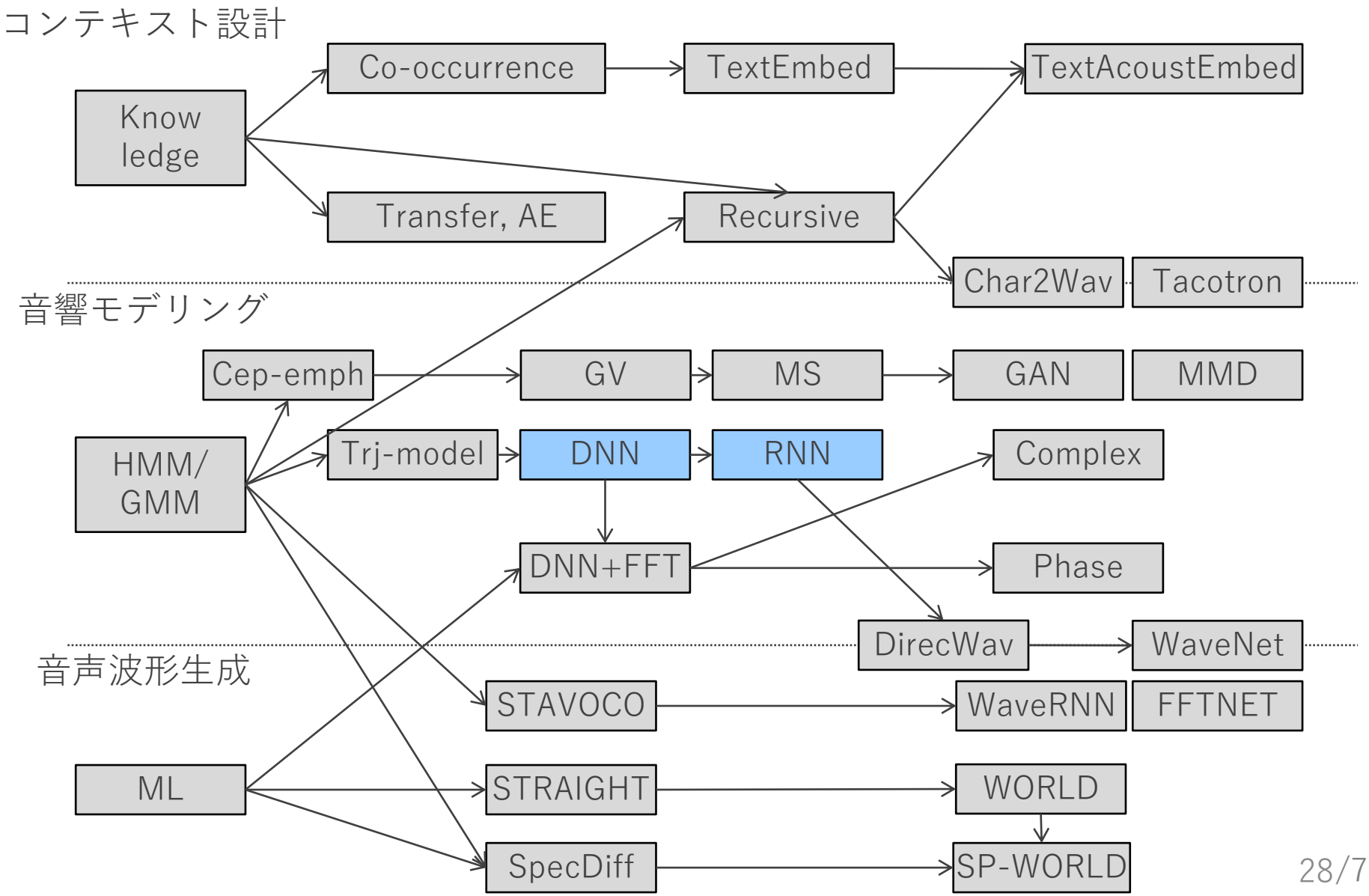
- AR-HMM[Shannon13]/DNN[Wang17]

MGE (minimum generation error) 学習

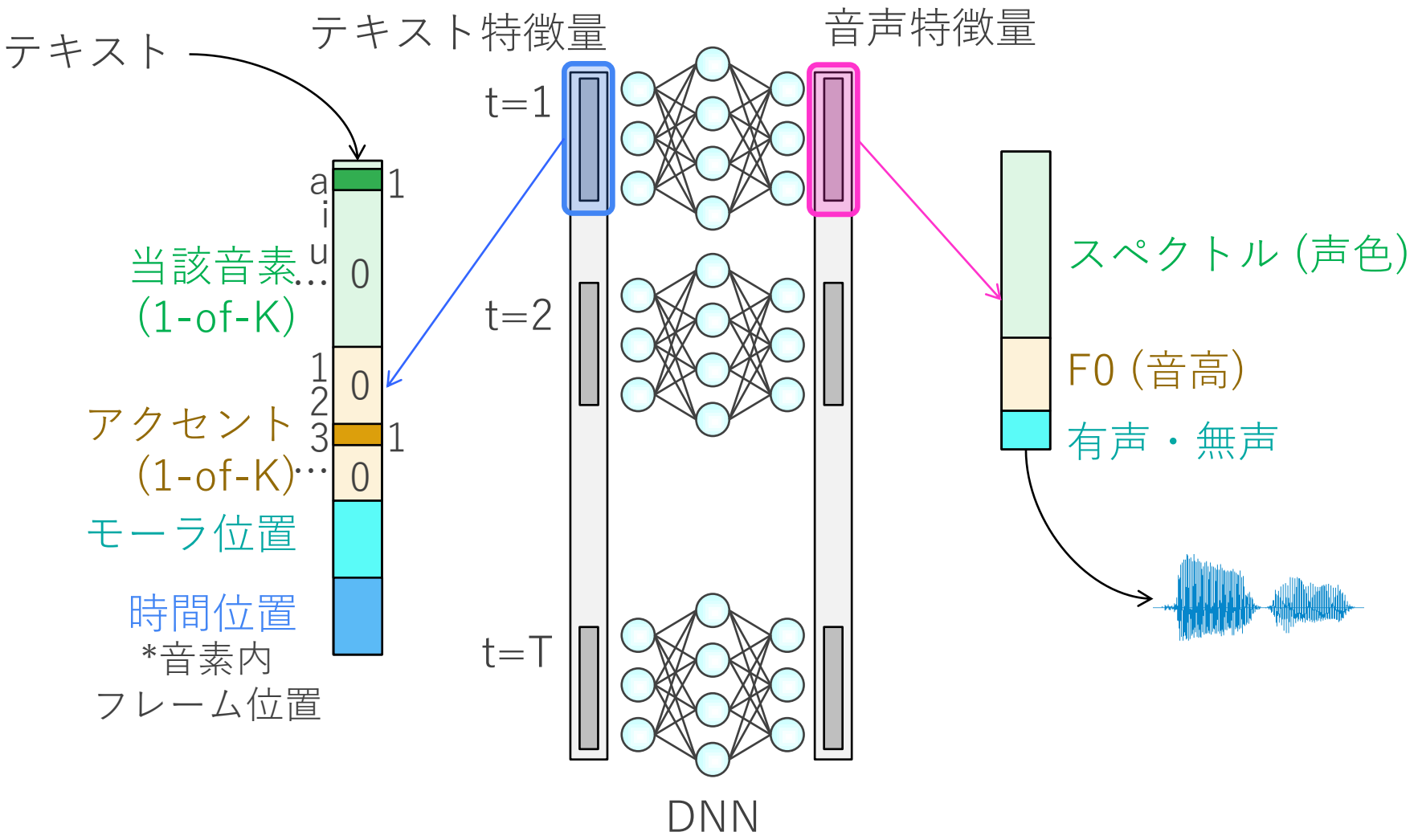
- トラジェクトリモデルの共分散を $\sigma^2 I$ で近似 (I は単位行列)
- MGE training for HMM[Wu06]/DNN[Wu16]

[zen09]
[Takamichi15] S. Takamichi et al., “Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion,” Proc. ICASSP, Apr. 2015.
[Hashimoto16]
[Kameoka15] H. Kameoka, “Modeling speech parameter sequences with latent trajectory hidden Markov model,” Proc. MLSP, Sep. 2015.
[Tobing16] P. L. Tobing et al., “Acoustic-to-articulatory inversion mapping based on latent trajectory Gaussian mixture model,” Proc. INTERSPEECH, Sep. 2016.
[Cai15] M.-Q. Cai et al., “Statistical parametric speech synthesis using a hidden trajectory model,” Speech Communication, 2015.
[Shannon13] M. Shannon et al., “Autoregressive models for statistical parametric speech synthesis,” IEEE Transactions, 2013
[Wang17] X. Wang et al., “An autoregressive recurrent mixture density network for parametric speech synthesis,” Proc. ICASSP, 2017.
[Wu06] Y.-J. Wu et al., “minimum generation error training for HMM-based speech synthesis,” Proc. ICASSP, 2006.
[Wu16] Z. Wu et al., “Improving Trajectory Modelling for DNN-based Speech Synthesis by using Stacked Bottleneck Features and Minimum Generation Error Training,” IEEE Transactions, 2016.

音声合成変換技術の変遷



DNN音声合成



DNNは自然音声特徴量との二乗誤差を最小化するように学習

[Zen13] Zen et al., “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, 2013.

HMM -> DNN -> RNN

HMM -> DNN で改善したこと[Zen13][Watt16]

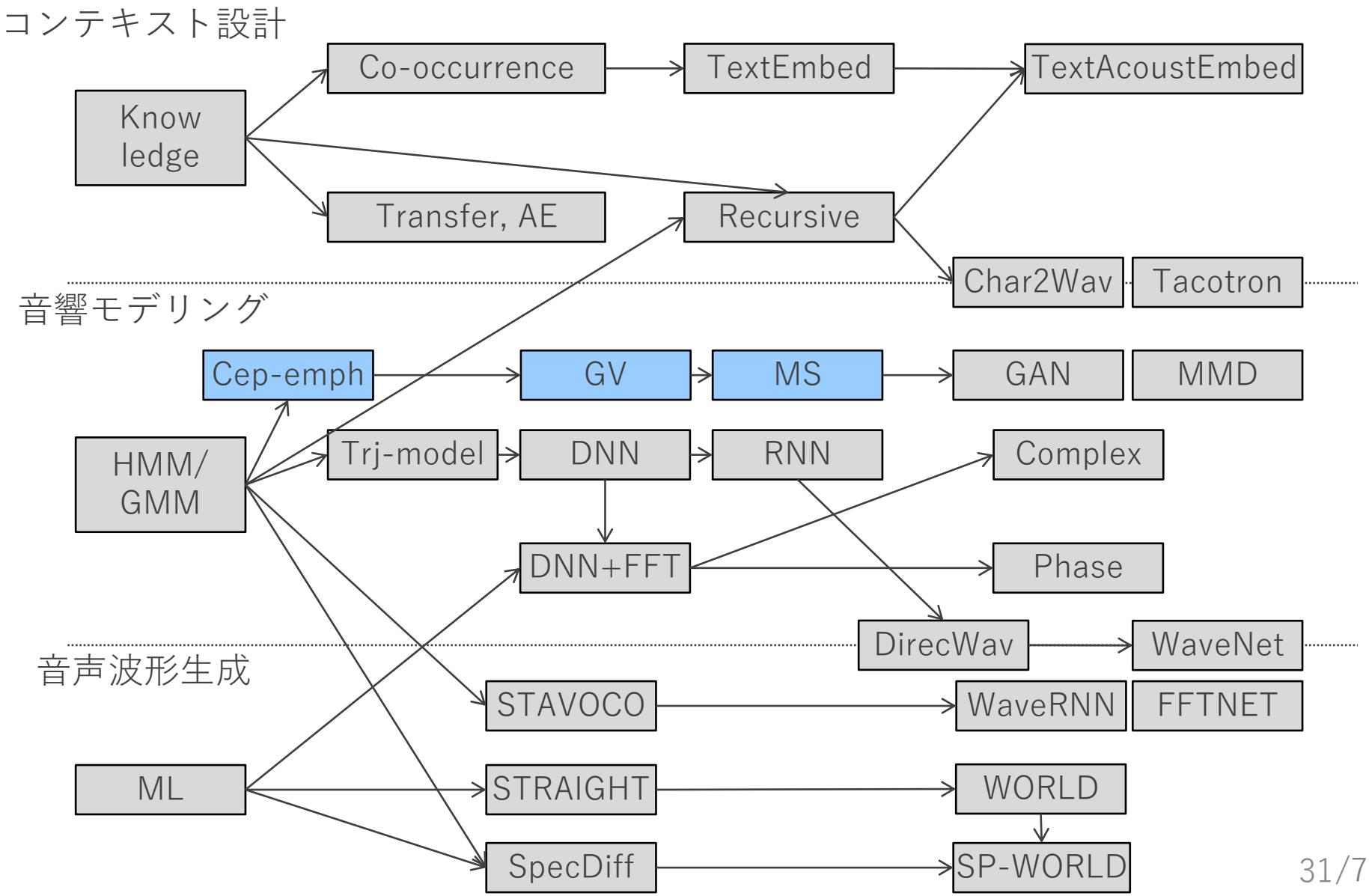
- 時間量子化の緩和：HMM状態 → フレーム
- 予測の精微化：クラスタリング→ 回帰
- 大規模データが利用可能に

DNN -> RNN (recurrent neural network) で改善したこと

- RNN: 時間的な再帰構造を持ったDNN
- 長期的な時間依存関係の獲得 [Fan14] (特にF0 [Wang16])
- 動的特徴量をモデルに内包 [Zen15]

[Zen13] Zen et al., “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, 2013.
[Watt16] Watts et al., “From HMMs to DNNs: where do the improvements come from?,” Proc. ICASSP, 2016.
[Fan14] Fan et al., ““TTS synthesis with bidirectional LSTM based recurrent neural networks,” Proc. INTERSPEECH, 2014.
[Wang16] Wang et al., “A Comparative Study of the Performance of HMM, DNN, and RNN based Speech Synthesis Systems Trained on Very Large Speaker-Dependent Corpora,” Proc. SSW, 2016.
[Zen15] Zen et al., “Unidirectional Long Short-Term Memory Recurrent Neural Network with Recurrent Output Layer for Low-Latency Speech Synthesis,” Proc. ICASSP, 2016.

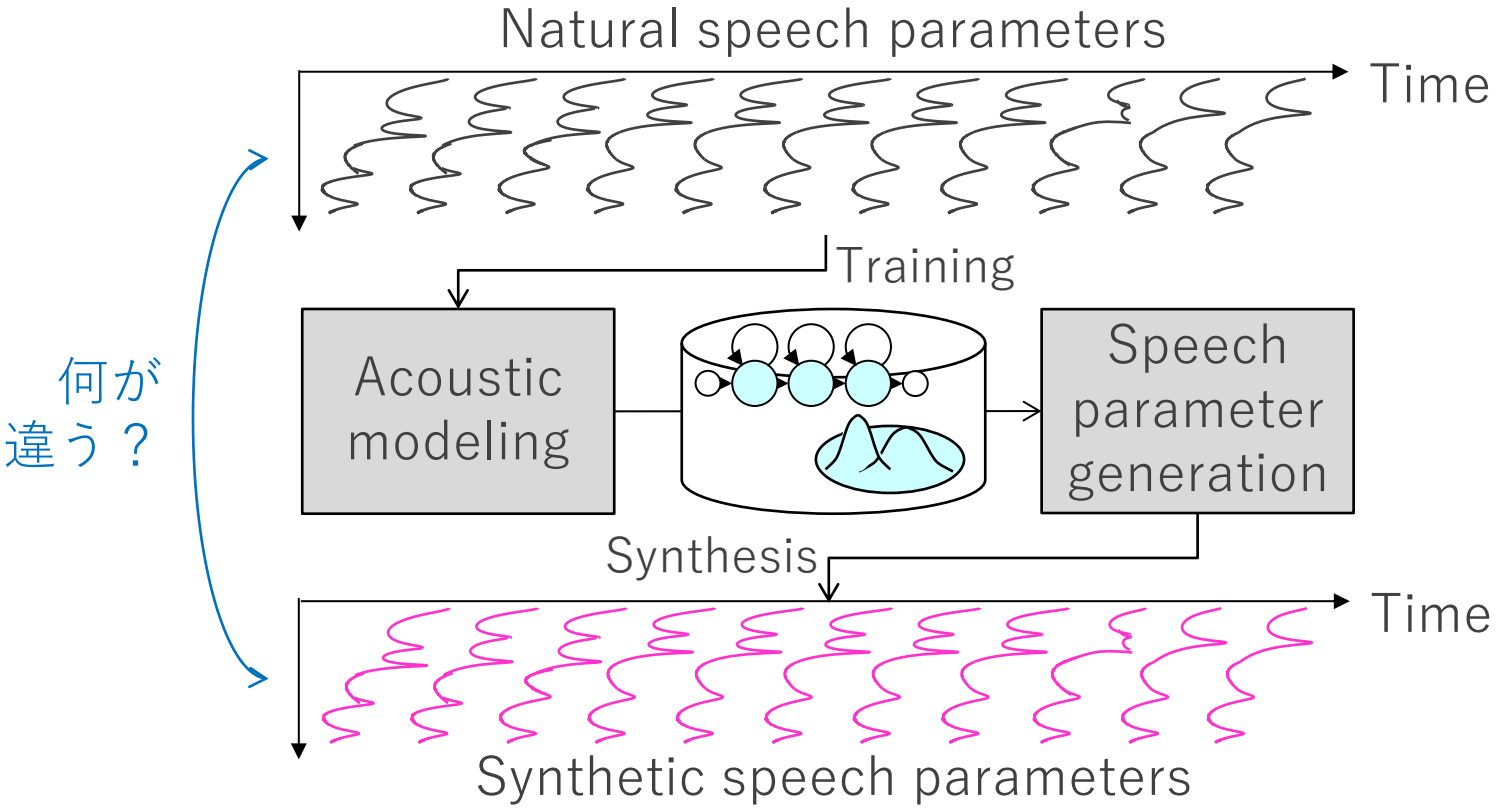
音声合成変換技術の変遷



生成パラメータの過剰な平滑化

過剰な平滑化とは

– 統計モデリングにおける平均化により，自然音声パラメータに含まれていた微細構造が消失すること．音質劣化の主要因

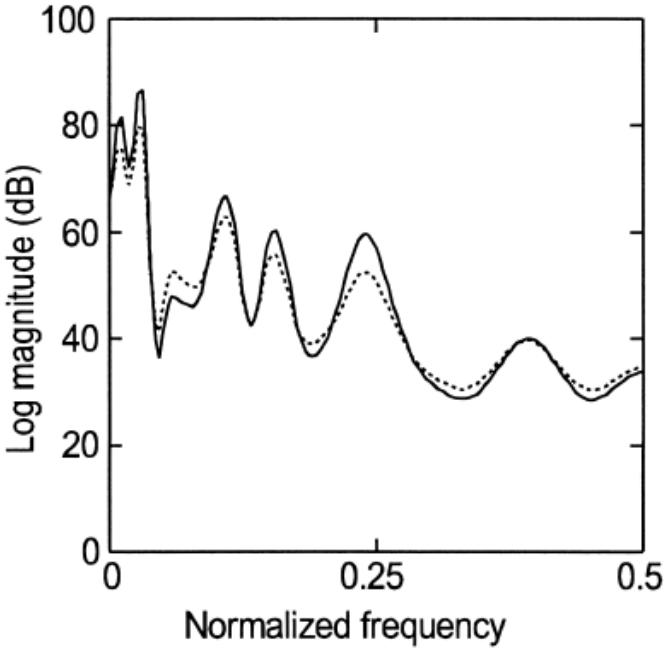


ケプストラム強調

ルールベースのフォルマント強調法

– ケプストラムの2次以上を定数倍

$$y'_t(d) = \beta y_t(d)$$ ($y_t(d)$は時刻t, d次元目 (d >= 2)のケプストラム)



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 $D(z) \cdot \bar{D}^{\beta}(z), \beta = 0.5$)

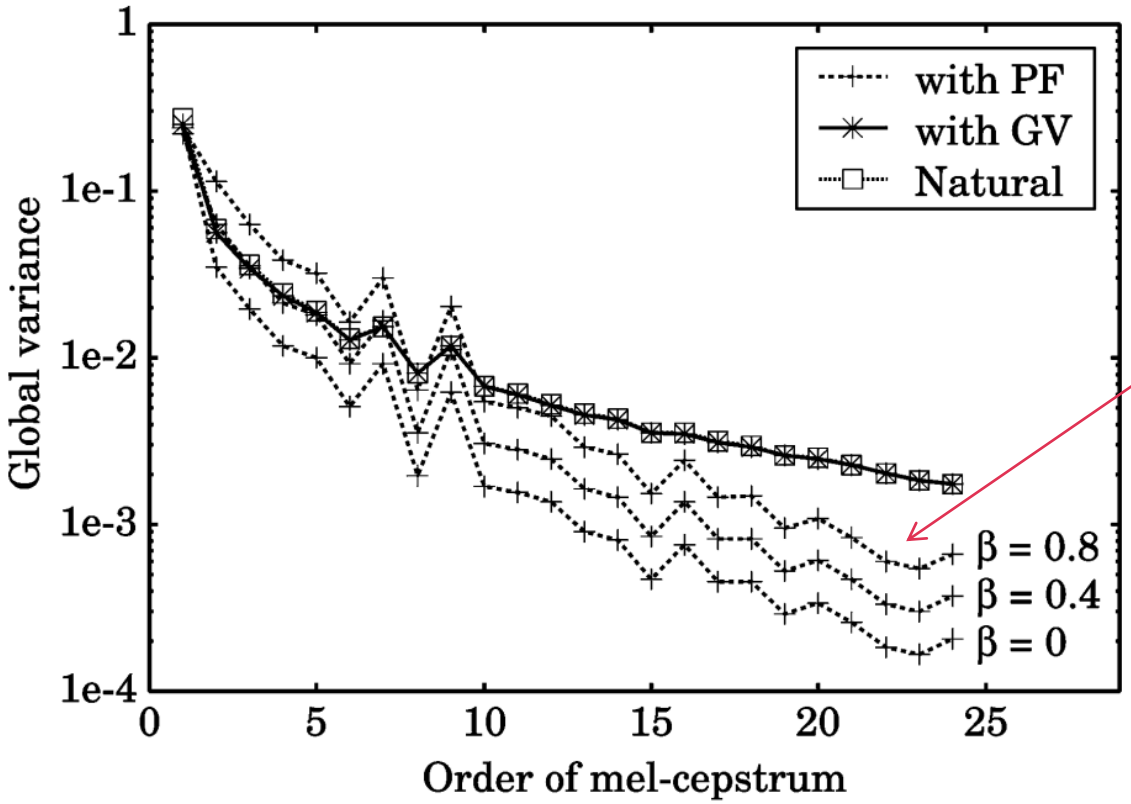
[Yoshimura04] Yoshimura et al., “Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis,” IEICE Transactions, 2004.

系列内変動 (Global Variance: GV)

時系列の “広がり” を捉えるデータドリブンの特徴量

– 定義：特徴量時系列の分散

$$v(d) = \mathrm{variance}([y_1(d), \cdots, y_t(d), \cdots, y_T(d)])$$



GV領域だと
ケプ強調は
バイアス

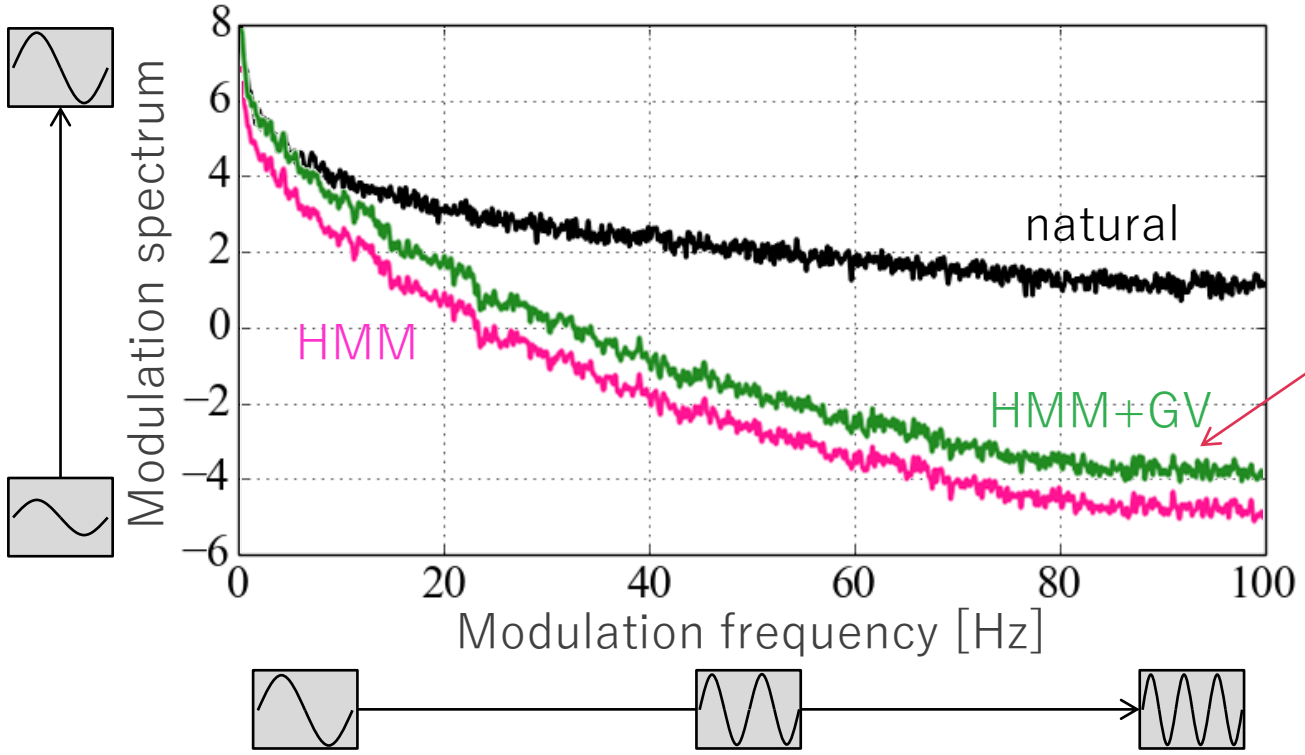
[Toda07] Toda et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Transactions, 2007.

変調スペクトル (Modulation Spectrum: MS)

時系列の “振動” を捉えるデータドリブンの特徴量

– 定義：特徴量時系列のパワースペクトル

$$\boldsymbol{s}(d) = |\mathrm{DFT}([y_1(d), \cdots, y_t(d), \cdots, y_T(d)])|^2$$



MS領域だと
GV強調は
バイアス

[Takamichi16] Takamichi et al., “Post-filters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis,” IEEE Transactions, 2016.

系列内変動・変調スペクトルの関連論文

音声パラメータ生成との統合

- GV/MSを補償する生成
- GV [Toda07]/MS [Takamichi15]尤度 とHMM/GMM/DNN尤度の Product of Experts

トラジェクトリモデルとの統合

- GV/MSを補償する学習
- GV [Toda09] / MS [Takamichi15-2] 制約付き(トラジェクトリ)学習

音質定量化の評価指標としての利用

- 音声合成 [Baljekar16]，歌声合成 [Blaauw17]

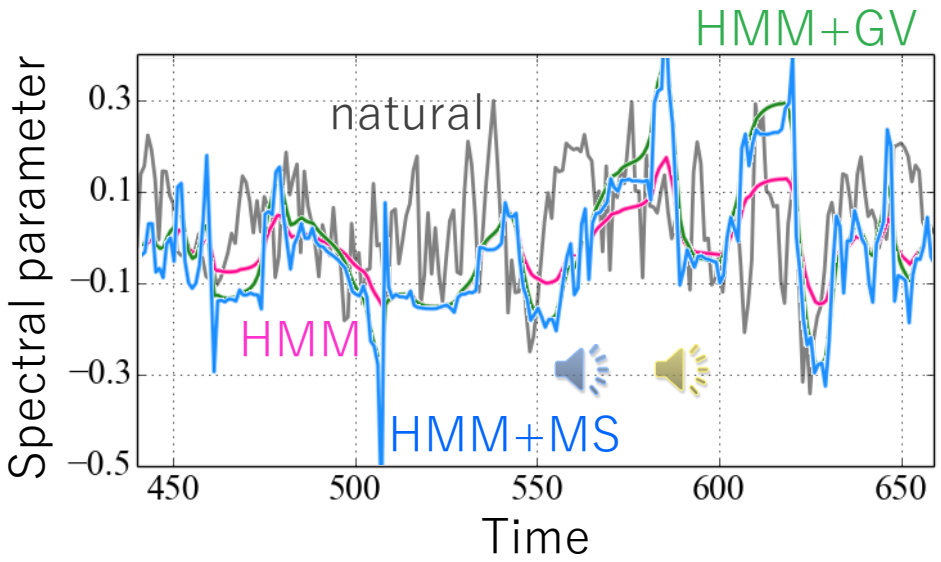
[Toda07] Toda et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Transactions, 2007.
[Takamichi15] S. Takamichi et al, “Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis,” Proc. ICASSP, 2015.
[Toda09] T. Toda et al., “Trajectory training considering global variance for HMM-based speech synthesis,” Proc. ICASSP, 2009.
[Takamichi15-2] S. Takamichi et al., “Modulation spectrum-constrained trajectory training algorithm for HMM-based speech synthesis,” Proc. ICASSP, 2015.
[Baljekar16] Baljekar et al., “Utterance Selection Techniques for TTS Systems Using Found Speech,” Proc. SSW9, 2016.
[Blaauw17] Blaauw et al., “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs,” Applied Science, 2017.

ケプストラム強調 vs.
系列内変動 vs. 変調スペクトル

変数・強調法としての違い（音質はMSが最も良い）

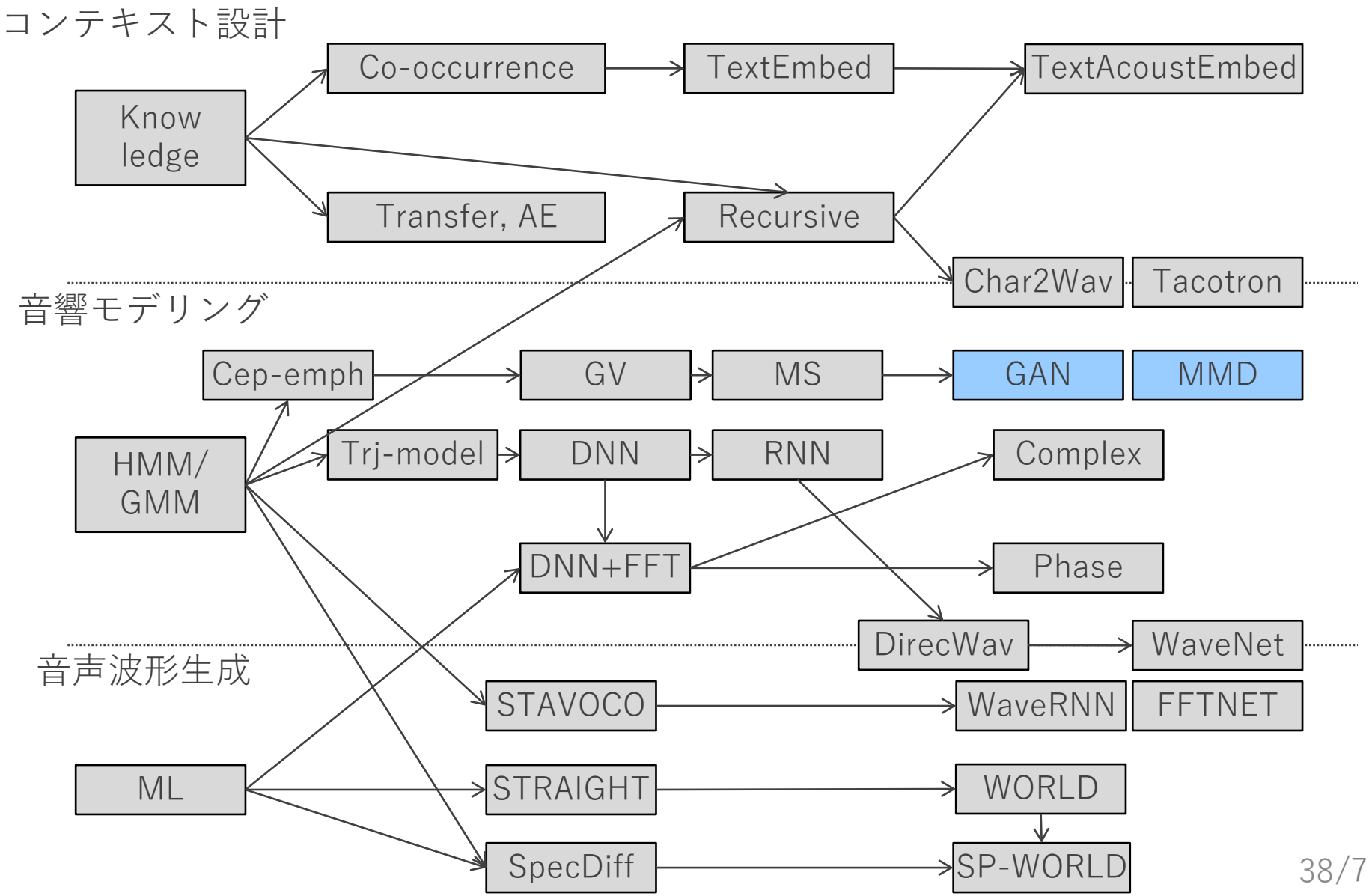
	ケプ強調	GV	MS
変数	スカラ	ベクトル	行列
特徴量毎の強調？	No	**Yes**	**Yes**
変調周波数毎の強調？	No	No	**Yes**
何を強調／復元？	フォルマント	スケール	振動

効果の違い



[Takamichi16] Takamichi et al., “Acoustic modeling and speech parameter generation for high-quality statistical parametric speech synthesis,” PhD thesis, 2016.

音声合成変換技術の変遷



深層生成モデルの利用へ

GV/MS

- Hand-crafted な特徴量
- 音質定量化には有効だが，音質改善効果に限界が.

敵対的学習 (GAN) の利用

- 二つのデータセット間の分布間距離を最小化 (分布補償)
  - GV/MS補償はモーメント補償なので近い手法とみなせる
- 複雑な分布に対しても適用可能であることが経験的に知られている

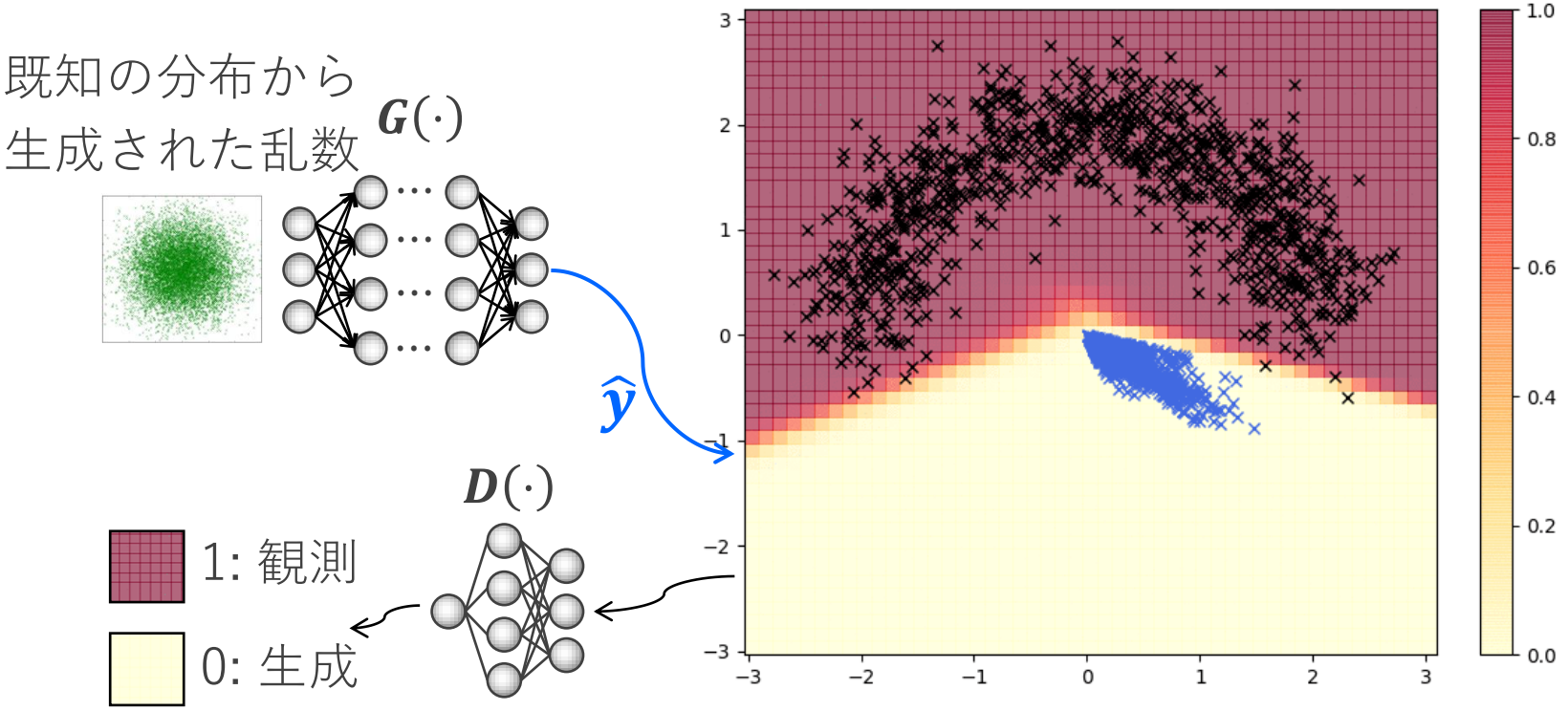
敵対的学習

Generative adversarial network [Goodfellow14]

– 分布間の近似 Jensen-Shannon divergence を最小化

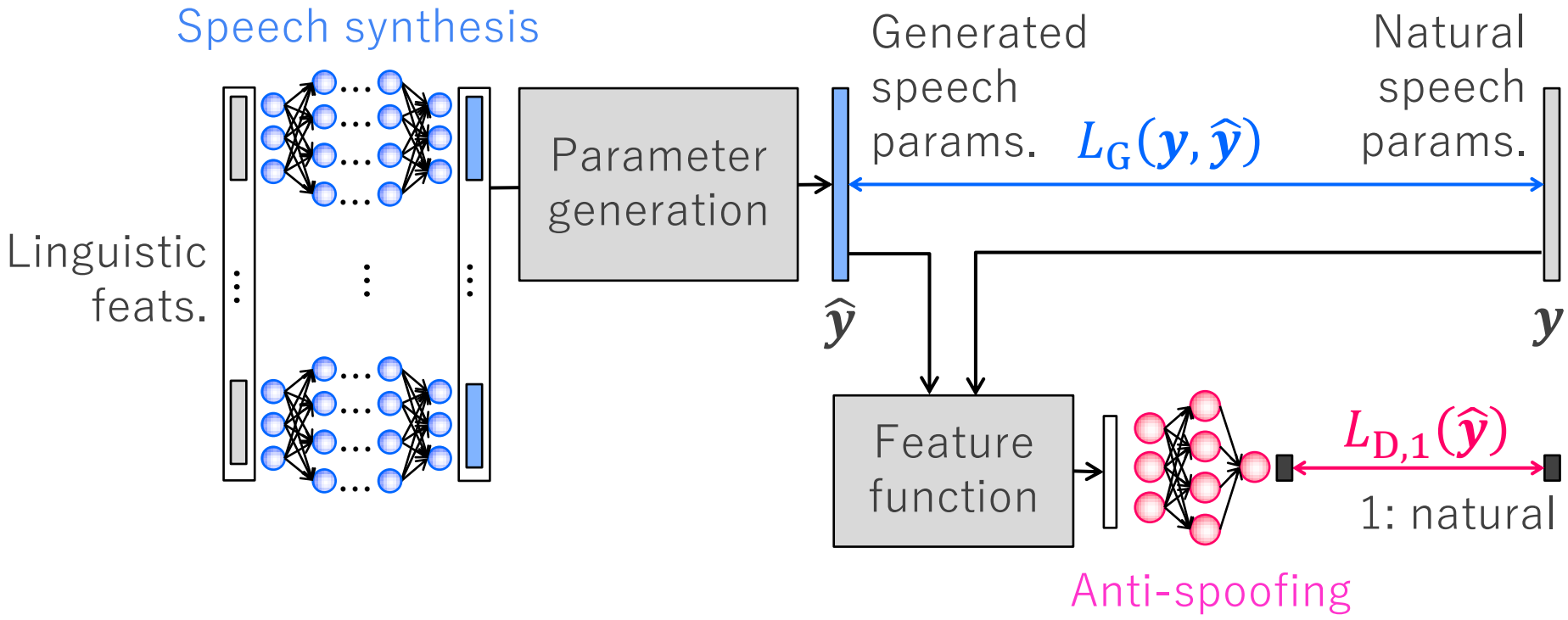
– 生成モデル $\boldsymbol{G}(\cdot)$ と，観測／生成データを識別するモデル $\boldsymbol{D}(\cdot)$ を敵対

$$\text{Loss} = E\left[\log\left(D(\boldsymbol{y})\right)\right] + E\left[\log\left(1 - D(\hat{\boldsymbol{y}})\right)\right] \text{（E[]は期待値）}$$



[Goodfellow14] Goodfellow et al., “Generative adversarial networks,” Proc. NIPS, 2014.

敵対的音声合成

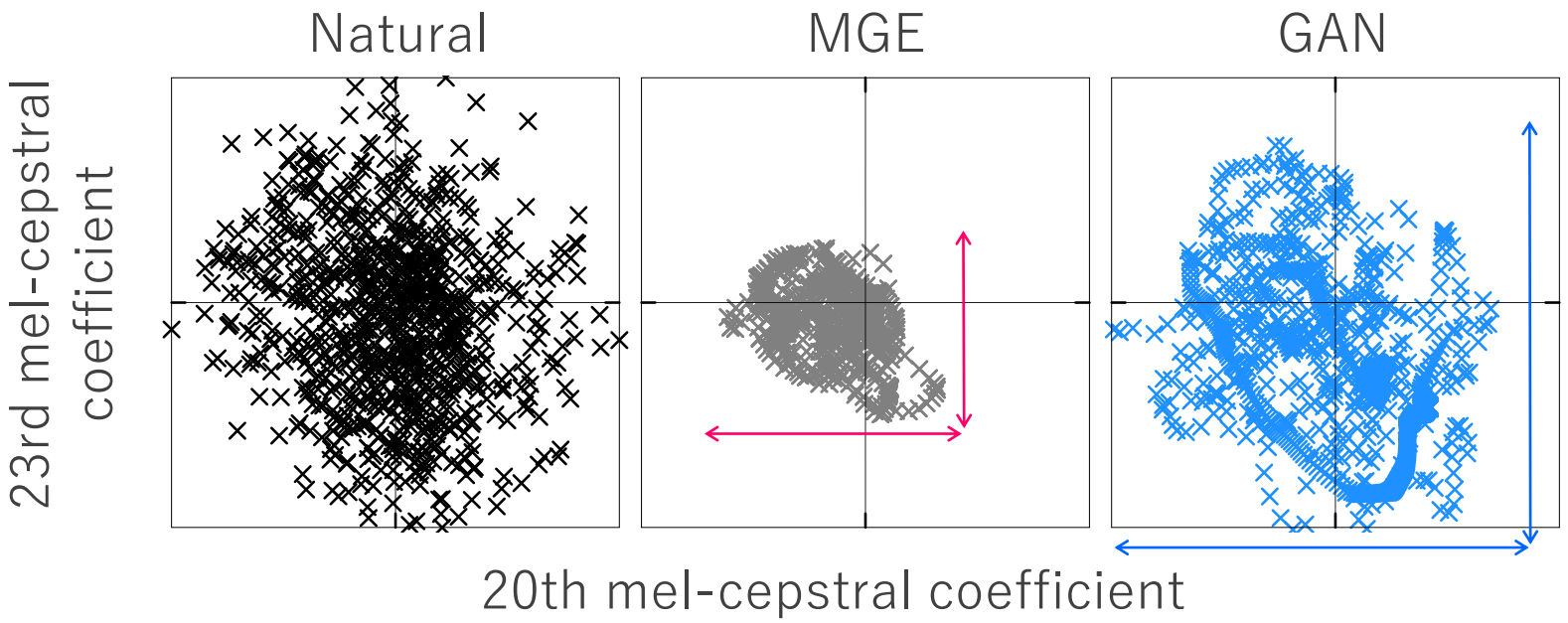


$L(y, \hat{y}) = L_G(y, \hat{y}) + \omega_D L_{D,1}(\hat{y})$ を最小化

生成誤差　　Anti-spoofingを騙す損失

[Saito18] Saito et al., “Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks,” IEEE Transactions, 2018.

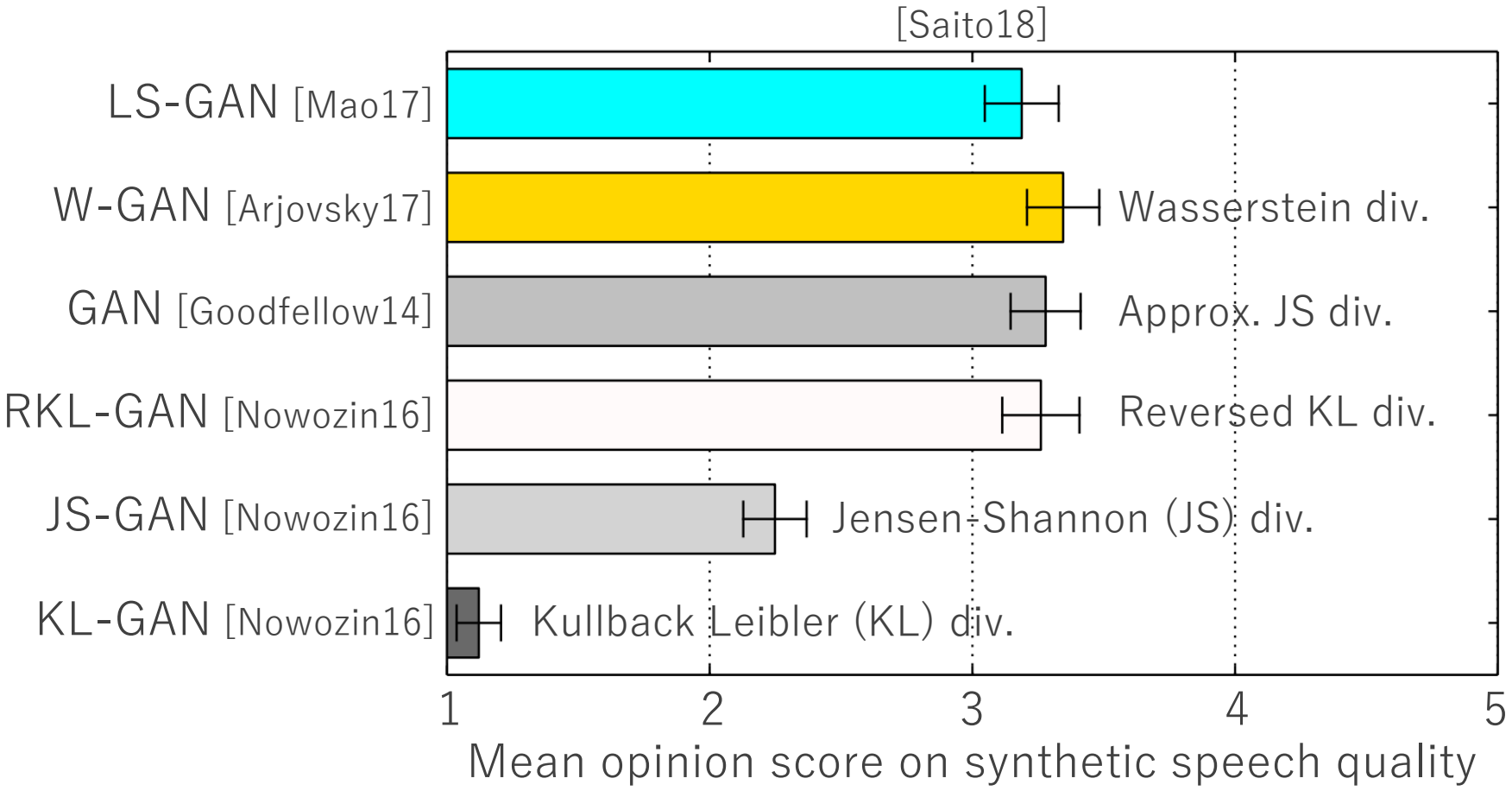
GANによる分布補償の効果



モーメントや分布を明示的に定義せずに分布を近づける

[Saito18] Saito et al., “Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks,” IEEE Transactions, 2018.

GANで最小化される距離規範の影響



[Mao17] Mao et al., ""Least squares generative adversarial networks," Proc. ICCV, 2017.
[Arjovsky17] Arjovsky et al., "Wasserstein GAN," Proc. ICML, 2017.
[Goodfellow14] Goodfellow et al., "Generative adversarial networks," Proc. NIPS, 2014.
[Nowozin16] Nowozin et al., "f-GAN: Training generative neural samplers using variational divergence minimization," Proc. NIPS, 2016.
[Saito18] Saito et al., "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks," IEEE Transactions, 2018.

音声合成×GANの最近の発展

DFTスペクトル・波形生成への応用

- DFTスペクトル（帯域分割 [Kaneko17]・帯域平均化 [Saito18-2]）
- 音声波形（1 frame 波形 [Juvela18]）

音声変換への応用

- DNN-based VC [Saito18]
- CycleGAN-based non-parallel VC [Kaneko18]
- StarGAN-based non-parallel VC [Kameoka18]

GAN以外の深層生成モデルの利用

- Generative moment-matching network

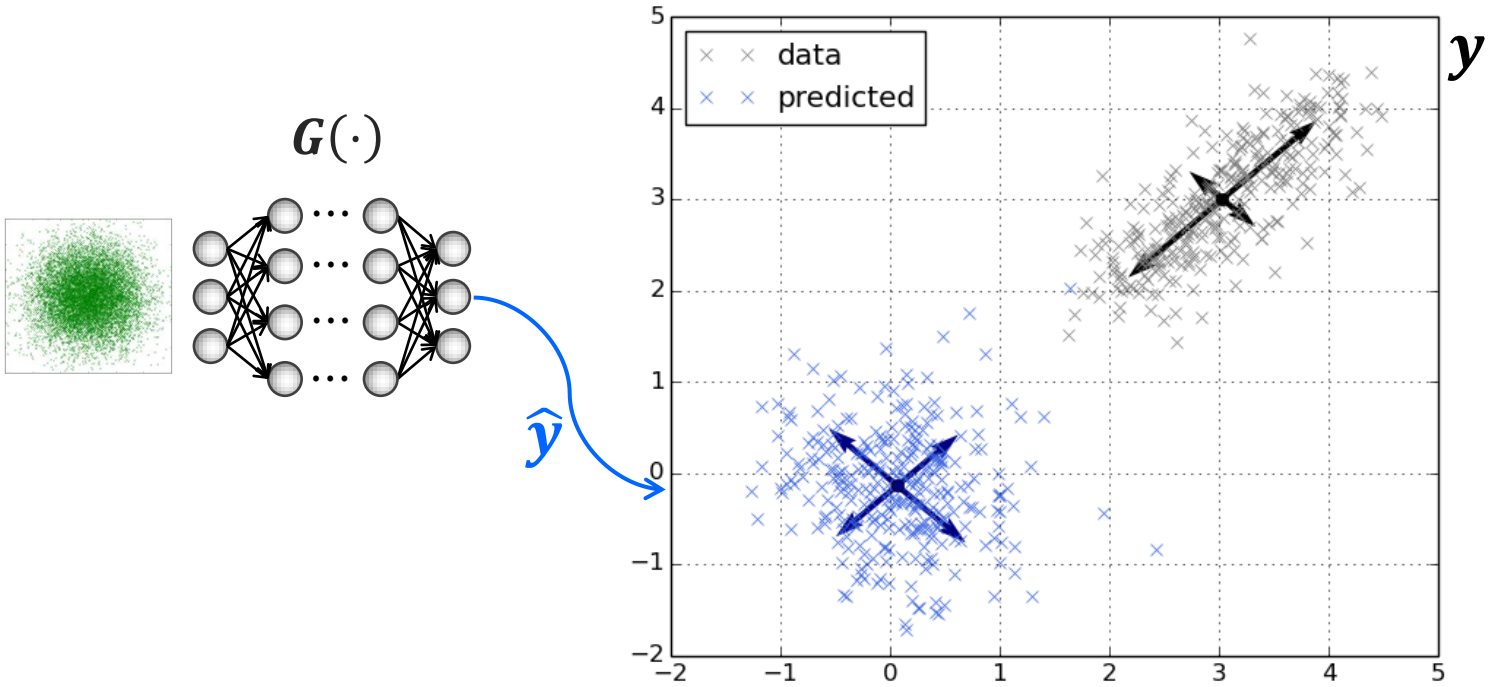
[Kaneko17] Kaneko et al., “Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks,” Proc. INTERSPEECH, 2017.
[Saito18-2] Saito et al., ““Text-to-speech synthesis using STFT spectra based on low-/multi-resolution generative adversarial networks,” Proc. ICASSP, 2018.
[Juvela18] Juvela et al., “Speech waveform synthesis from MFCC sequences with generative adversarial networks,” Proc. ICASSP, 2018.
[Saito18] Saito et al., “Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks,” IEEE Transactions, 2018.
[Kaneko18] Kaneko et al., “Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks,” arXiv, 2018.
[Kameoka18] Kameoka et al., “StarGAN-VC: Non-parallel Many-to-Many Voice Conversion with Star Generative Adversarial Networks,” arXiv, 2018.

Maximum mean discrepancy (MMD)

Moment matching network [Li15] [Ren16]

– 分布のモーメント間の二乗距離を最小化

$$L = \mathrm{tr}\left(\mathbf{1} \cdot \boldsymbol{K}_{\boldsymbol{y},\boldsymbol{y}}\right) + \mathrm{tr}\left(\mathbf{1} \cdot \boldsymbol{K}_{\hat{\boldsymbol{y}},\hat{\boldsymbol{y}}}\right) - 2\mathrm{tr}\left(\mathbf{1} \cdot \boldsymbol{K}_{\boldsymbol{y},\hat{\boldsymbol{y}}}\right) \left(\boldsymbol{K}_{\boldsymbol{y},\boldsymbol{y}}\text{は}\boldsymbol{y},\boldsymbol{y}\text{間のグラム行列}\right)$$



[Li15] Li et al., “Generative moment matching network,” Proc. ICML, 2015.
[Ren16] Ren et al., ““Conditional generative moment-matching networks,” Proc. NIPS, 2016.

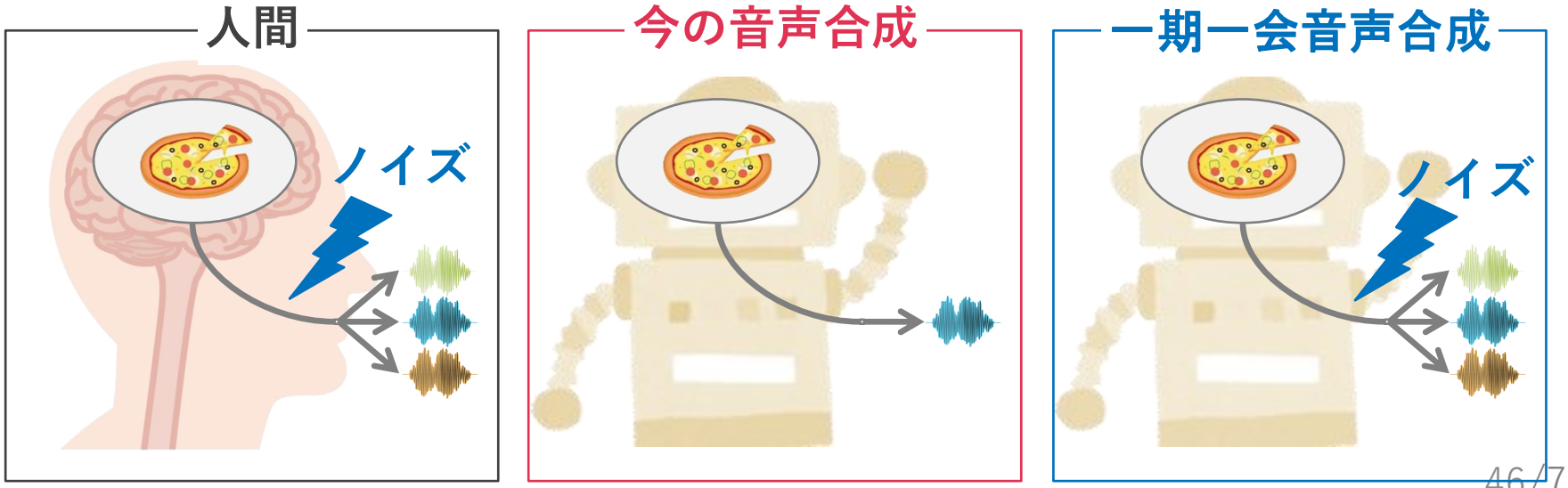
(余談) 一期一会音声合成への拡張

Moment-matching network の利点

- 単なる最小化問題なので，GANに比べ安定して学習
- モーメントを明示的に取り入れられる
- 音質劣化なしで音声パラメータをランダムサンプリング [Takamichi17]

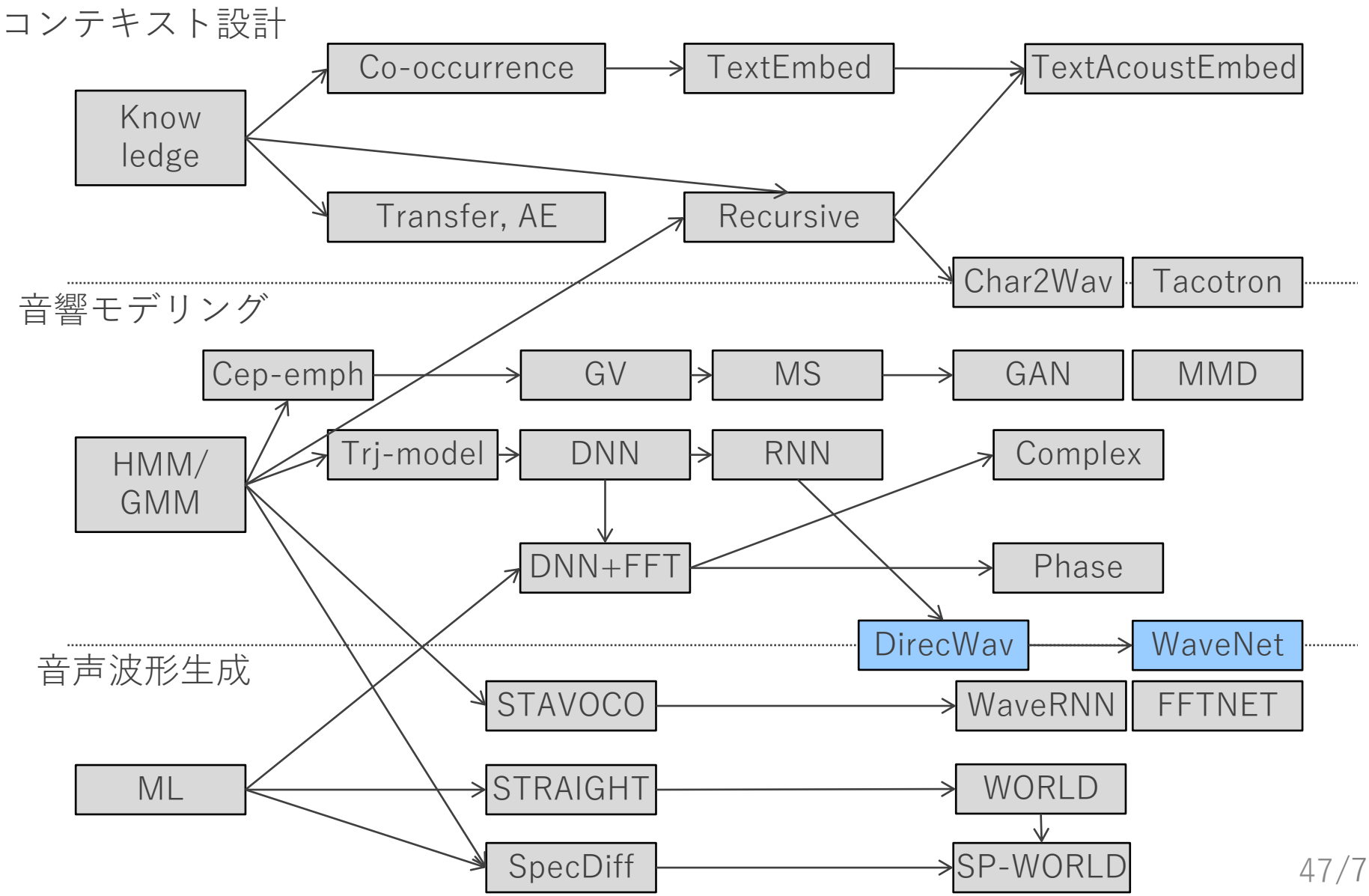
一期一会音声合成への応用 [Takamichi17]

- 人間の発話間変動を再現する音声合成



[Takamichi17] Takamichi et al., “Sampling-based speech parameter generation using moment-matching networks,” Proc. INTERSPEECH, 2017.

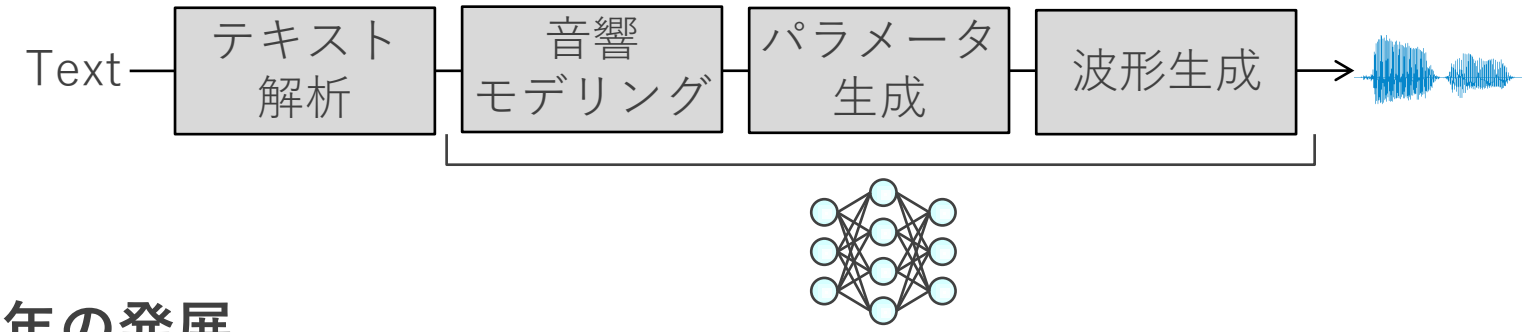
音声合成変換技術の変遷



波形を直接出力するDNNへ

音声パラメータ生成から波形生成へ

– 各モジュールの個別学習から同時学習へ



近年の発展

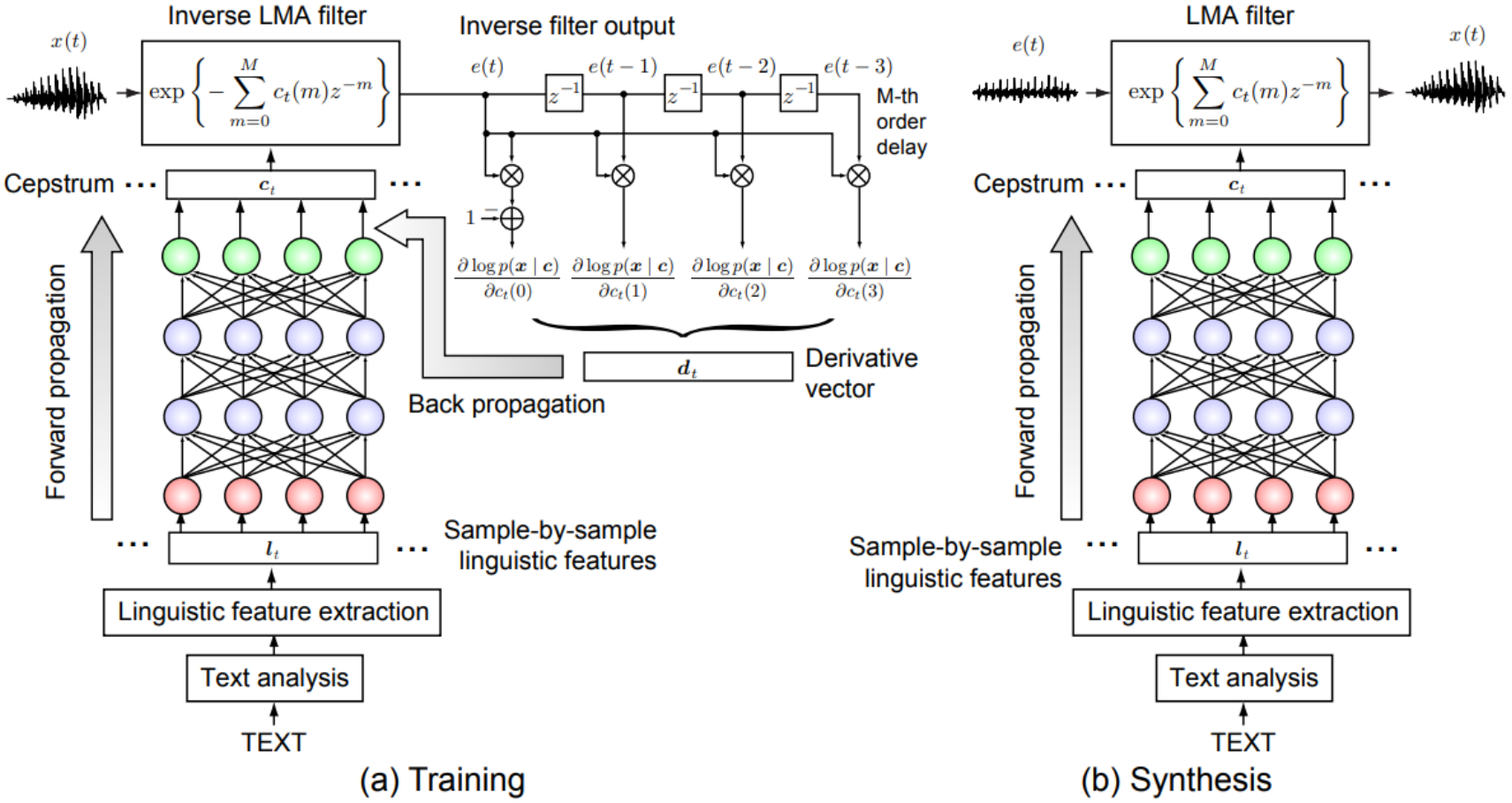
– Integration of feature extraction and modeling [Nakamura14]

– Direct waveform modeling by DNNs [Tokuda15]

– WaveNet [Oord16]

フレーム分析とガウス分布を仮定した Direct Waveform modeling

– フレーム毎の声道フィルタパラメータをDNNで予測
– 音声波形のガウス性を仮定して最尤推定

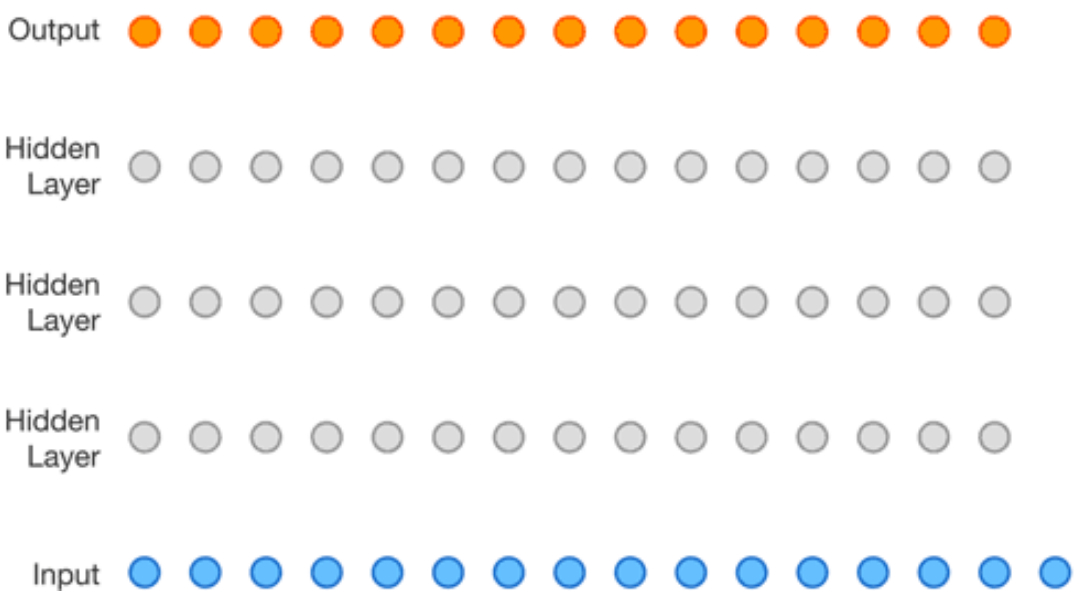


[Tokuda15] Tokuda et al., “Directly Modeling Speech Waveforms by Neural Networks for Statistical Parametric Speech Synthesis,” Proc. ICASSP, 2015.
[Tokuda16] Tokuda et al., “Directly Modeling Voiced and Unvoiced Components in Speech Waveforms by Neural Networks,” Proc. ICASSP, 2016.

WaveNet

離散化された波形を1サンプル毎に予測する深層生成モデル

- Receptive field を広げるための dilated convolution
- AR (auto-regressive) 過程による音サンプル生成
- 下の灰色の点は，dilated conv., gated activation, 1x1 conv., residual networkから成る



[Webpage] https://deepmind.com/blog/wavenet-generative-model-raw-audio/
[Oord16] Oord et al., “WaveNet: a generative model for raw audio,” arXiv, 2016.

WaveNetに関する近年の発展

生成の高速化

- Parallel WaveNet [Oord17] … ARを過程をMA過程で近似
- Subband WaveNet [Okamoto17] … 帯域分割で並列生成

挙動の分析

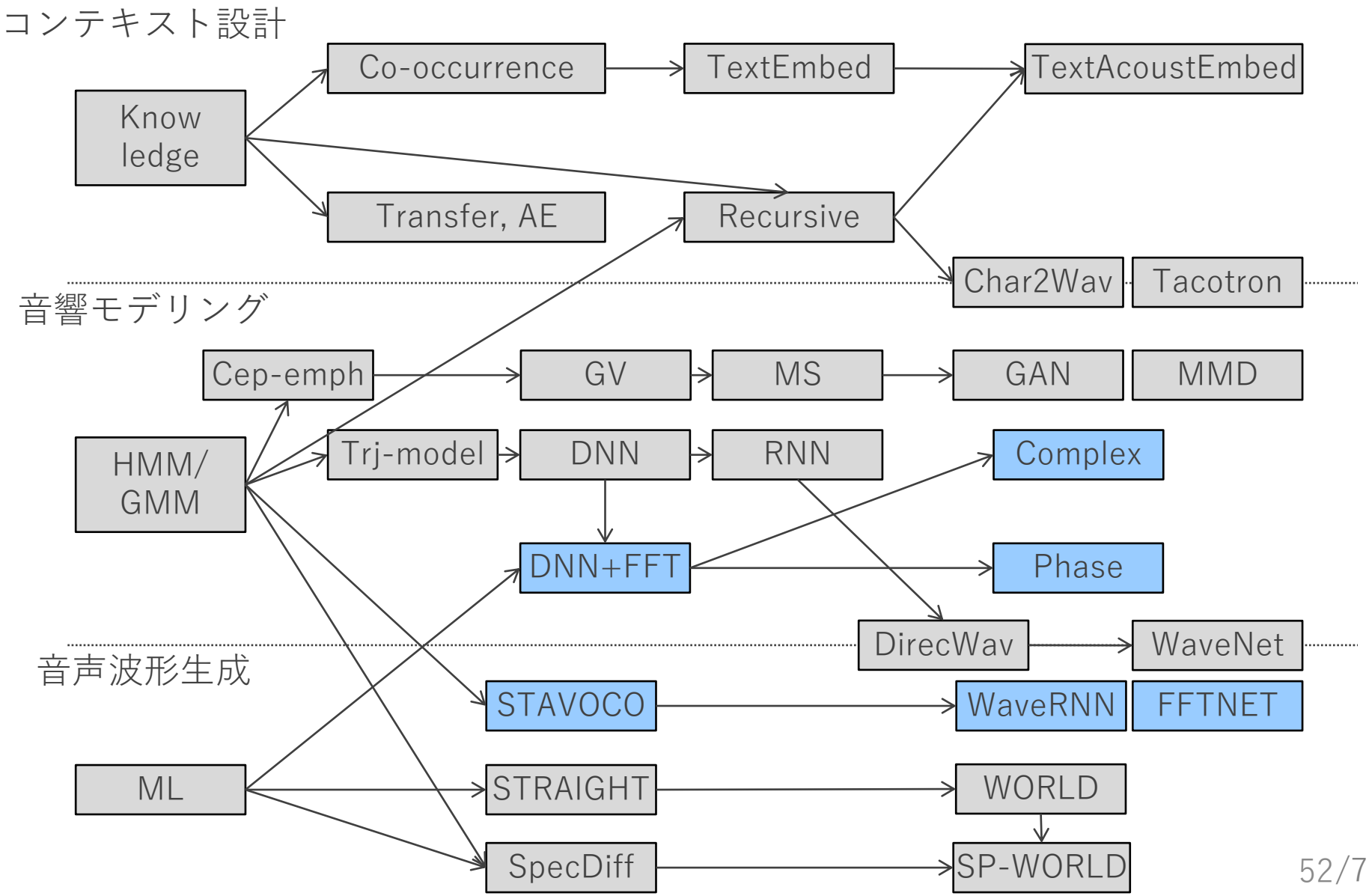
- Data size [Vit18] … 学習データ量と品質の調査
- Interpretation [Hua18] … モデルパラメータの挙動を調査

応用展開

- Low-rate speech coding [Kleijn17]
- Bayesian WaveNet-based speech enhancement [Qian17]

[Oord17] Oord et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” arXiv, 2017.
[Okamoto17] Okamoto et al., “Subband WaveNet with overlapped single-subband filterbanks,” Proc. ASRU, 2017.
[Vit18] Vit et al., “On the analysis of training data for WaveNet-based speech synthesis,” Proc. ICASSP, 2018.
[Hua18] Hua, “Do WaveNets Dream of Acoustic Waves?,” arXiv, 2018.
[Kleijn17] Kleijn et al., “Wavenet based low rate speech coding,” arXiv, 2017.
[Qian17] Qian et al., “Speech Enhancement Using Bayesian Wavenet,” Proc. INTERSPEECH, 2017.

音声合成変換技術の変遷



信号処理／統計的ボコーダ と DFTスペクトル

ボコーダの役割

- 波形生成DNNが現れる中でも，ボコーダは，音声を直感的なパラメータ空間で扱える，重要な手段である

信号処理ボコーダ (STRAIGHT, WORLDなど)

- 利点：学習データ不要なのでポータビリティが高い
- 欠点：再合成音声の音質が，自然音声よりわずかに低い

統計的ボコーダ

- 利点：自然音声の音質に近い合成音声を復元可能
- 欠点：一定量の学習データが必要でポータビリティは低い

統計的ボコーダ

統計的ボコーダのさきがけ

– STAVOCO [Toda08] … Factored trajectory HMMベース

DNNベースの統計的ボコーダ

– WaveNet vocoder [Tamamori17] … WaveNet TTS をボコーダに
– SampleRNN vocoder [Ai18] … SampleRNN TTS をボコーダに
– WaveRNN vocoder [Kalchbrenner18] … 時間subscaleによる高速生成
– FFTNET vocoder [Jin18] … Deep Cooley-Tukey型FFT的な

[Toda08] Toda et al., “Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM,” Proc. ICASSP, 2008.
[Tamamori17] Tamamori et al., “Speaker-dependent WaveNet vocoder,” Proc. INTERSPEECH, 2017.
[Ai18] Ai et al., “SampleRNN-based neural vocoder for statistical parametric speech synthesis,” Proc. ICASSP, 2018.
[Kalchbrenner18] Kalchbrenner et al., “Efficient Neural Audio Synthesis,” arXiv, 2018.
[Jin18] Jin et al., “FFTNET: a real-time speaker-dependent neural vocoder,” Proc. ICASSP, 2018.

FFTNET vocoder

WaveNet と Fast Fourier Transform (FFT) の共通点？

- 「 Dilated conv. と Cooley-Tukey 型FFTの構造って似てるよね」
- Dilated conv. の各層は，一種のダウンサンプリング (とみなせる)

FFTNET

- バタフライ演算機構に影響された1x1 conv. のstack.
- WaveNet vocoder に比べ省パラメータ (リアルタイム合成可)

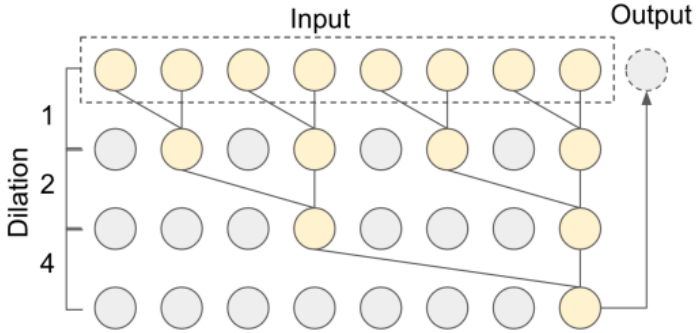
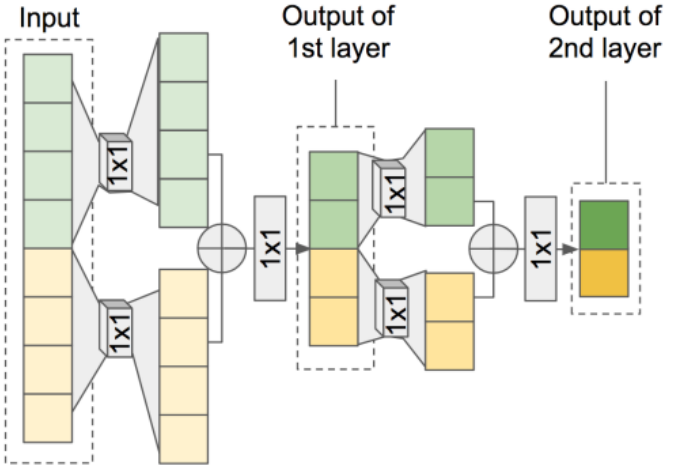


Fig. 1. Dilated convolution in WaveNet



[Jin18] Jin et al., “FFTNET: a real-time speaker-dependent neural vocoder,” Proc. ICASSP, 2018.

DFTスペクトルを直接生成するDNNへ

ボコーダ特徴量 vs. DFTスペクトル

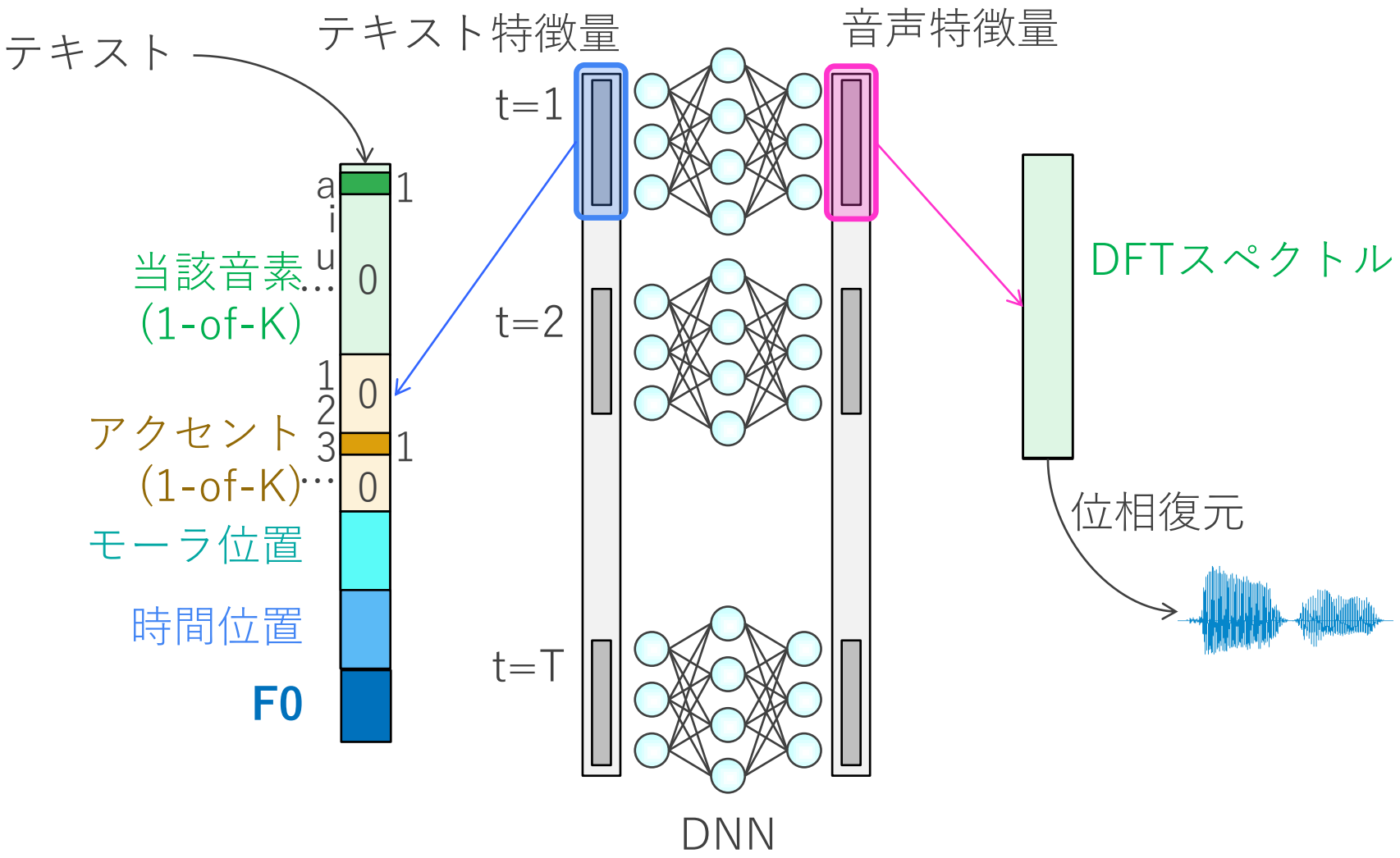
- 相反する利点・欠点がある
- ボコーダ：他の音声処理との接続性が悪いが，低次元＆直感的
- DFT：高次元特徴量だが，他の音声処理との接続性は良い

DFTスペクトルを直接生成する方式へ

- 振幅スペクトルを生成するDNN [Takaki17]
  - 位相は，Griffin-Lim位相復元法など [Griffin84] で別途推定

[Takaki17] Takaki et al., “Direct Modeling of Frequency Spectra and Waveform Generation Based on Phase Recovery for DNN-Based Speech Synthesis,” Proc. INTERSPEECH, 2017.
[Griffin84] Griffin et al., “Signal estimation from modified short-time Fourier transform,” IEEE Transactions, 1984.

DFTスペクトルを直接生成するDNN



[Takaki17] Takaki et al., “Direct Modeling of Frequency Spectra and Waveform Generation Based on Phase Recovery for DNN-Based Speech Synthesis,” Proc. INTERSPEECH, 2017.

複素数表現・位相表現

波形を生成するために，位相情報も扱えないか？

スペクトルの複素表現に基づくDNN

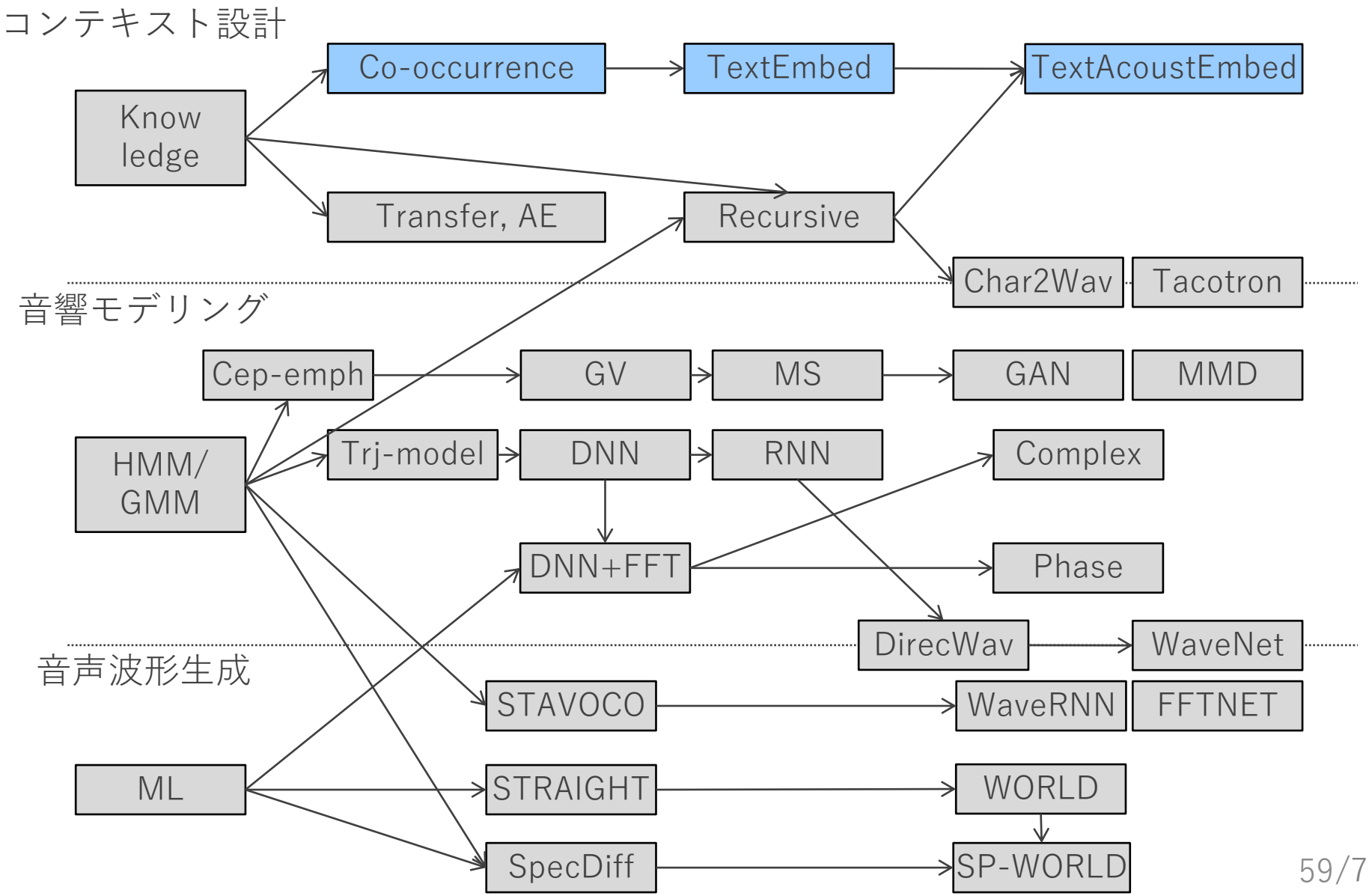
- Complex-valued Feed-Forward [Hu16]
- Complex-valued RBM [Nakashika17] … 複素ガウシアン

スペクトルの極座標表現に基づくDNN

- von Mises分布DNNに基づく位相推定 [高道18]
  - 周期変数をモデル化する深層生成モデル

[Hu16] Hu et al., “Initial investigation of speech synthesis based on complex-valued neural networks,” Proc. ICASSP, 2016.
[Nakashika17] Nakashika et al., “Complex-valued restricted Boltzmann machine for direct learning of frequency spectra,” Proc. INTERSPEECH, 2017.
[高道18] 高道 他, “von Mises分布DNNに基づく振幅スペクトログラムからの位相復元," 情報処理学会研究報告, 2018.

音声合成変換技術の変遷



テキスト音声合成のためのコンテキスト

通常，コンテキストは言語知識に基づいて設計されてきた

- 音素
  - 前後の音素，当該音素
- シラブル／モーラ
  - {前の／当該／後ろの}シラブルの音素数・位置
  - {前の／当該／後ろの}シラブルのアクセント・ストレス
  - 当該単語内のシラブル位置
- 単語
  - {前の／当該／後ろの}単語のシラブル数・位置
  - 当該フレーズ内の単語位置
- フレーズ・文

low-resource language*では利用困難・データドリブンでない

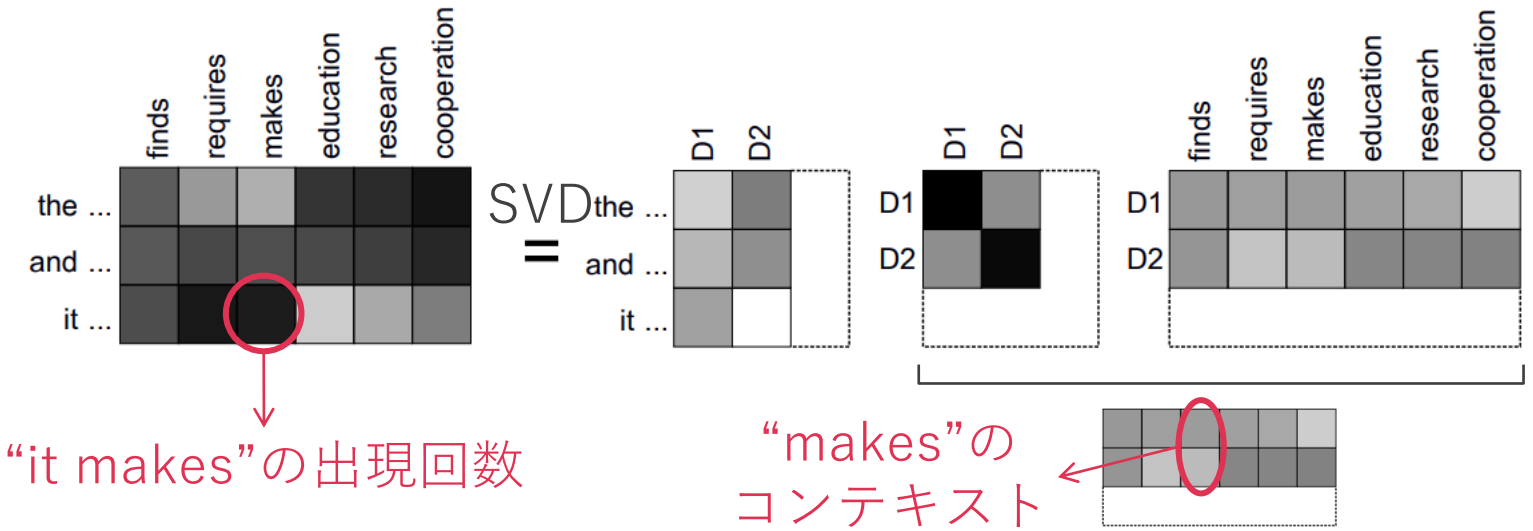
- *言語知識の整理されていない希少言語

[Yoshimura99] Yoshimura et al., “Simultaneous modeling of spectrum, pitch, and duration in HMM-based speech synthesis,” Proc. EUROSPEECH, 1999. (for Japanese)
[Tokuda02] Tokuda et al., “An HMM-based speech synthesis system applied to English,” Proc. ICASSP, 2002. (for English)
[Qian06] Qian et al., “An HMM-based Mandarin Chinese text-to-speech system,” Proc. ISCSLP, 2006. (for Chinese)

単語共起頻度に基づく分散表現

単語バイグラム (2-gram)の頻度行列化と低次元化

- 「近い共起頻度を持つ単語は近いコンテキストを持つ」ことを仮定
- 頻度行列をSVD (特異値分解) などで低次元圧縮.

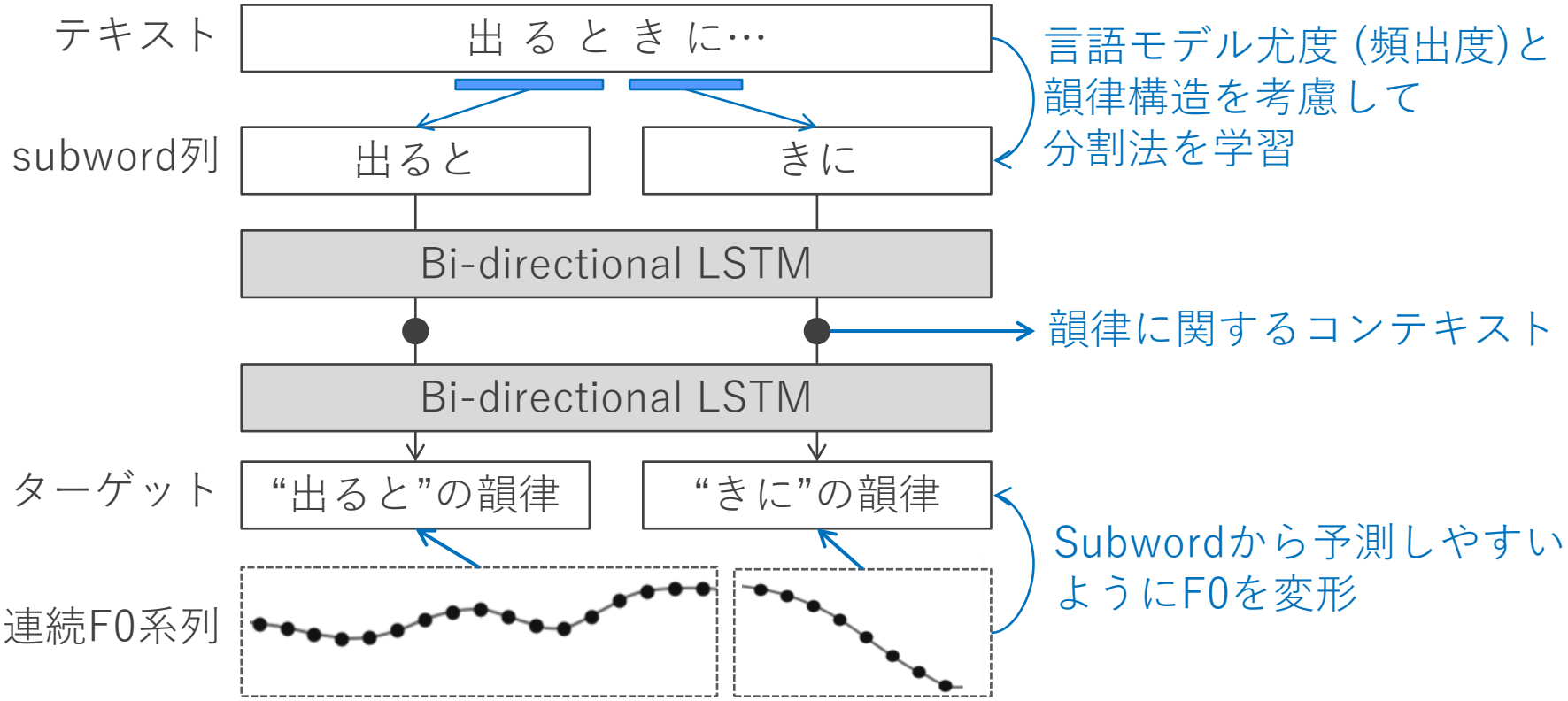


[Watts12] Watts et al., “Unsupervised Learning for Text-to-Speech Synthesis,” Ph.D thesis, 2012.

Subword分割と音響的サブワード埋め込み

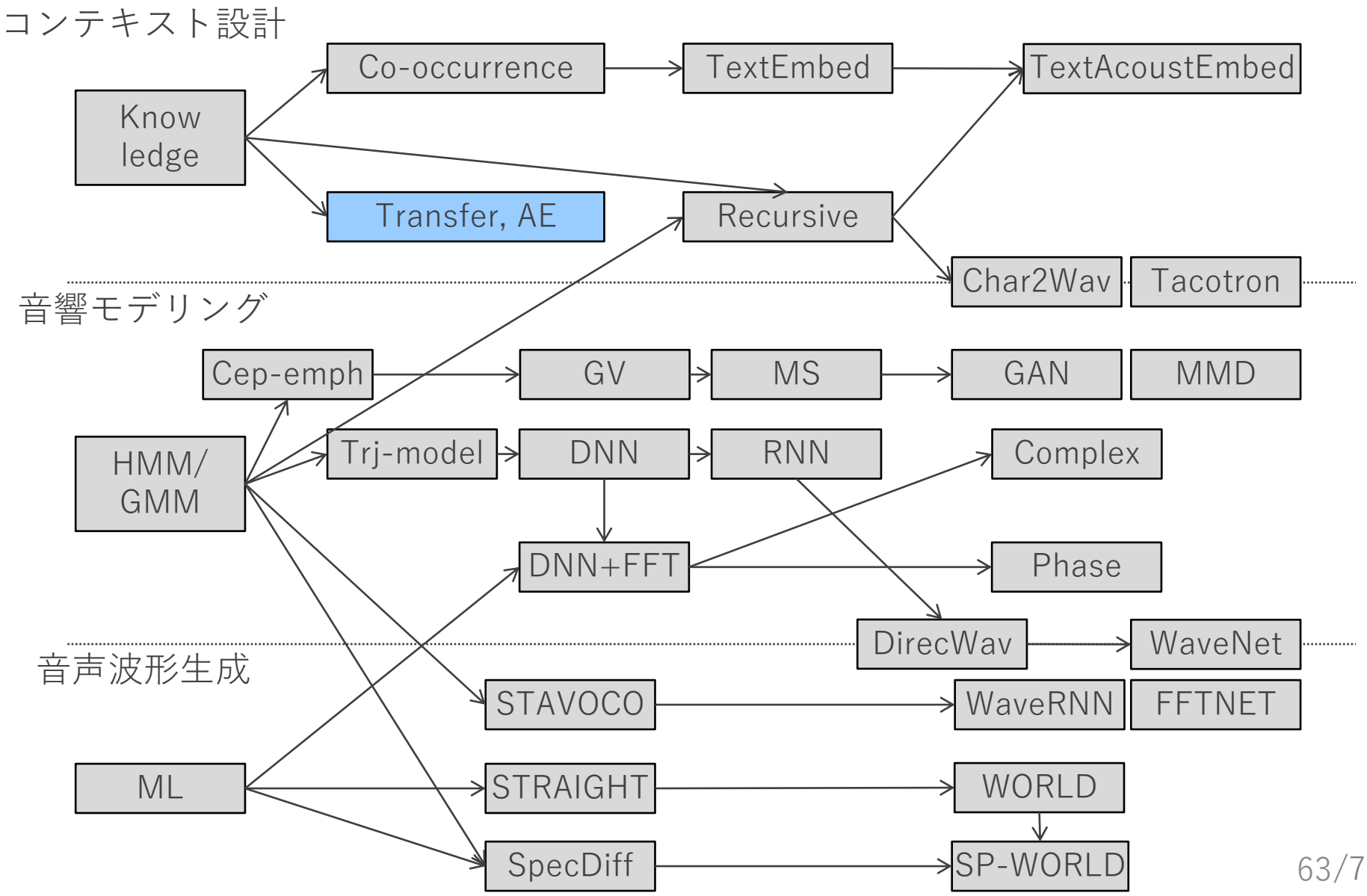
単語に代わる分割法とDNNに基づく埋め込み

- 「近いF0を持つサブワードは近いコンテキストを持つ」ことを仮定
- 単語数爆発に伴う学習の困難さを教師なし分割法で緩和



[高道18] 高道 他, “"日本語韻律構造を考慮したprosody-aware subword embeddingとDNN多方言音声合成への適用," 情報処理学会研究報告, 2018

音声合成変換技術の変遷

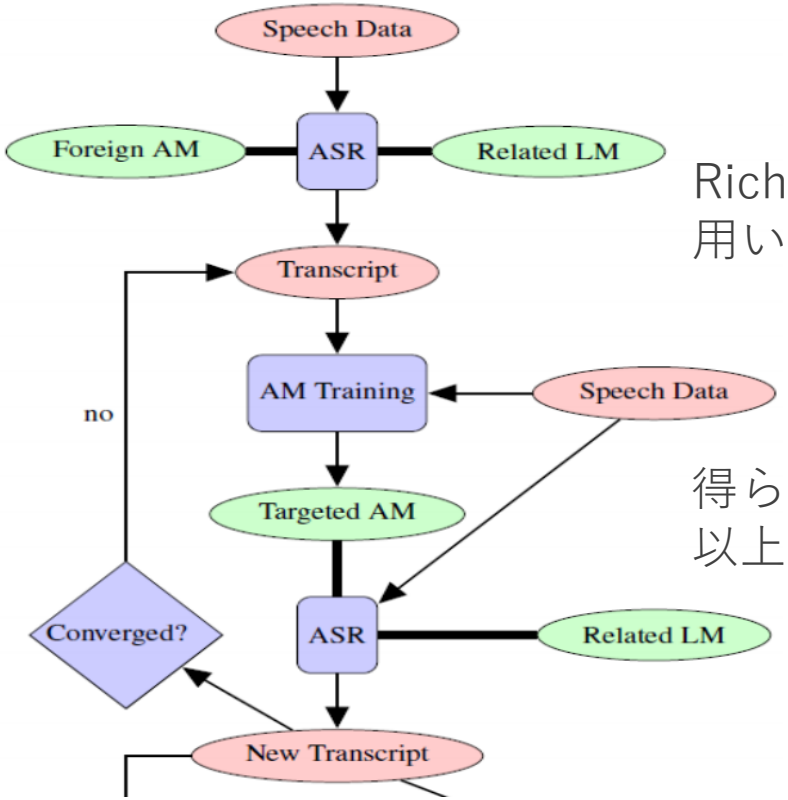


Rich-resourced languageのモデルを用いた, low-resourced language の音声合成

希少言語のコンテキストをどう作るか？

– 正書法のない(= written form が定まっていない) 言語もある.

主要言語の言語／音響モデルを利用 [Sitaram13]



AM/LM … 音響モデル・言語モデル

Rich-resourced languageの音声認識 (ASR) を用いてテキスト (transcript)を推定

得られたテキストを用いて音声合成.
以上を繰り返し.

[Sitaram13] Sitaram et al., “Text to Speech in New Languages without a Standardized Orthography,” Proc. SSW8, 2013.

Auto-encoderに基づく音声の圧縮・変換

Auto-encoder (AE) による次元圧縮

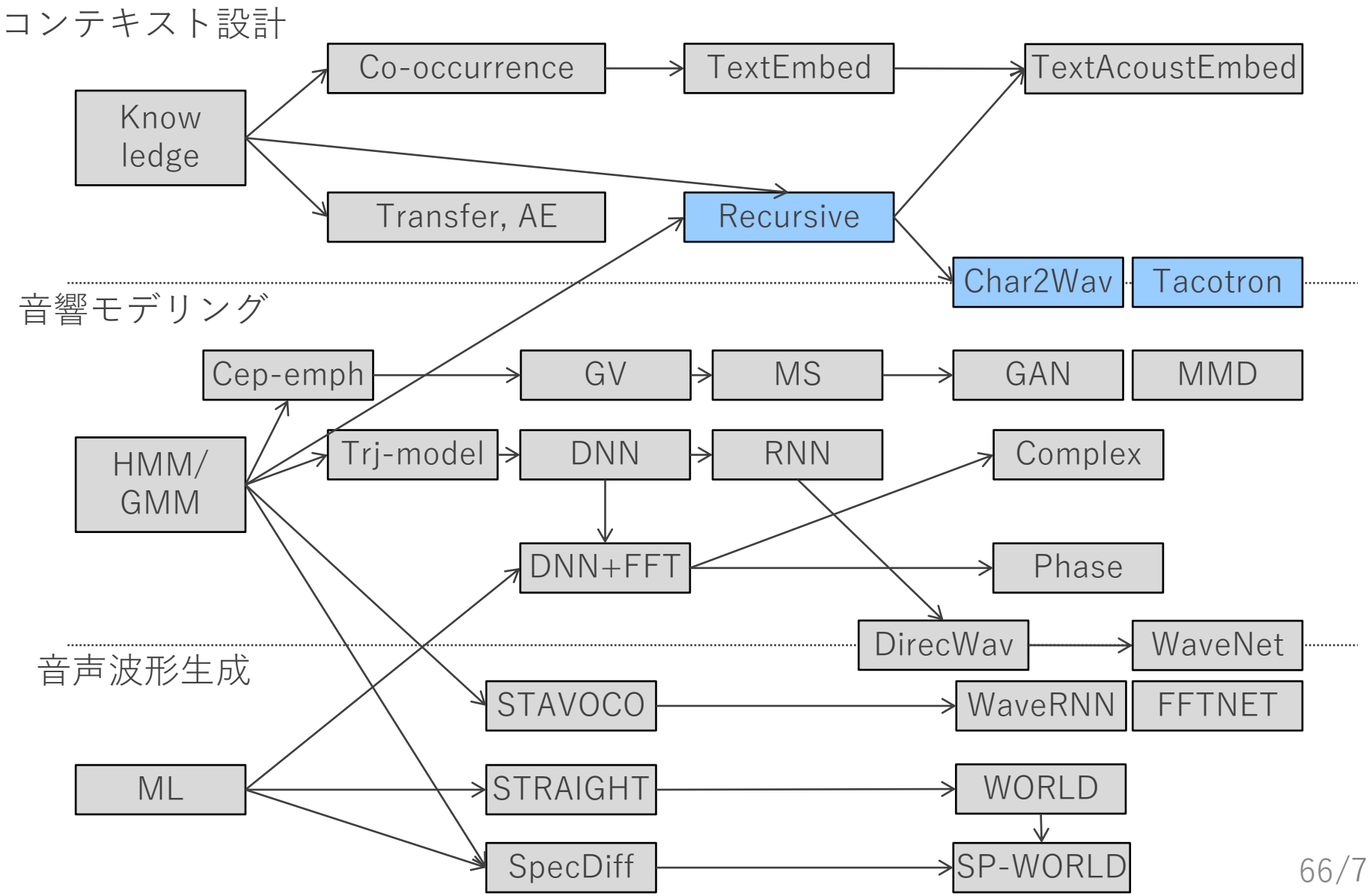
- 信号処理ベースの圧縮 (ケプストラムなど) から機械学習ベースへ
- Stacked AE に基づくスペクトル圧縮
- AEを用いたスペクトル変換 [Takaki16]

それ以降の発展

- What-where auto-encoder に基づくスペクトル圧縮 [Hu17]
- Siamese auto-encoder-based [Hamidreza17]
- Variational auto-encoder (VAE) に基づくスペクトル変換 [Hsu16]
- VQ-VAE に基づく音声変換 [Oord17]

[Takaki16] Takaki et al., “A Deep Auto-encoder based Low-dimensional Feature Extraction from FFT Spectral Envelopes for Statistical Parametric Speech Synthesis,” Proc. ICASSP, 2016.
[Hu17] Hu et al., “Extracting structural spectral features using what-where auto-encoders for statistical parametric speech synthesis,” Proc. ICASSP, 2017.
[Hamidreza17] Hamidreza et al., “Siamese Autoencoders for Speech Style Extraction and Switching Applied to Voice Identification and Conversion,” Proc. INTERSPEECH, 2017.
[Hsu16] Hsu et al., ““Voice conversion from non-parallel corpora using variational auto-encoder,” Proc. APSIPA, 2016.
[Oord17] Oord et al., “Neural discrete representation learning,” Proc. NIPS, 2017.

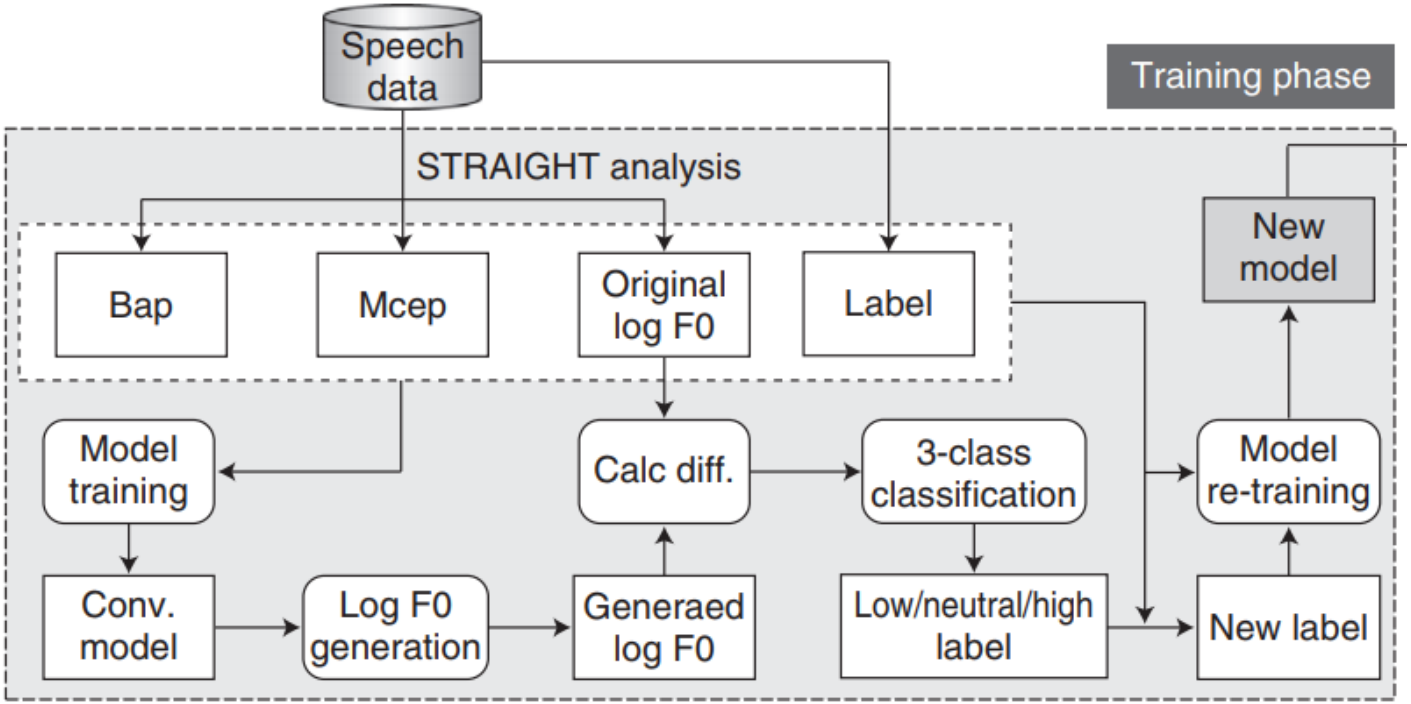
音声合成変換技術の変遷



反復的なラベル推定

既存コンテキストからの差分を用いてラベルを教師なし推定

- この論文では，HMM感情音声合成の韻律生成に適用.
- 直感的に言えば，読み上げ形式のF0からの差分を用いた感情韻律クラスタリング

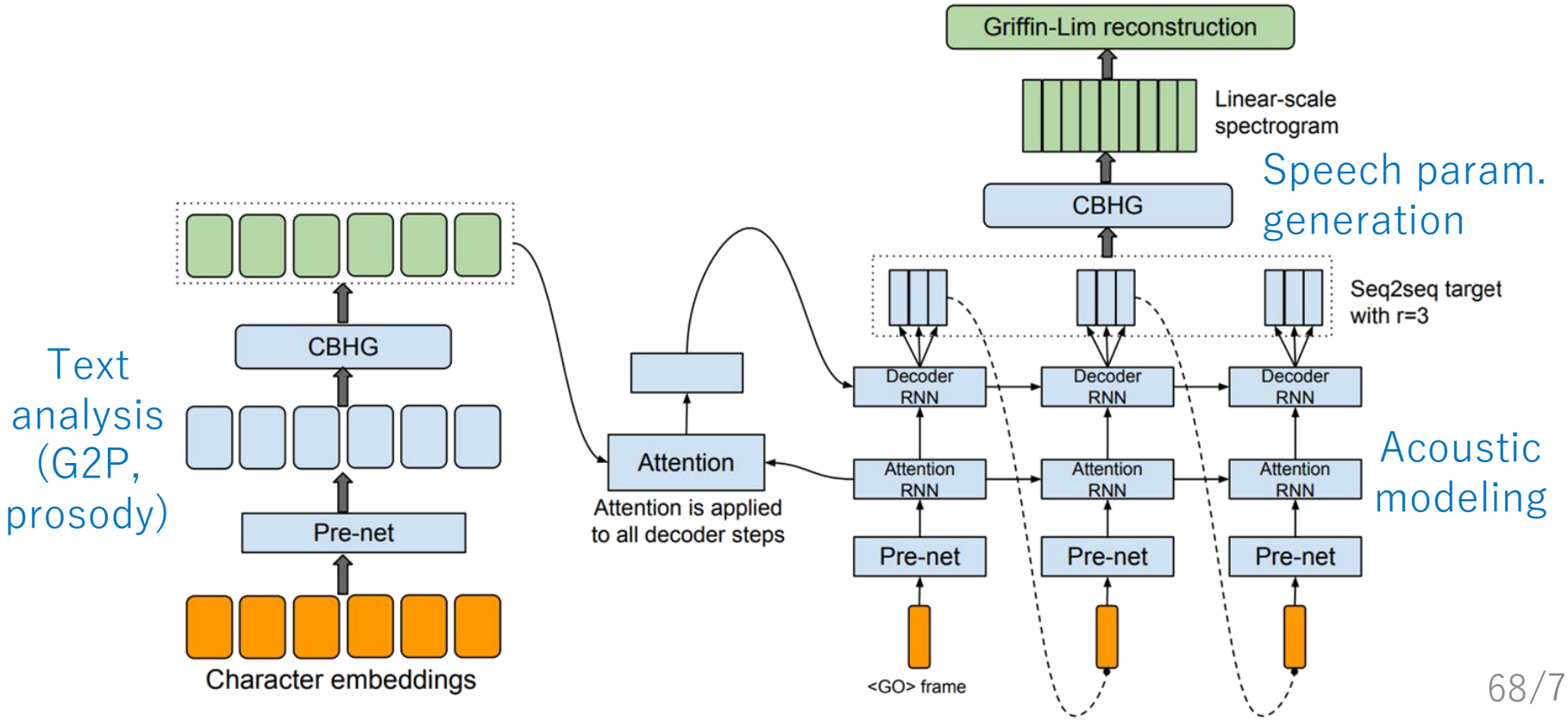


[Maeno14] Maeno et al., “Prosodic Variation Enhancement Using Unsupervised Context Labeling for HMM-based Expressive Speech Synthesis,” Speech Communication, 2014.

Tacotron: towards end-to-end speech synthesis

End-to-End型音声合成に向けたDNN構造

- 音響モデル部にAttention構造を導入（出力側は時間方向に圧縮）
- Interpretability は低下するが，Character embedding で pronunciation や prosody を予測しやすい言語では有効



[Maeno14] Wang et al., “Tacotron: Towards End-to-End Speech Synthesis,” Proc. INTERSPEECH, 2017.

End-to-End関連の関連論文

Tacotron の attention 行列の monotonic さを考慮した高速化

- Monotonicity regularization [Tachibana18]
- Forward attention [Zhang18]

その他のEnd-to-end型（っぽい）音声合成（紹介だけ）

- Char2Wav from MILA [Sotelo17]
- DeepVoice from Baidu [Ping18]
- Tacotron2 from Google [Shen18]

[Tachibana18] Tachibana et al., “Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention,” Proc. ICASSP, 2018.
[Zhang18] Zhang et al., “Forward attention in sequence-to-sequence acoustic modeling for speech synthesis,” Proc. ICASSP, 2018.
[Sotelo17] Sotelo et al., “CHAR2WAV: END-TO-END SPEECH SYNTHESIS,” Proc. ICLR, 2017.
[Ping18] Ping et al., “DEEP VOICE 3: SCALING TEXT-TO-SPEECH WITH CONVOLUTIONAL SEQUENCE LEARNING,” Proc. ICLR, 2018.
[Shen18] Shen et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” arXiv, 2018.

まとめ

まとめ

今日説明したこと

- 近年のコンテキスト設計・音響モデル・波形生成
- それらの統合

説明しなかったこと

- 多言語・話者モデリング，モデル適応，他の音声処理との統合など.
- 音声合成は自然言語処理・音声信号処理・機械学習などの複合技術なので，学ぶことはまだまだ沢山あります.

これからの音声合成

音声合成の役目は，音声を正確に出すこと？

- **答えはNo.**（もちろん，正確に出すことも大事）

音声合成の役目は，音声コミュニケーションを拡張すること

- 音声の芸術性を満たすには？（感性工学？）
- 音声生成・聴取との関連？（物理学？）
- セキュリティとの関連？（セキュリティ工学？）
  - 声の肖像権はどうあるべき？
- 人間を組み込んだ音声合成？（ヒューマンコンピュテーション？）
- IoA（Internet of Ability）としての音声合成？
  - 身体・時空間・文化の多様性を認めつつ，それらを拡張できる？