Statistical Speech Synthesis Integrating Human's Speech Information Processing Abilities

(人間の音声情報処理能力を統合した統計的音声合成)

システム情報学専攻 齋藤 佑樹 (48-187303)
指導教員: 猿渡 洋 教授

予備審査での主なご指摘と修正点

指摘1: 博士論文の題目が適切ではないのでは?

→ 本論文全体のフィロソフィーを明確に述べる題目に変更

指摘2: 音声合成一般の問題点と, 本論文の主眼を述べるべき

→ 説明を追加 (1.1節)

指摘3: GV 補償後の合成音声特徴量の散布図も示すべき

→ 図を修正 (3.3.5節)

指摘4: GAN で最小化される擬距離に関する考察を追加すべき

→ 学習の性質に関する考察を追加 (3.4.10節)

指摘5: 類似度行列埋め込みがうまくいかない理由を述べるべき

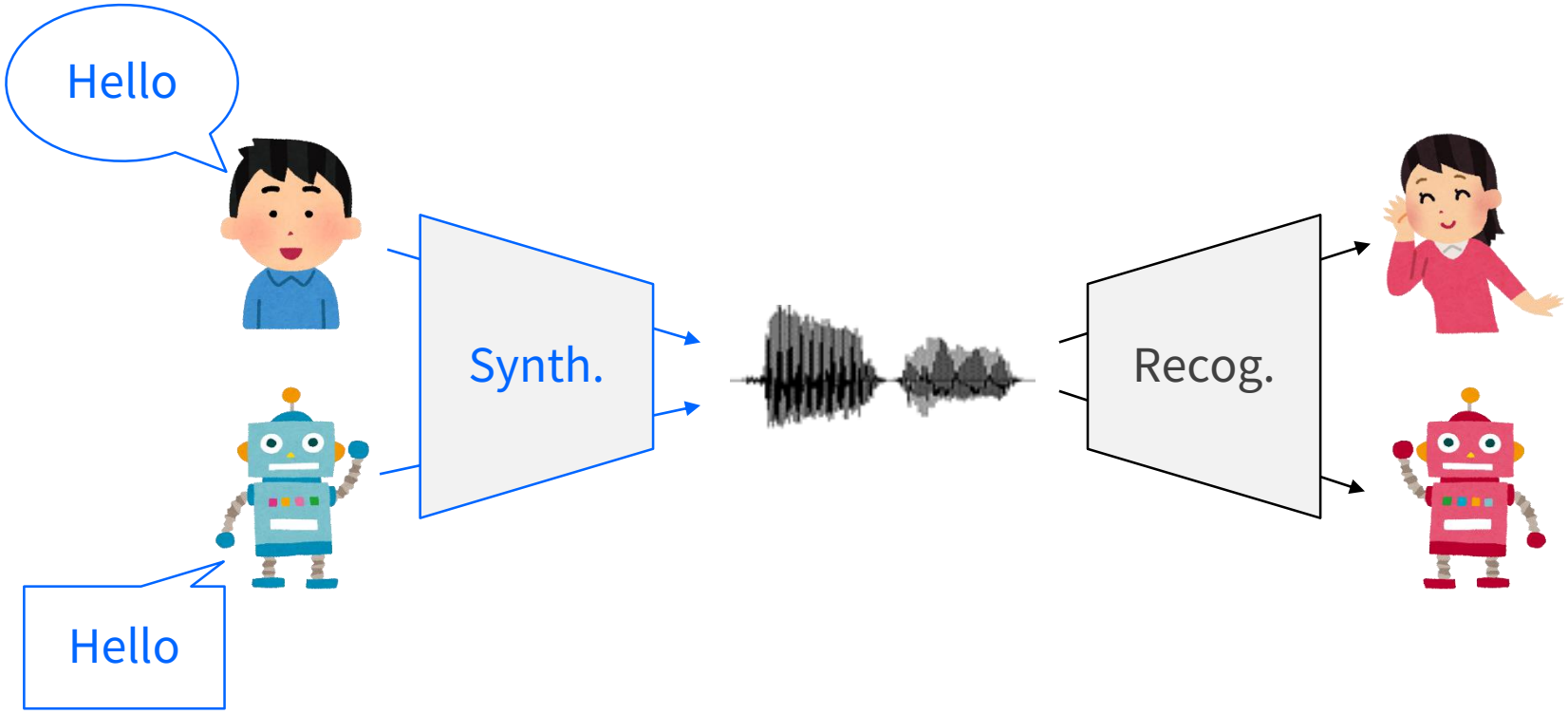
→ 話者空間の可視化に基づく議論・考察を追加 (6.3.5節)

研究背景

コミュニケーションにおける音声メディアの重要性

発話者の意図や感情などの情報を, より自然に受聴者に伝達可能

人間同士だけでなく, 人間とロボットの間での情報伝達も実現

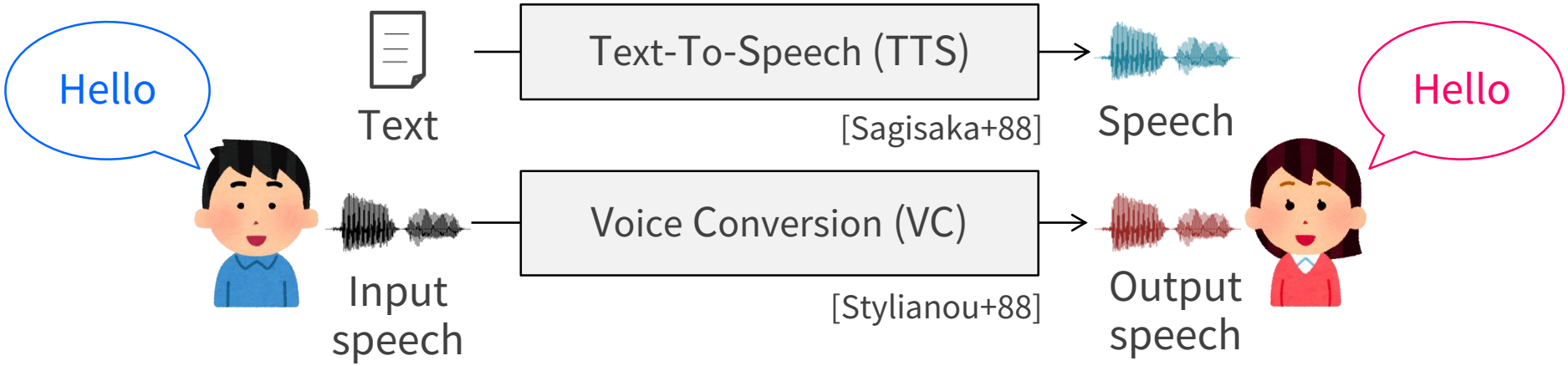


研究分野

研究分野: 音声合成 (speech synthesis)

コンピュータで人間の音声を人工的に生成・変換・加工する技術

物理的/物質的な制約を超えた音声コミュニケーションを実現



本論文の対象: 統計的音声合成 [Zen+09]

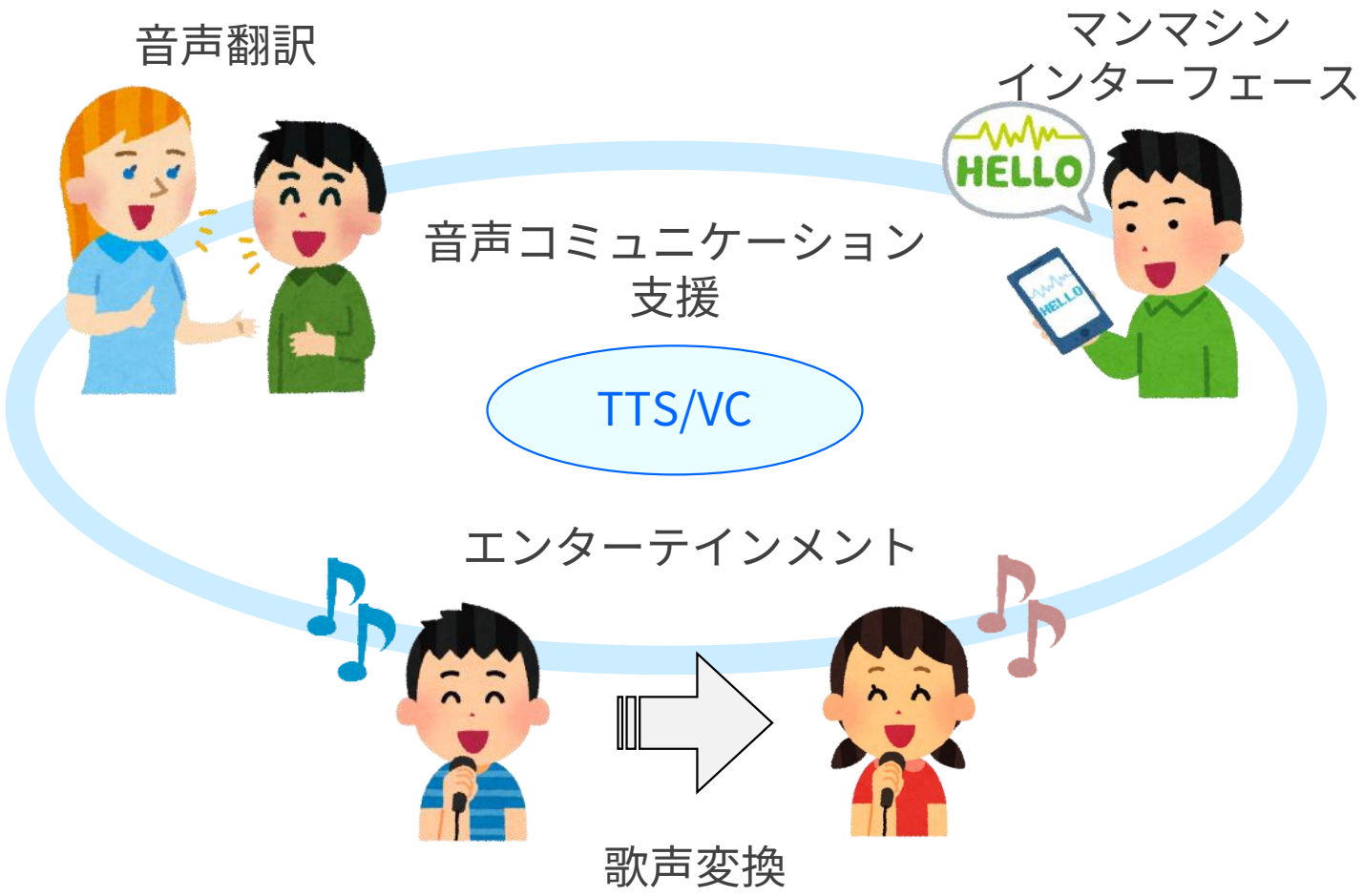
入力 → 音声 の対応関係を統計的音響モデルで学習

近年では, DNN 音声合成 [Zen+13] が主流

従来技術 (HMM-TTS, GMM-VC) よりも高精度な音響モデリングを実現
[Tokuda+13] [Toda+07]

DNN: Deep Neural Network [Hinton+12]

人間社会における音声合成の応用



あらゆる人・モノが, あらゆる声で
自在にコミュニケーション可能な社会を実現

音声合成における解決すべき課題

[修正]

合成音声の高品質化

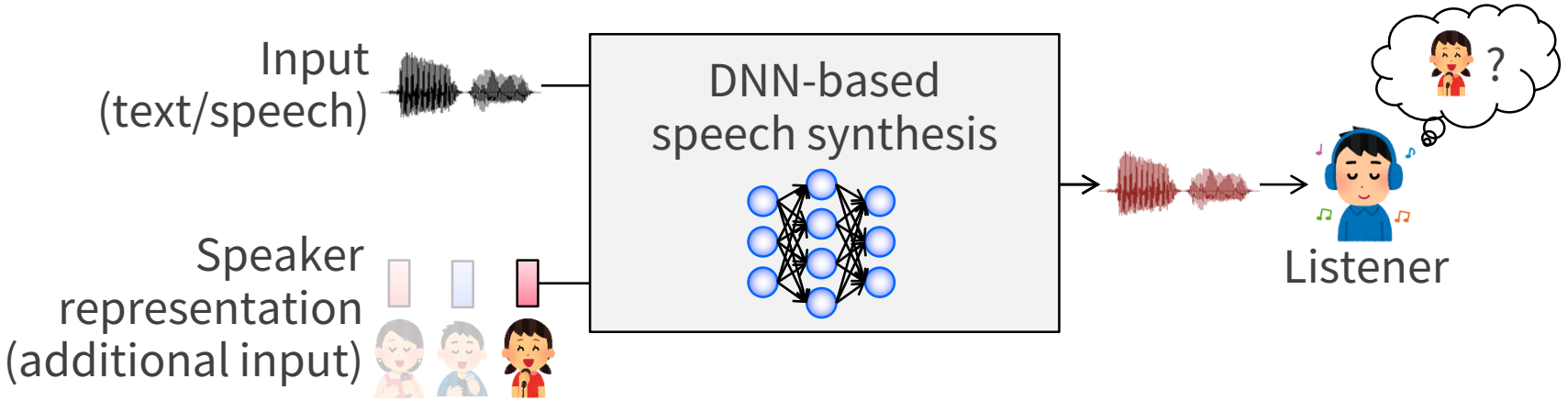
→ 人間の自然音声と合成音声の違いをどう最小化するか?

合成音声の多様化

→ 言語・話者性・感情などが異なる多様な音声をどう生成するか?

合成音声の制御性向上

→ 人間が直感的に制御しやすい音声合成をどう実現するか?



本論文では, 単一言語・多数話者の読み上げ音声の合成/変換を対象

本論文のアプローチ

目的: 人間の音声情報処理能力を統合した音声合成理論の構築

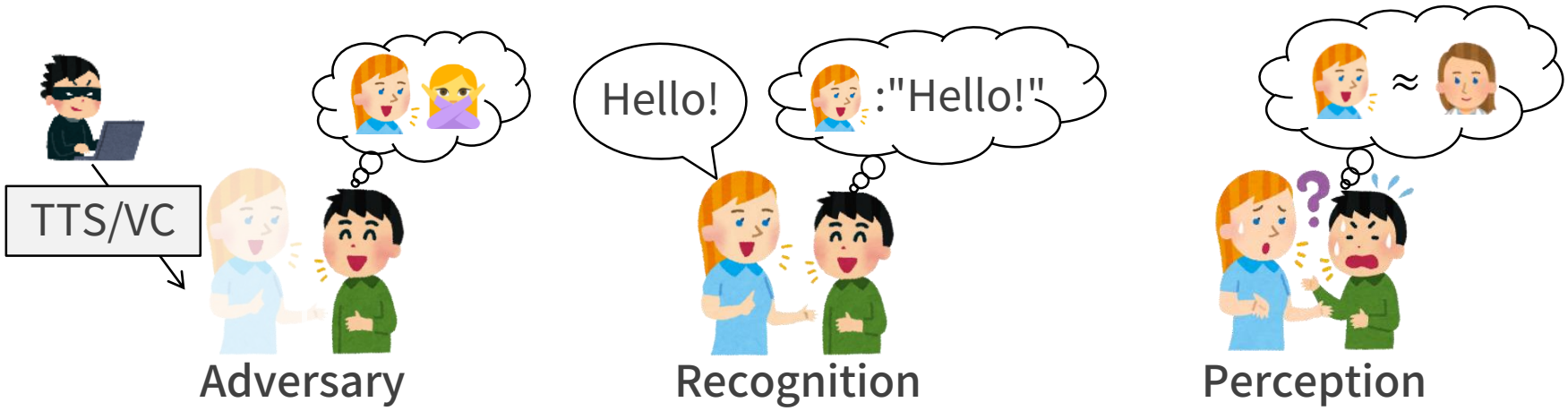
人間からヒントを得て, 音声合成の品質・多様性・制御性を改善

本論文での提案法

音声敵対過程を統合した高品質な音声合成 (第3章 & 第4章)

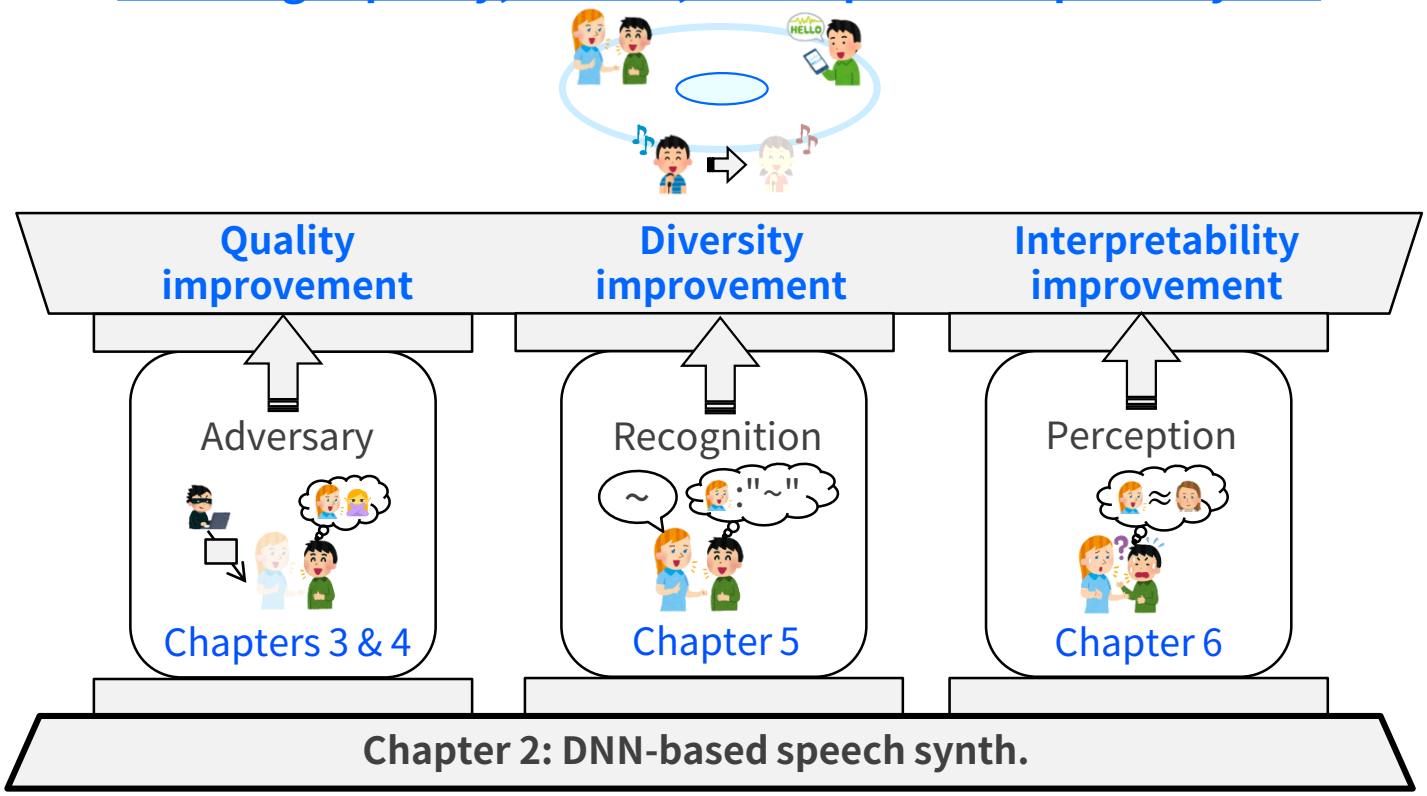
音声認識過程を統合した多様な音声合成 (第5章)

音声知覚過程を考慮した解釈性の高い話者表現学習 (第6章)



第2章: DNN 音声合成の基礎

Goal: High-quality, diverse, & interpretable speech synth.



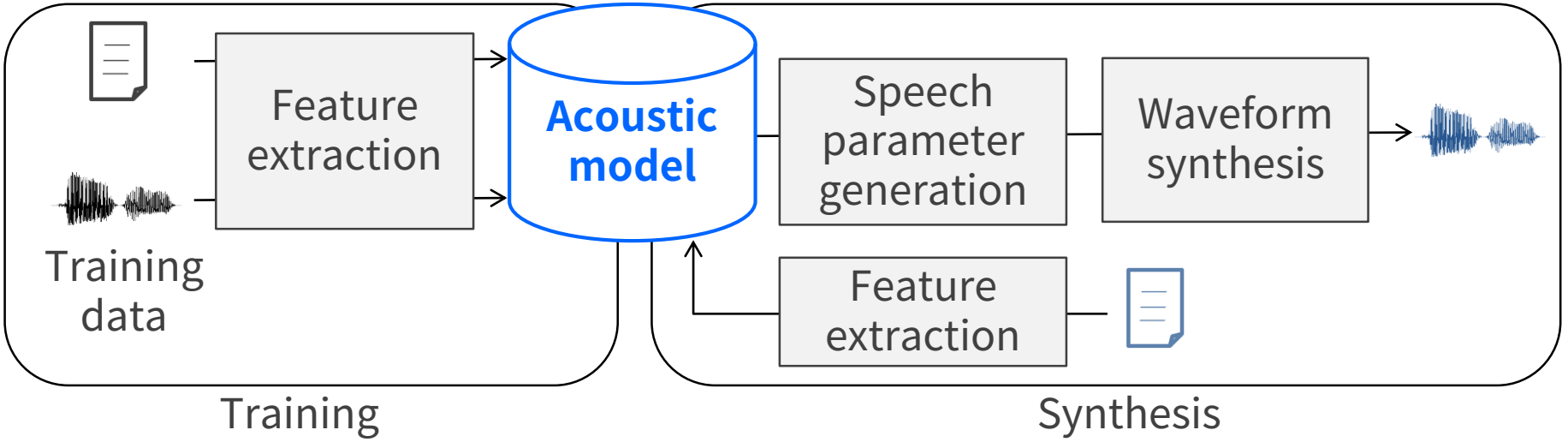
DNN 音声合成の基本的な枠組み (TTS の場合)

学習時

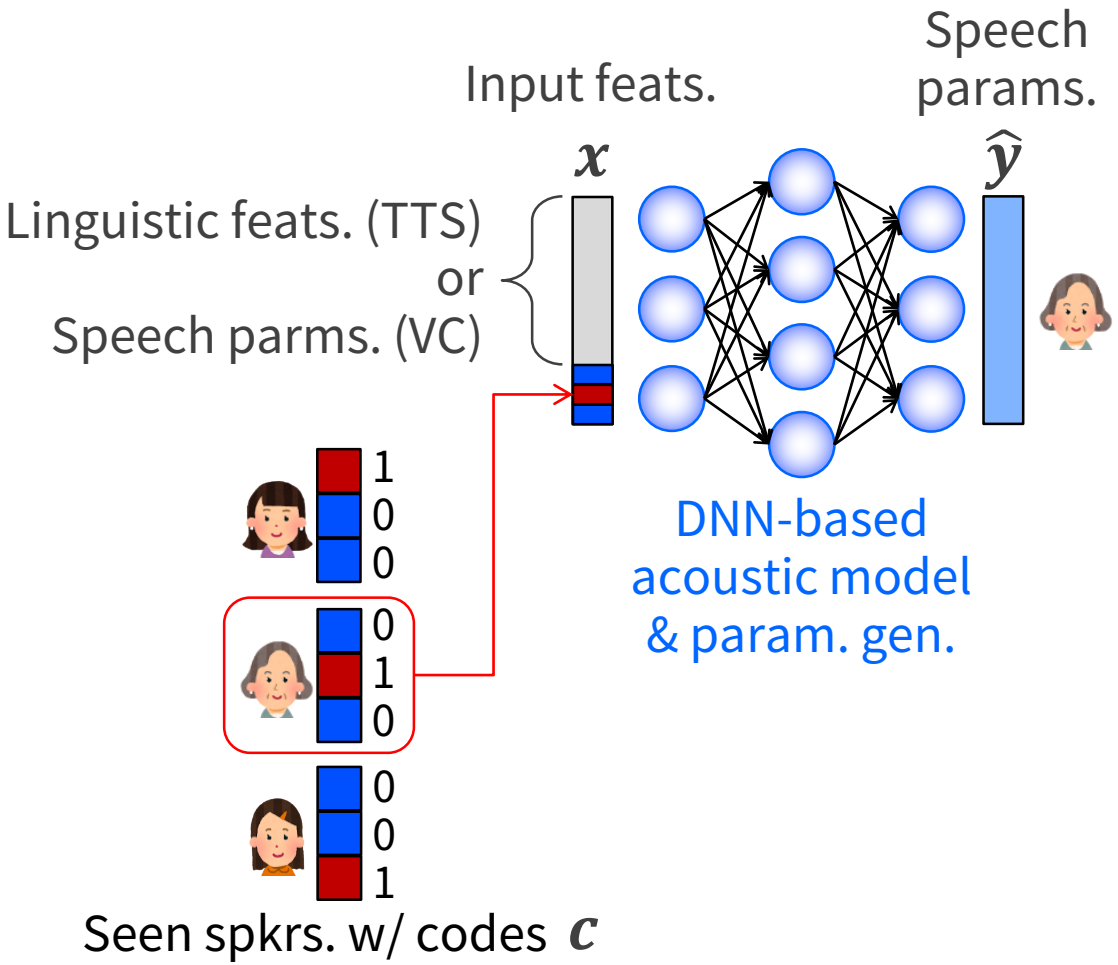
1. テキスト/音声波形から言語特徴量/音声パラメータを抽出
2. 言語特徴量から音声パラメータを生成する DNN 音響モデルを学習

生成時

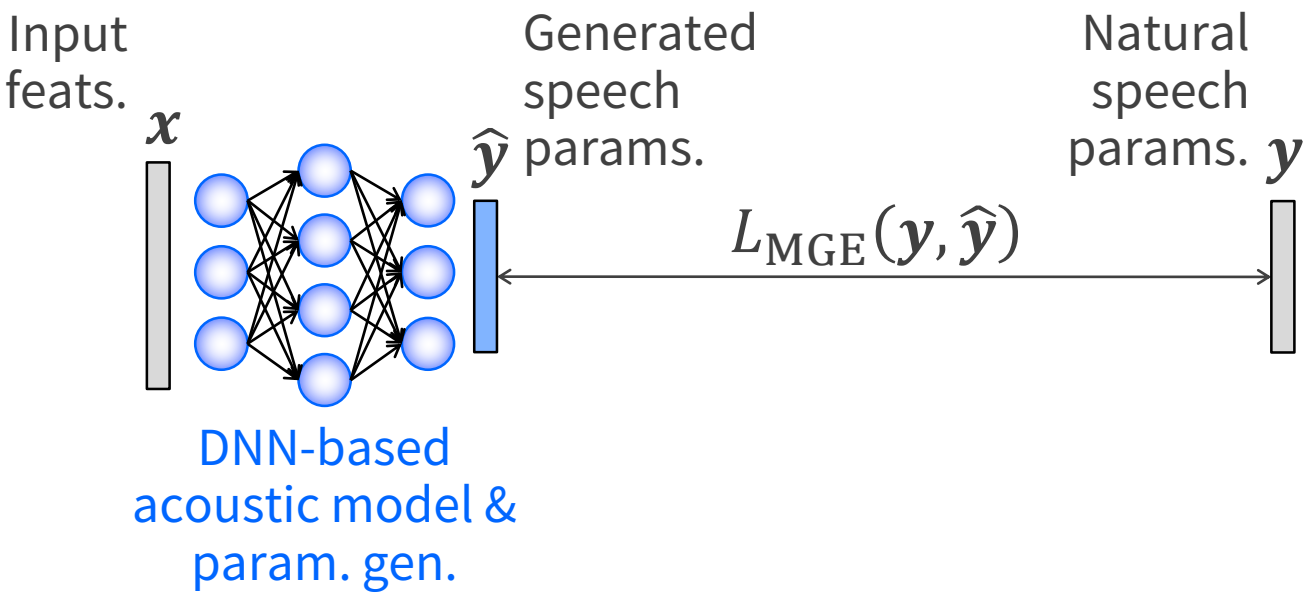
1. テキストから言語特徴量を抽出
2. 学習後の DNN 音響モデルを用いて合成音声パラメータを生成
3. 合成音声パラメータから合成音声波形を生成



話者コードを用いた DNN ベース多話者音響モデリング [Hojo+18]



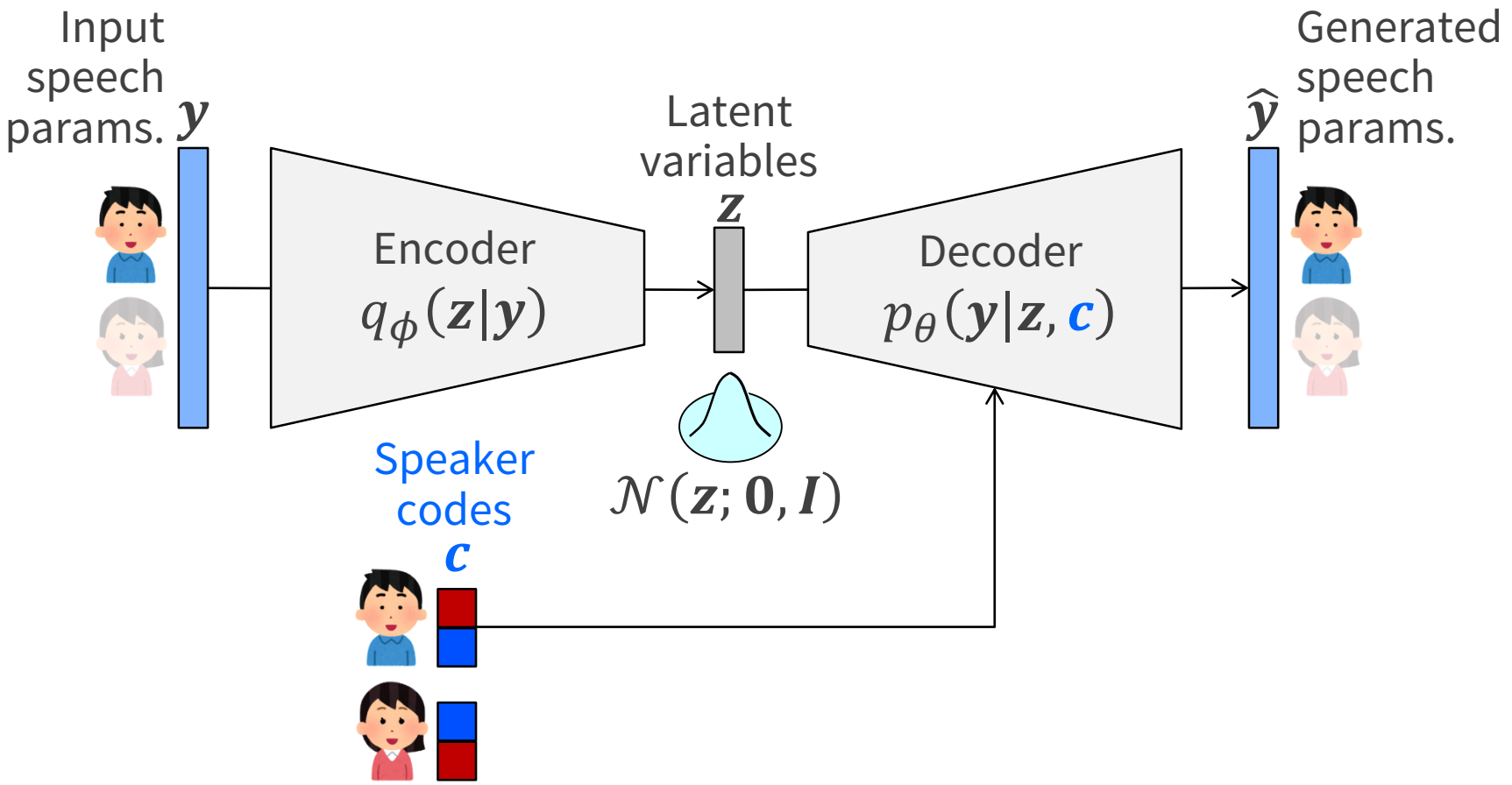
従来の音響モデル学習法: MGE 学習 [Wu+16]



$$L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\top}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \rightarrow \text{Minimize}$$

従来の学習法 = 音声合成の損失 (自乗誤差) のみを最小化

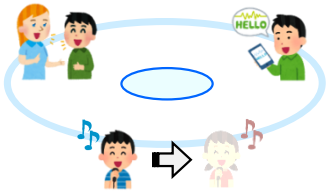
話者コードを用いた VAE ベース多話者音響モデリング [Hsu+16]

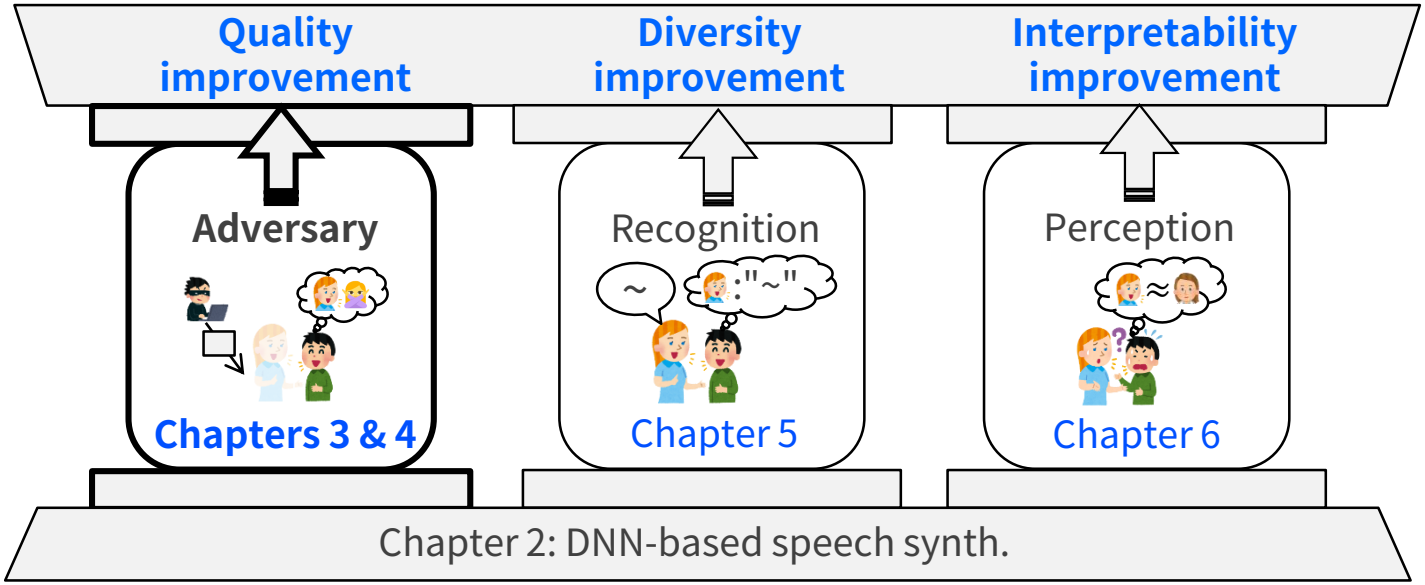


$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{y},\boldsymbol{c}) = \underbrace{-D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}) \,||\, \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})\right)}_{\text{潜在変数 } \boldsymbol{z} \text{ に対する正則化項}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c})]}_{\text{音声パラメータ } \boldsymbol{y} \text{ の再構築誤差}}$$

第3章 & 第4章:
音声敵対過程を統合した高品質な音声合成法

Goal: High-quality, diverse, & interpretable speech synth.





第3章の概要

本章の目的: ボコーダベース音声合成の品質改善

合成音声パラメータの過剰な平滑化への対処が必要

従来法: 過剰な平滑化の度合いを定量化する特徴量の補償

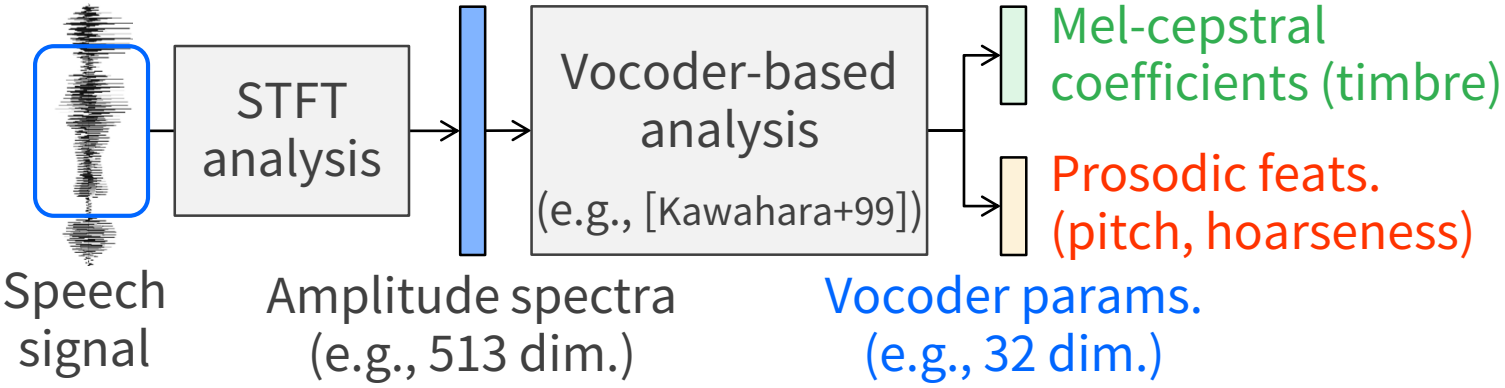
例) 音声パラメータ系列の分布の2次モーメント (GV)

→ 2次モーメント以外の統計的な違いは補償不可 (音質改善に限界あり)

提案法: 自然/合成音声パラメータの分布間差異の補償

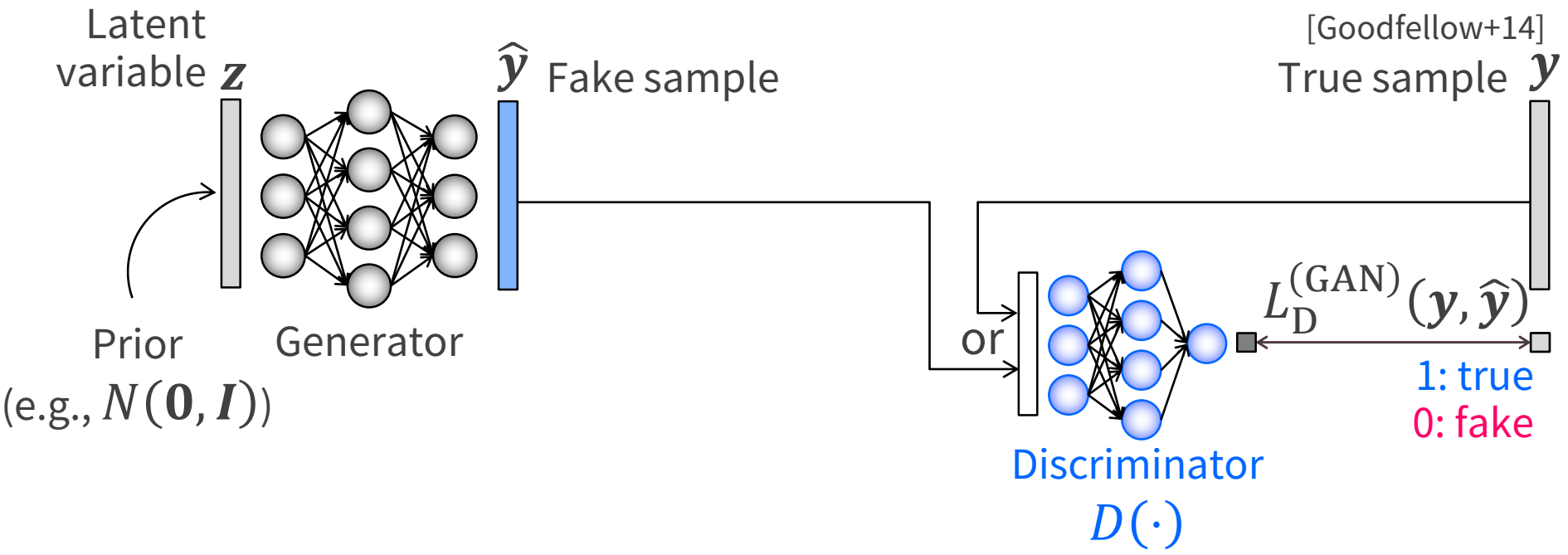
人間の「自然/合成音声を識別する能力」を模擬 & 統合した学習法

GAN の枠組みを導入し, 自然/合成音声パラメータの擬距離を最小化



GV: Global Variance [Toda+07], GAN: Generative Adversarial Network [Goodfellow+14]

GAN の学習: discriminator の更新



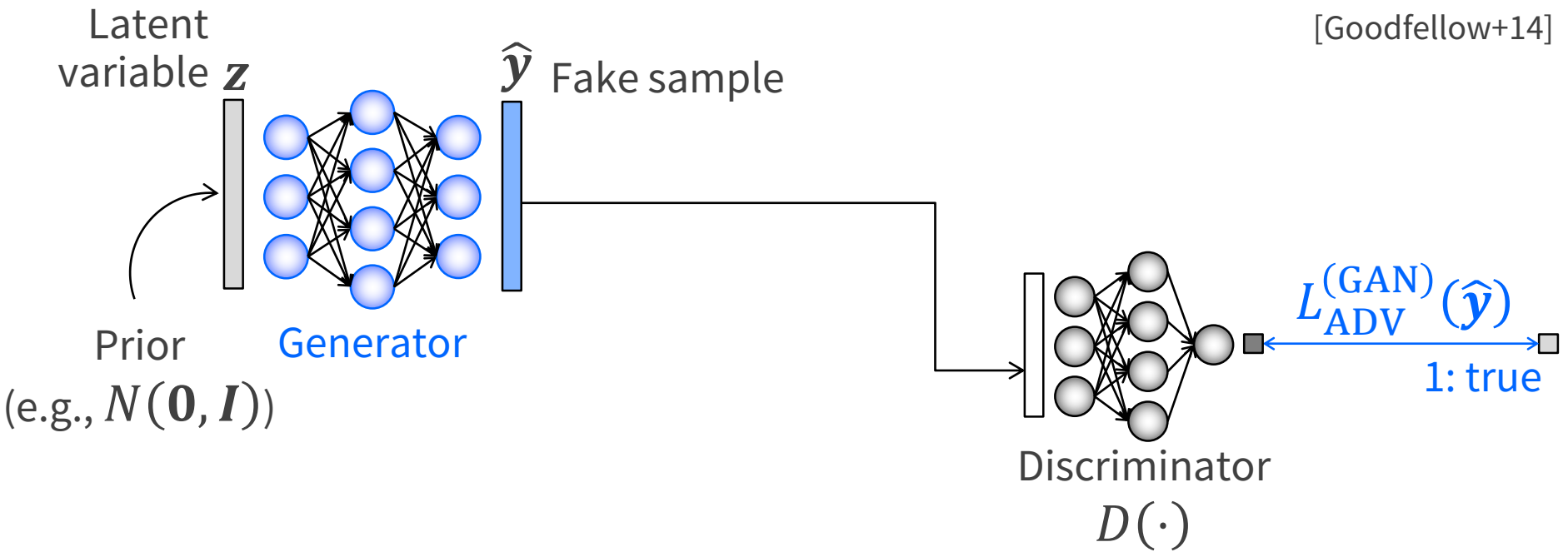
真のサンプル y を
真のサンプルとして識別

偽のサンプル $\hat{y}$ を
偽のサンプルとして識別

$$L_{\mathrm{D}}^{(\mathrm{GAN})}(y, \hat{y}) = L_{\mathrm{D},1}^{(\mathrm{GAN})}(y) + L_{\mathrm{D},0}^{(\mathrm{GAN})}(\hat{y}) \rightarrow \text{Minimize}$$

$L_{\mathrm{D}}^{(\mathrm{GAN})}(y, \hat{y})$ は, 2値分類の cross-entropy と等価

GAN の学習: generator の更新



[Goodfellow+14]

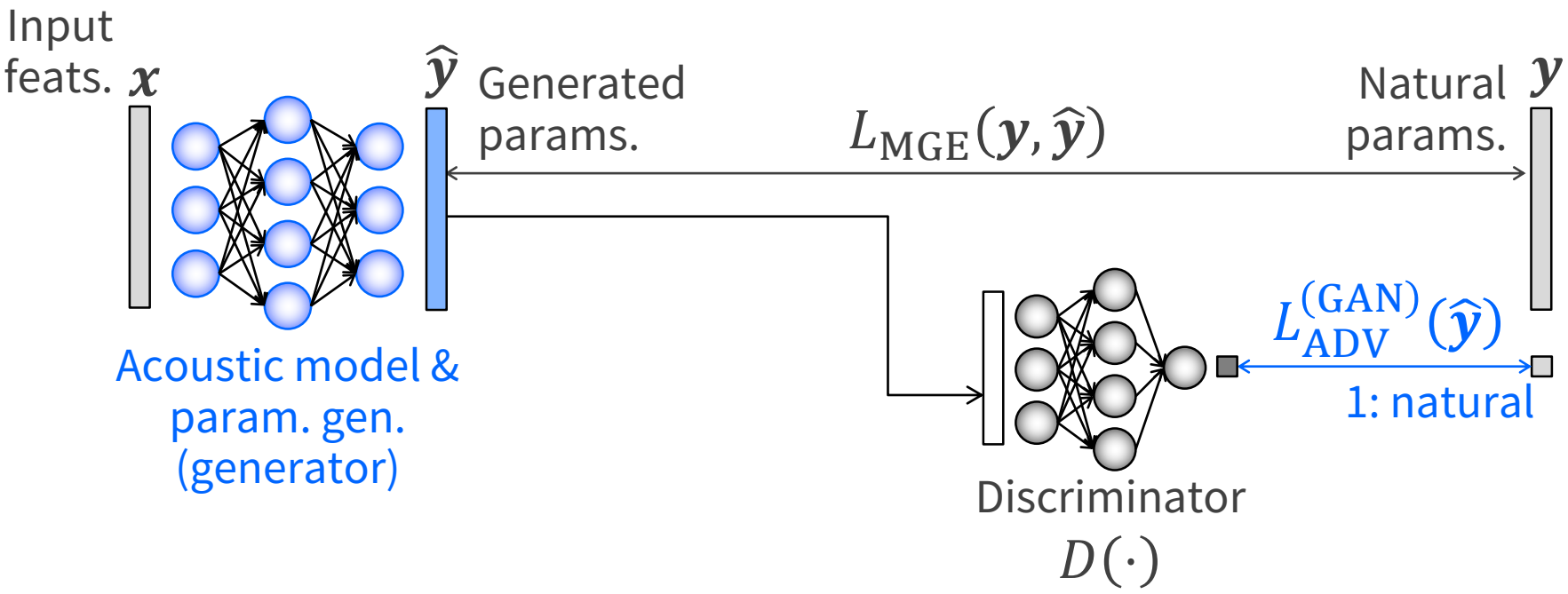
偽のサンプル $\hat{y}$ を
真のサンプルとして識別

$$L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{y}) = L_{\mathrm{D},1}^{(\mathrm{GAN})}(\hat{y}) \rightarrow \text{Minimize}$$

y と $\hat{y}$ の分布間の (近似された) JS 擬距離最小化

JS: Jensen-Shannon

提案法:
音声敵対過程を統合した GAN ベース音響モデル学習



合成音声パラメータ $\hat{\boldsymbol{y}}$ を
自然音声パラメータとして識別

$$L_{\mathrm{G}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \rightarrow \text{Minimize}$$

ω_{D}: 第2項の重み, E_{L_*}: L_*の期待値 (損失の adaptive scaling)

GAN ベース音響モデル学習に関する考察

パラメトリックな統計的差異を最小化する学習の一般化

提案法は, GV などを明示的に考慮せずに**分布間の擬距離を最小化**

本論文では, GAN で最小化される擬距離の影響も調査

MGE 最小化と GAN のマルチタスク学習

GAN の敵対損失を正則化項として音声合成に導入した世界初の試み

Anti-spoofing (音声なりすまし検出) と敵対する音声合成の学習

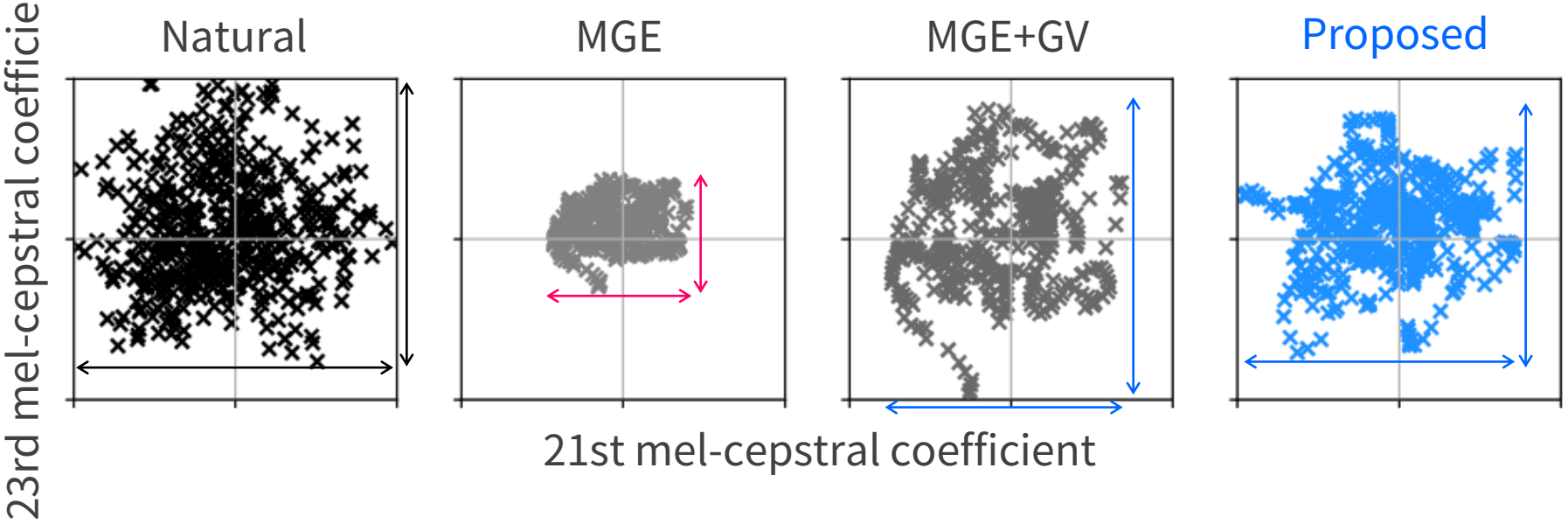
敵対技術の知見を活用し, より効率的な音響モデル学習も実現可能

例) 逆メルスケールで伸縮したスペクトル [Sahidullah+15] の補償 → 第4章

$$L_{\mathrm{G}}^{(\mathrm{GAN})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) \rightarrow \text{Minimize}$$

自然/合成音声パラメータの分布間差異補償

[修正]



GAN の導入により, 自然/合成音声パラメータの分布の違いを補償！

実験条件
(GAN ベースの音響モデル学習を用いた TTS)

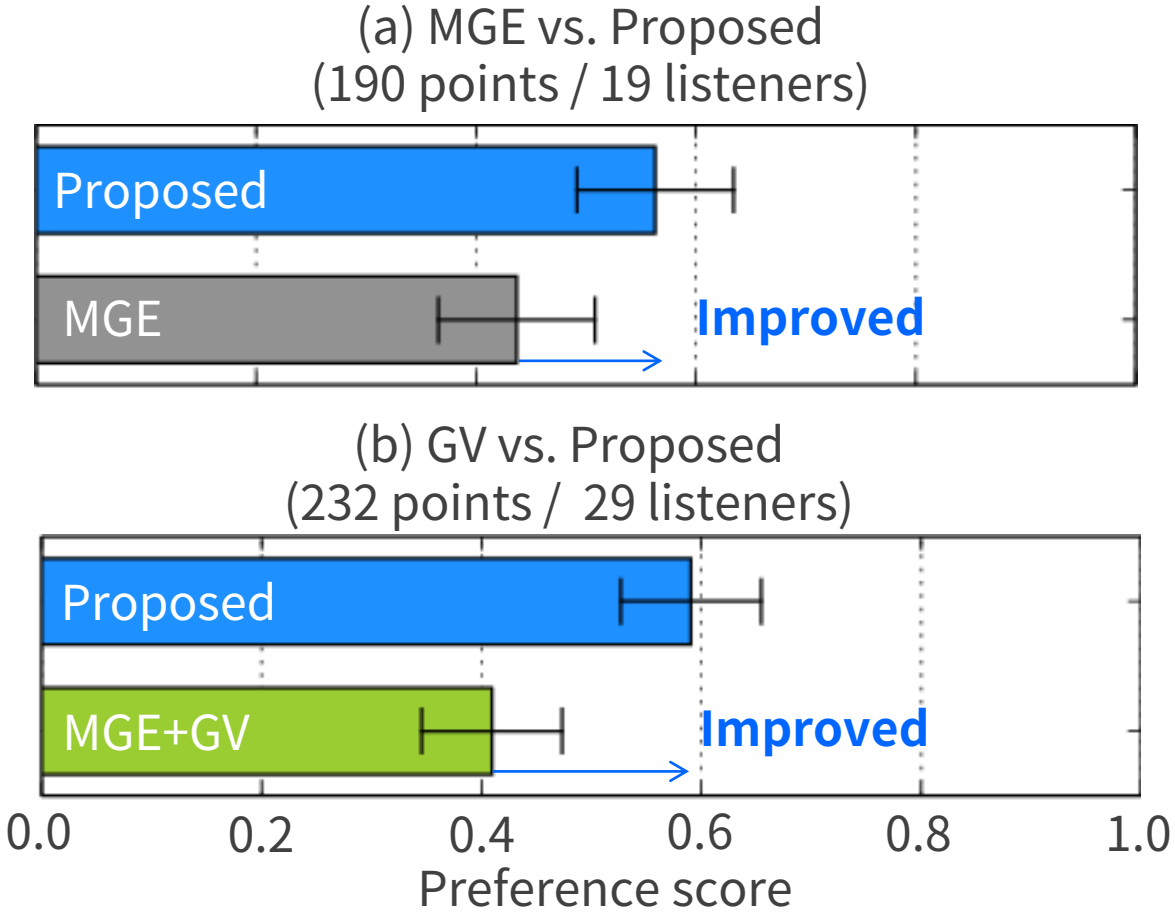
Q. GAN に基づく音響モデル学習は, 合成音声の音質を改善するか？

学習/評価データ	男性話者1名 450文/53文 (16 kHz sampling)
言語特徴量	442次元ベクトル (音素, アクセント, etc...)
音声パラメータ	24次メルケプストラム, F0, 非周期性指標
DNN アーキテクチャ	すべて Feed-Forward 型ネットワーク
提案法の重み ω_{D}	**1.0 (3.4.2節 & 3.4.4節の結果を参考に設定)**
比較手法	**MGE** [Wu+16], **MGE+GV** [Toda+07], **Proposed**

VC でも同様な実験を実施し, 提案法の有効性を確認 (3.5節)

合成音声の音質に関する主観評価 (ボコーダベース TTS)

プリファレンス AB テスト: 音質が良いと感じたサンプルを選択



MGE
(over-smoothed)

MGE+GV
(over-emphasized)

Proposed
(natural)

提案法により, 合成音声の音質が有意に改善！

エラーバーは95%信頼区間

学習時に最小化される擬距離の影響

[修正]

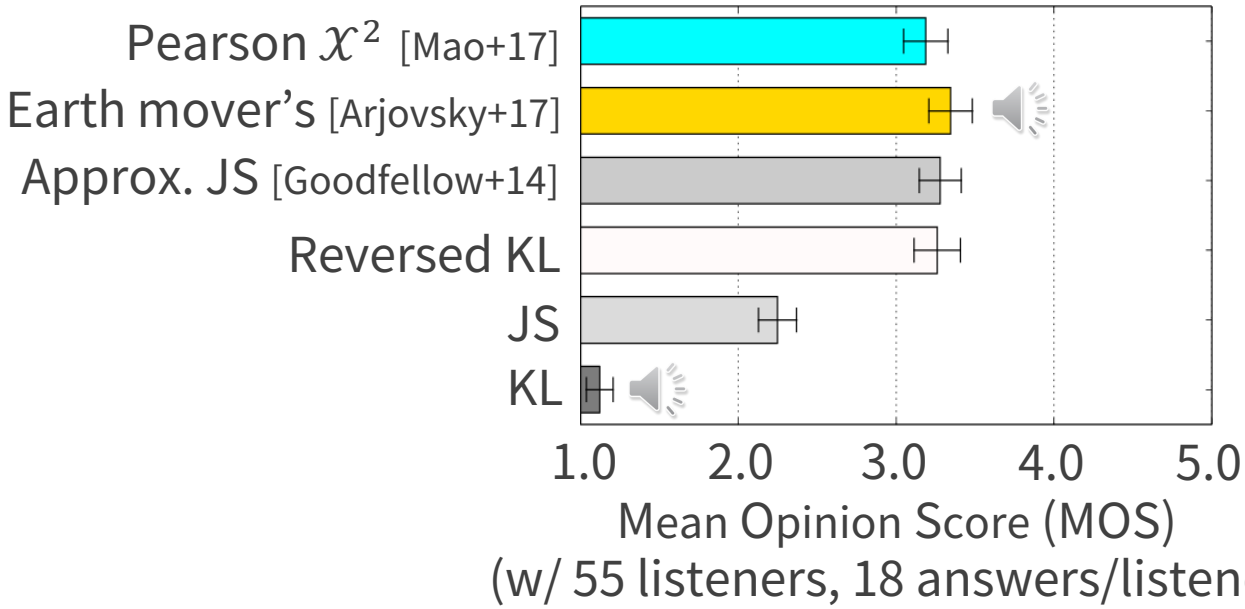
GAN で最小化する擬距離は, 音質に大きく影響

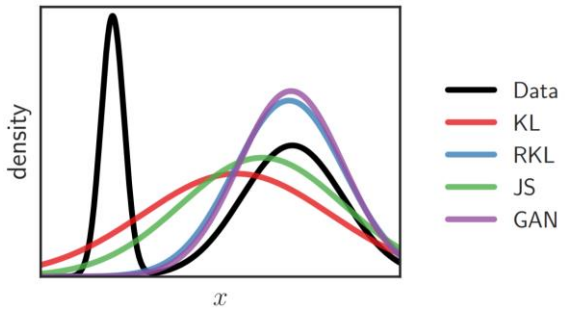
JS/KL 擬距離の最小化は, 音質が有意に劣化

→ Mode-covering ベースの学習は音声合成において不適切?

Earth mover's 距離の最小化が最も高い MOS を達成

→ GAN 学習の安定化は, 画像生成だけでなく音声合成においても有効





右図は [Poole+16] Figure 1 (b) より引用

第4章の概要

本章の目的: ボコーダフリー音声合成の品質改善

第3章とは異なり, どのような特徴量の補償が有効かの検討も必要

課題: 高次元で複雑な振幅スペクトルの統計モデリングの難しさ

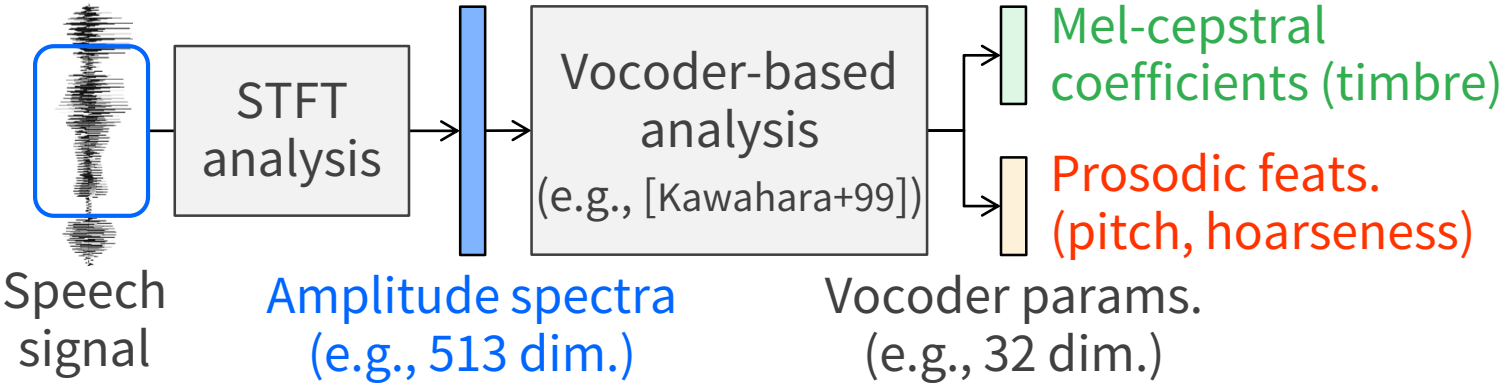
音韻・音高に関する情報を両方含む複雑な特徴量分布

提案法: 周波数伸縮/解像度圧縮を導入した音響モデル学習

Discriminator に着目させる特徴を明示的に制御して分布間差異を補償

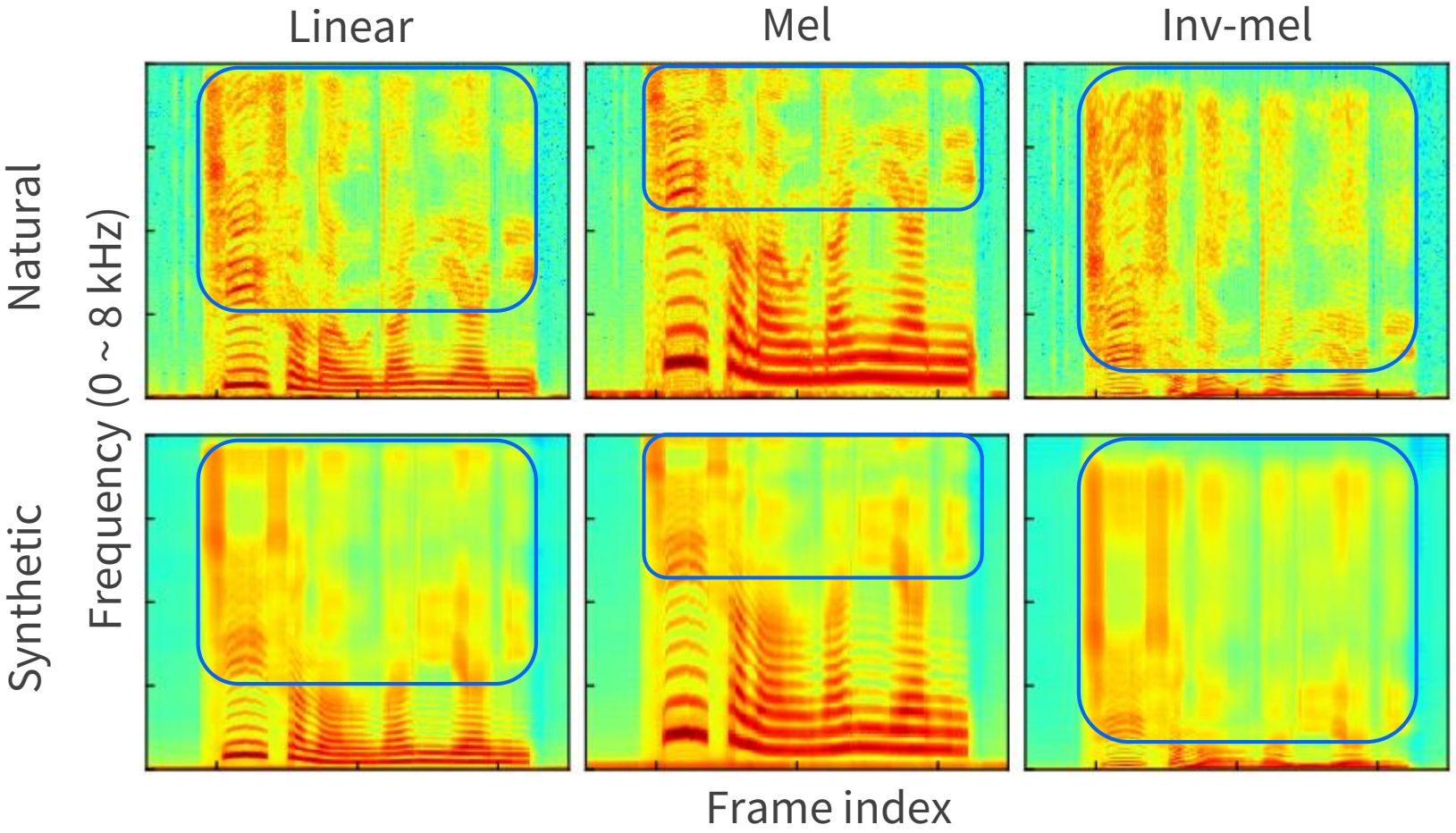
周波数伸縮 → スペクトルのどの帯域を重視するか

解像度圧縮 → スペクトルの包絡成分をどれだけ重視するか



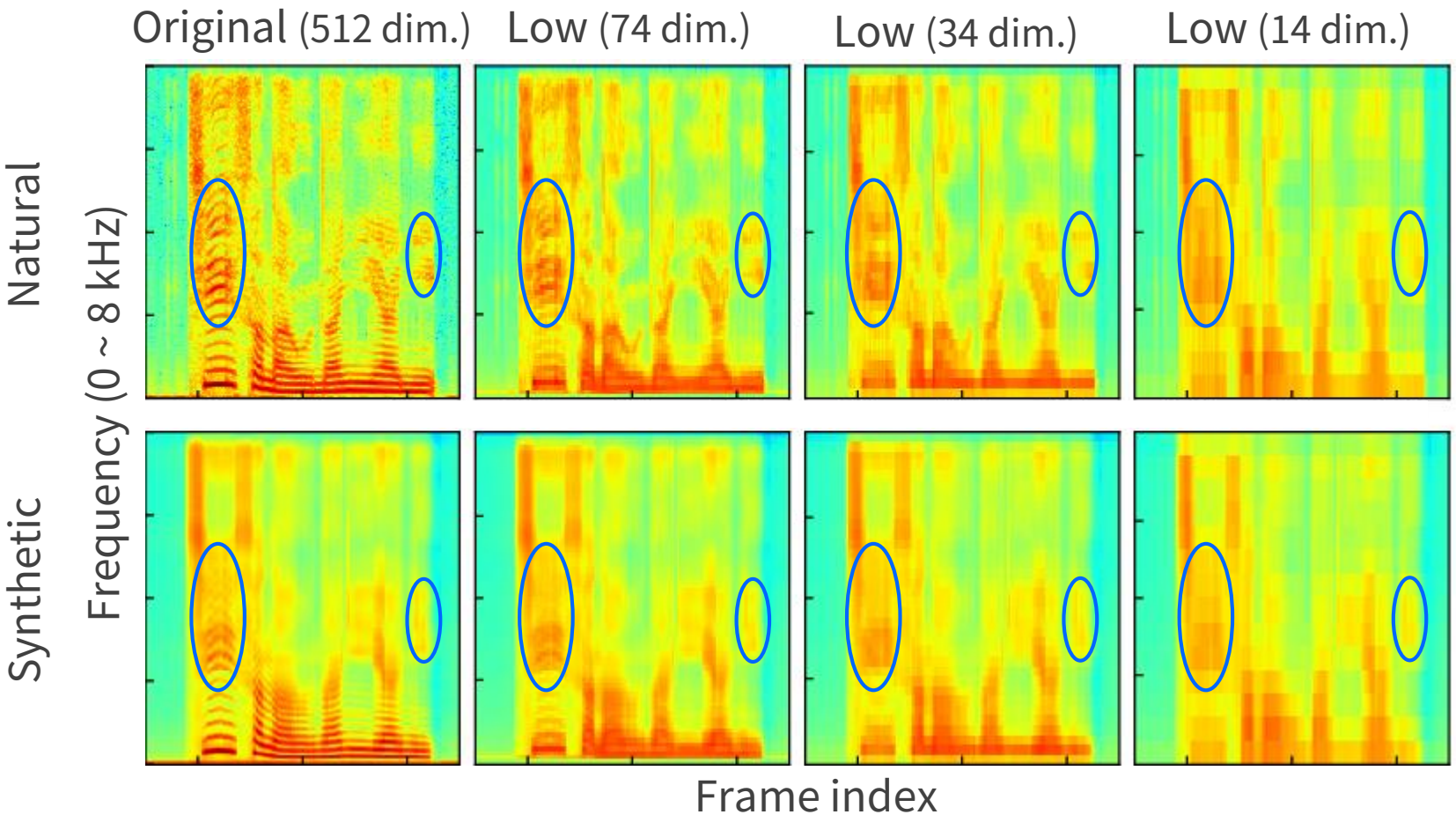
異なる周波数スケールにおける自然/合成音声スペクトルの違い

自然/合成音声スペクトルの違いは, 高域において顕著

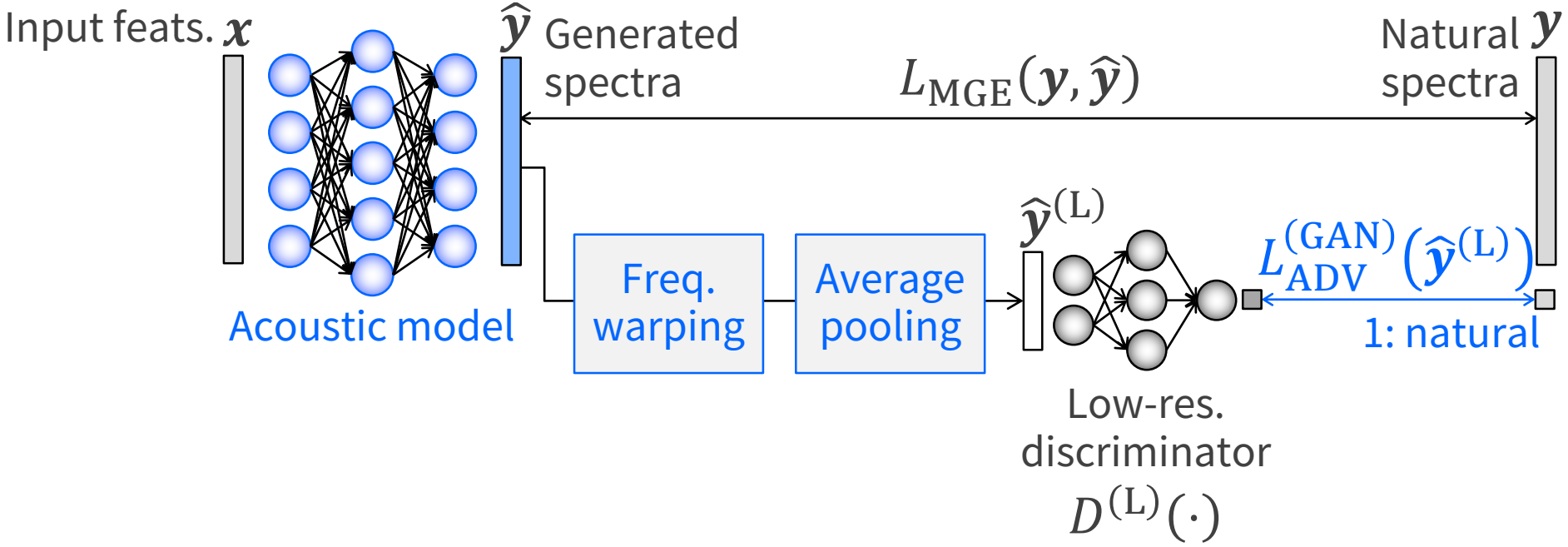


異なる周波数解像度における自然/合成音声スペクトルの違い

自然/合成音声スペクトルの違いは, 包絡成分のピークで顕著



提案法: 周波数伸縮/解像度圧縮を導入した GAN ベース音響モデル学習



$$L_{\mathrm{G}}^{(\mathrm{L})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}^{(\mathrm{L})} \frac{E_{L_{\mathrm{MGE}}}}{E_{L_{\mathrm{ADV}}}} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}^{(\mathrm{L})}) \rightarrow \text{Minimize}$$

合成音声の音質に関するプリファレンス AB テスト (ボコーダフリー TTS)

周波数解像度を低くしてから GAN を適用すると有意に音質改善

	Method A	A vs. B	Method B	
提案法	Original	0.180 vs. **0.820**	Baseline	従来法 [Takaki+17]
	Low	**0.608** vs. 0.392	Baseline	
	Original+Low	0.252 vs. **0.748**	Baseline	

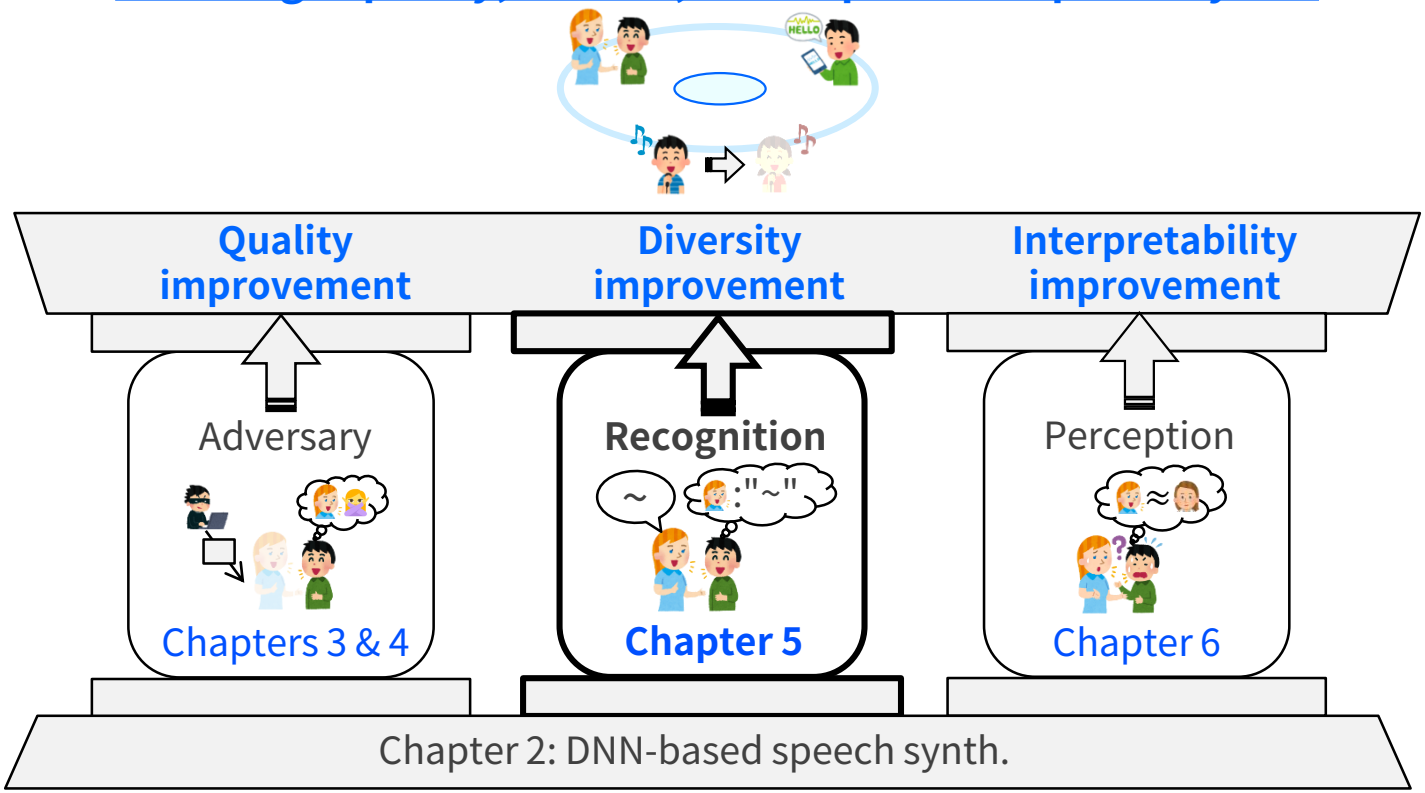
逆メルスケールで周波数伸縮した "Low" → さらに音質改善

Method A	A vs. B	Method B
Linear	**0.752** vs. 0.248	Mel
Mel	0.272 vs. **0.728**	Inv-mel
Inv-mel	**0.576** vs. 0.424	Linear

(w/ 25 listeners & 10 answers / listener. 表中の太字は, $p < 0.05$ で有意差あり)

第5章:
音声認識過程を統合した多様な音声合成法

Goal: High-quality, diverse, & interpretable speech synth.



第5章の概要

本章の目的: 音声を合成可能な話者の多様性向上

学習に用いられていない未知話者の音声も合成可能

課題: 音韻と話者性の分離の難しさ & 合成可能な話者数の制約

従来技術: 合成音声の品質劣化 & 未知話者の音声は合成不可能

提案法: 音声認識/話者認識の中間表現を用いた音響モデリング

人間の「音声の発話内容/話者を正確に認識する」能力を模擬 & 統合

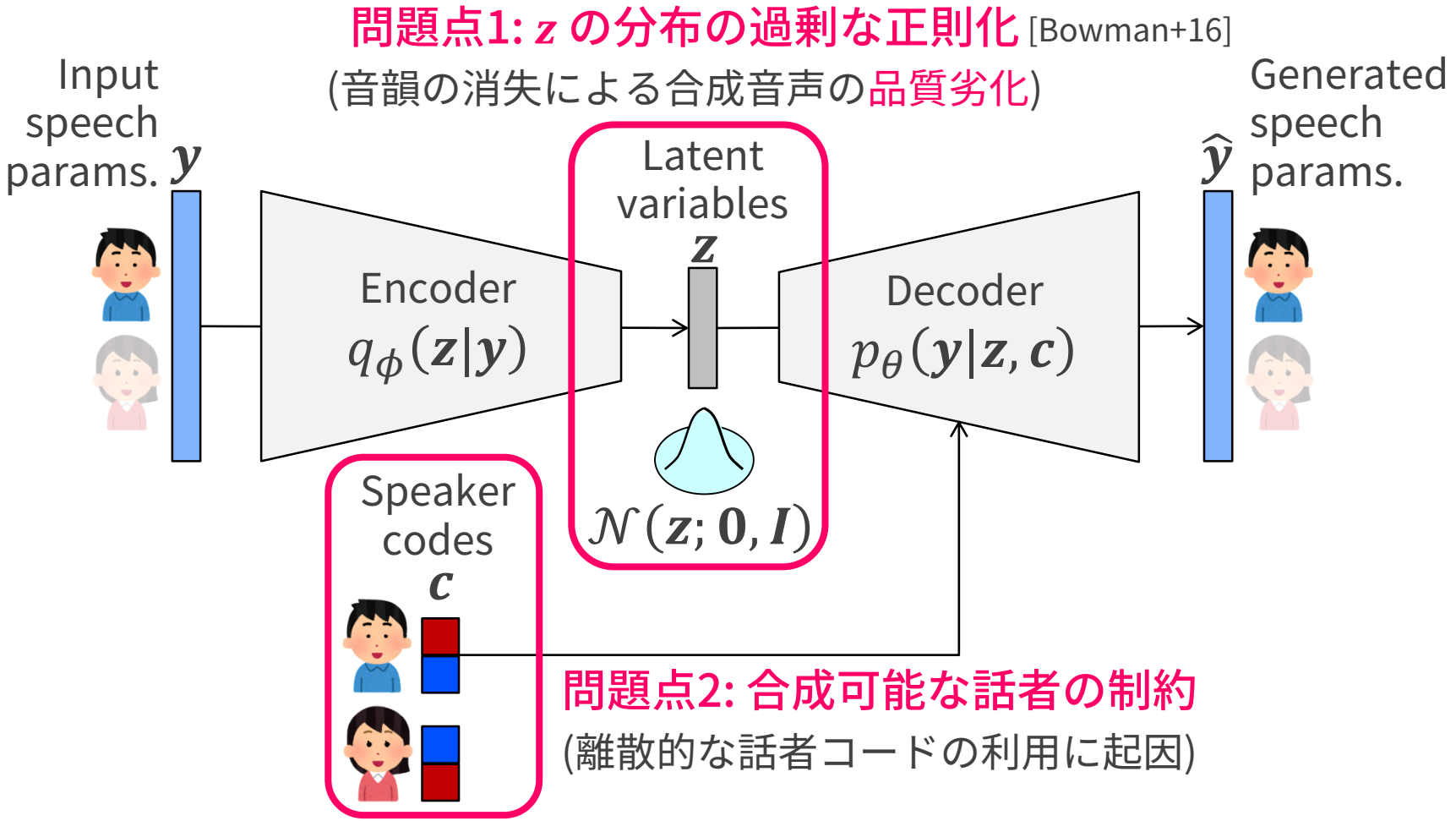
多話者音響モデリングに, 事前学習済みの音声/話者認識モデルを利用

音声認識モデル → 音韻に関する中間表現を獲得

話者認識モデル → 話者性に関する中間表現を獲得

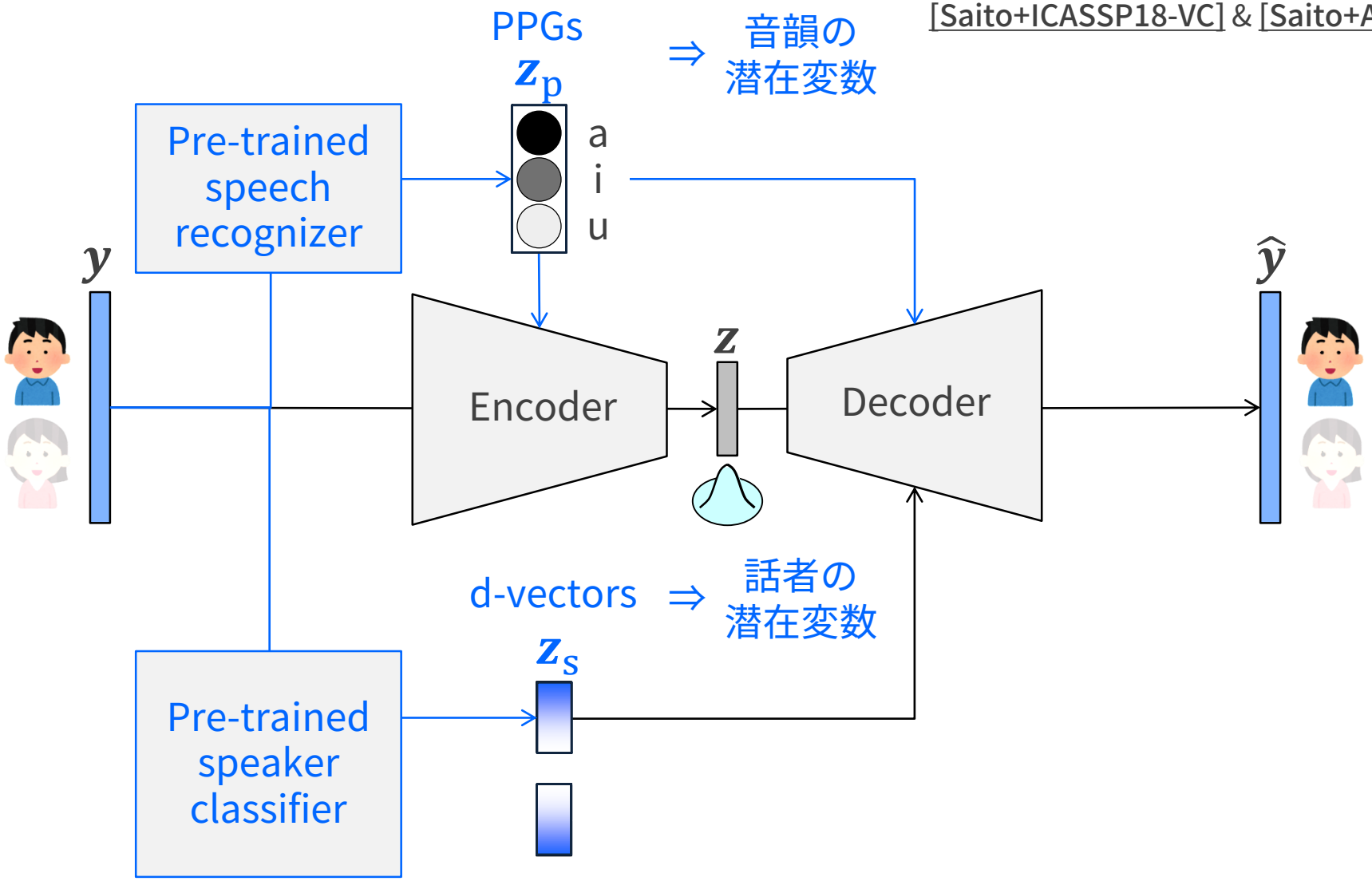
任意話者の音声を, 別の任意話者のものに変換可能な枠組みを提案

従来の VAE ベース多話者音響モデリングの問題点



提案法: 音声認識/話者認識過程を統合した VAE ベース多話者音響モデリング

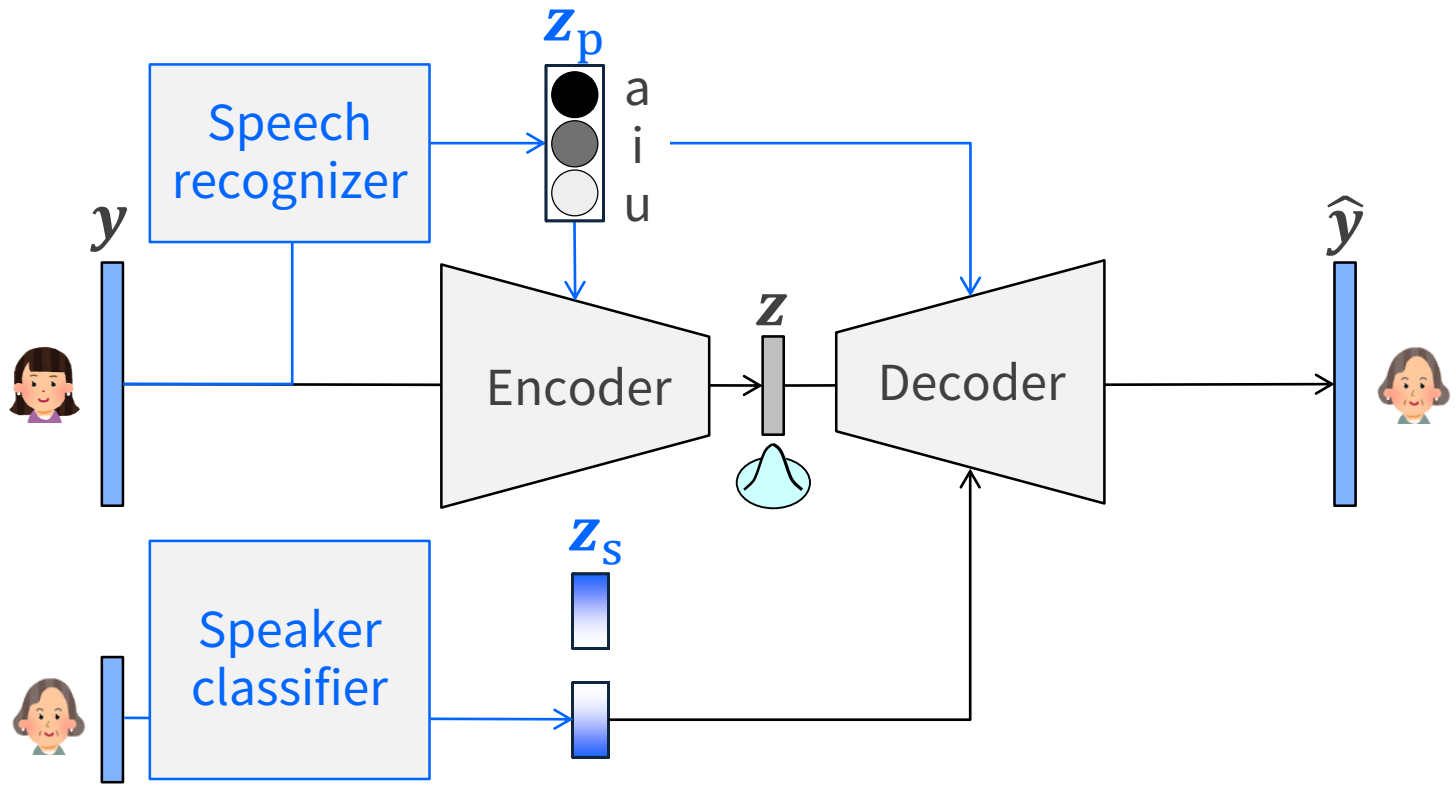
[Saito+ICASSP18-VC] & [Saito+AST20]



PPG: Phonetic PosteriorGram [Sun+16], d-vector [Variani+14]

VAE ベース多対多 VC への応用

1. 音声認識/話者認識の事前学習 (認識エラー最小化)
2. VAE 音響モデルの学習 (VAE 目的関数最適化)
3. 任意話者から任意話者への VC



VAE ベースの多話者音響モデリングに関する考察

音声合成の逆過程 (音声認識/話者認識) を統合した音響モデリング

提案法における認識モデル: 事前学習後は固定

→ **End-to-End 学習** [Zhang+17][Heigold+16] **による同時最適化**も可能

半教師あり学習への拡張

提案法は, 音声認識の事前学習用ラベル (テキスト書き起こし) が必要

→ Conditional VAE [Kingma+14] の枠組みに基づく**半教師あり学習**が可能

提案法のハイパーパラメータ: 学習話者数と d-vector 次元

前者は音響モデリングの頑健性に, 後者は話者空間の自由度に影響

学習話者数が少ない場合, d-vector 次元は大きいほうが better

学習話者数が多くなると, d-vector 次元は8次元まで小さくできる

→ **学習話者数と d-vector 次元のトレードオフを示唆 (5.3.4節)**

実験条件
(音声認識過程を統合した VAE-based VC)

Q. 音声認識過程の統合で, 多様な話者の音声は合成できるか？

学習/評価データ (22.05 kHz sampling)	**音声認識/話者認識モデル構築用: 260名** (男性130, 女性130, 約31時間) **VC の音響モデル構築用: 6名 (男性3, 女性3)** 同一内容 (パラレル) 425発話を3分割 (1—200: 変換元, 201—400: 変換先, 401—425: 評価)
音声パラメータ	40次元メルケプストラム, F0, 非周期性指標
DNN アーキテクチャ	すべて Feed-Forward 型ネットワーク
PPG の音素数	56
d-vector	16次元
VAE の潜在変数	64次元
手法 **VAE-*-PPG は,** **多対多 VC でも評価**	**FFNN: パラレル発話*で学習した DNN (理想条件)** **VAE-SC: 話者コードで条件付けした VAE** [Hsu+16] **VAE-SC-PPG: 話者コードと PPG で条件付けした VAE** **VAE-DV-PPG: d-vector と PPG で条件付けした VAE**

*学習データ収集時のコスト: パラレル発話 > ノンパラレル発話

変換音声の話者類似性に関する評価結果

DMOS テスト: 変換先話者との類似性を5段階で評価

変換元・変換先話者の性別 (男/女) ごとに評価を実施し, 結果を平均

1対1 VC: 変換元・変換先話者2名のみのデータで学習

多対多 VC: **変換元・変換先話者以外**の260名のデータで学習

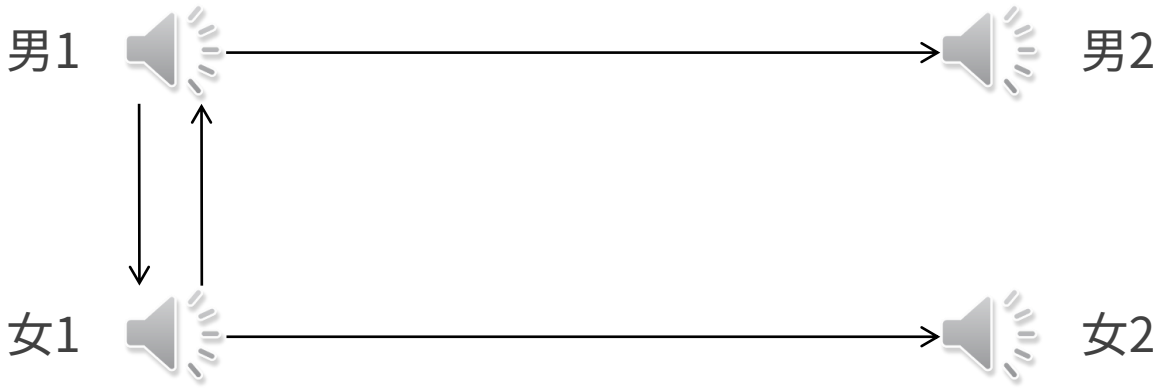
VAE-SC-PPG による多対多 VC では, 話者コード適応 [Luong+17] を利用

VC	手法	男-to-男	男-to-女	女-to-男	女-to-女
1対1	FFNN	3.49±0.16	3.20±0.14	3.13±0.14	2.99±0.15
	VAE-SC	1.28±0.10	1.21±0.08	1.24±0.09	1.27±0.09
	VAE-SC-PPG	**3.23±0.14**	**2.90±0.13**	**3.17±0.14**	**2.91±0.13**
	VAE-DV-PPG	3.13±0.14	2.87±0.13	3.04±0.13	2.84±0.14
多対多	VAE-SC-PPG	2.08±0.12	1.92±0.12	2.14±0.12	2.09±0.13
	VAE-DV-PPG	2.31±0.12	1.94±0.11	2.33±0.11	2.10±0.12

提案法により, 学習に使われない未知話者の音声も合成可能に！

各評価につき50名が参加 (24 answers/listener, 自然性に関しても同様の結果)

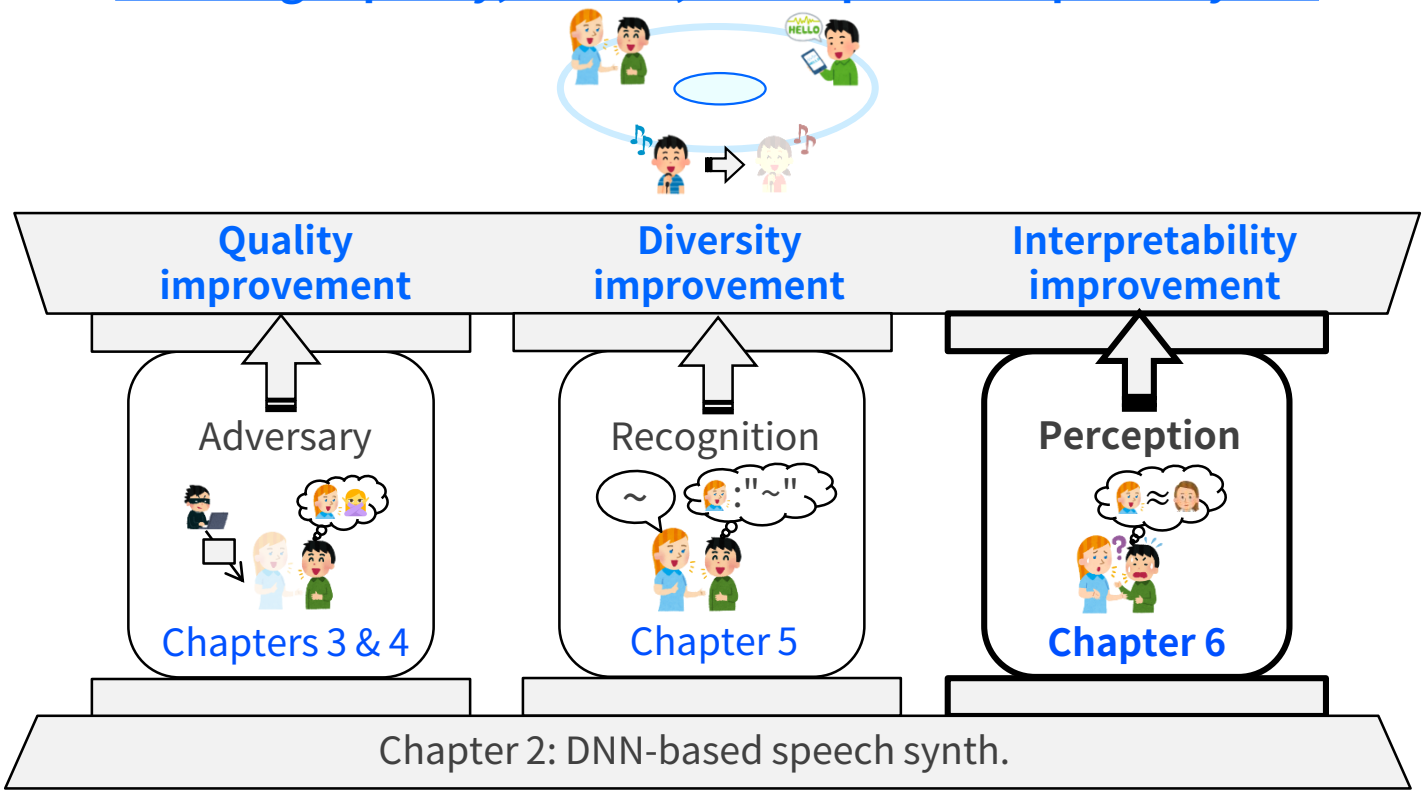
変換音声のサンプル



VC	手法	男1-to-男2	男1-to-女1	女1-to-男1	女1-to-女2
1対1	FFNN				
	VAE-SC				
	VAE-SC-PPG				
	VAE-DV-PPG				
多対多	VAE-SC-PPG				
	VAE-DV-PPG				

第6章:
音声知覚過程を考慮した解釈性の高い話者表現学習法

Goal: High-quality, diverse, & interpretable speech synth.



第6章の概要

本章の目的: 多話者音響モデリングにおける話者表現の解釈性向上

合成音声の話者性をより直感的に制御可能

課題: 人間の知覚をどのように統計的にモデル化するか

人間の知覚に関するデータをどのように収集するか?

収集されたデータで話者表現をどのように学習すべきか?

提案法: 話者間類似度の主観スコアリングに基づく話者表現学習

人間の「話者間の知覚的な関係性を記憶する」能力を模擬 & 統合

複数話者の音声の主観的な類似度に関するスコアリングを実施

得られたスコアに基づいて話者表現を生成する DNN を学習

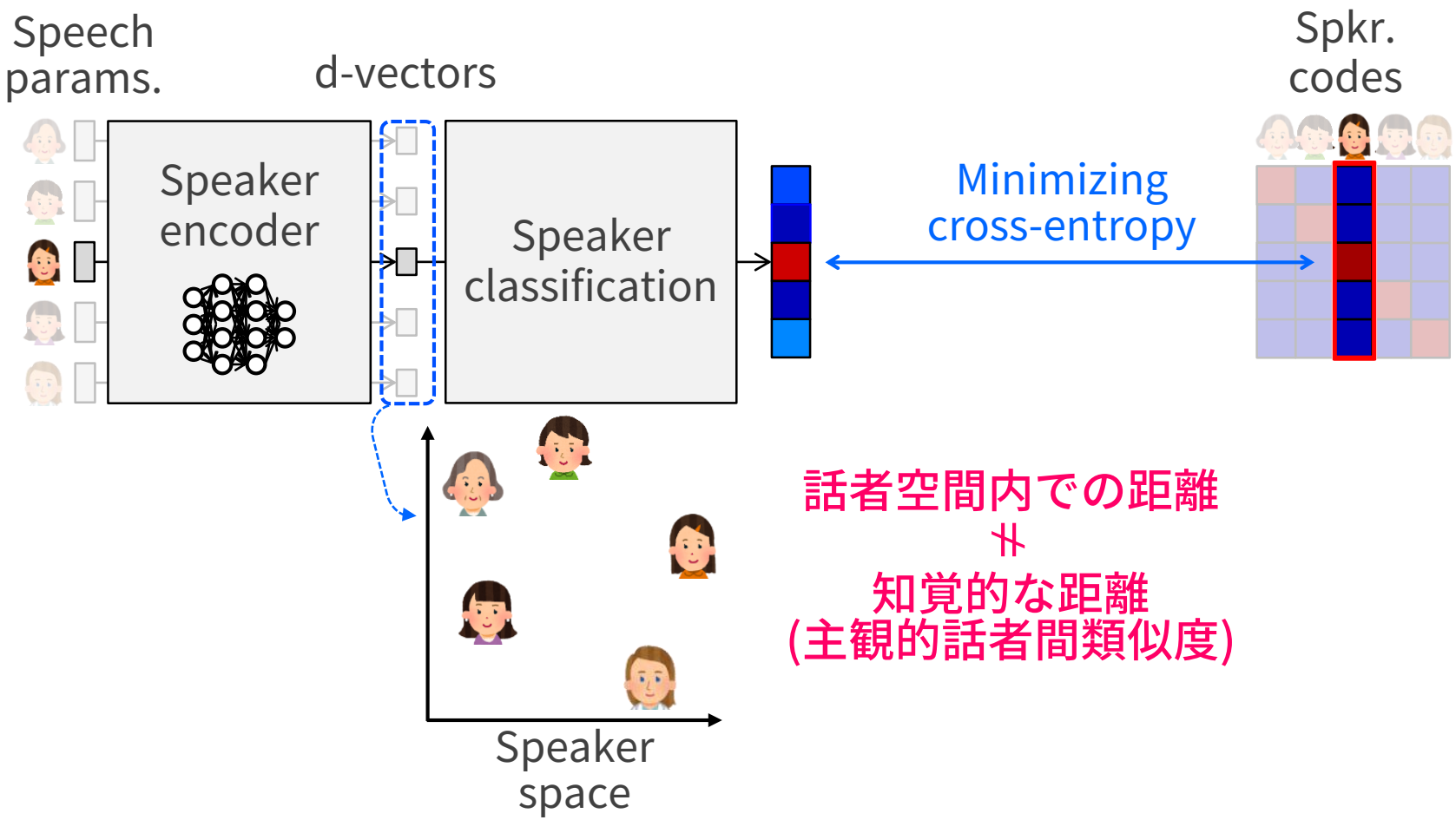
スコアの表現形: 類似度 { スコアベクトル, スコア行列, グラフ }

類似度スコアリングと DNN 学習を反復する active learning も提案

d-vector の問題点:
主観的な話者間の類似度を無視した話者表現

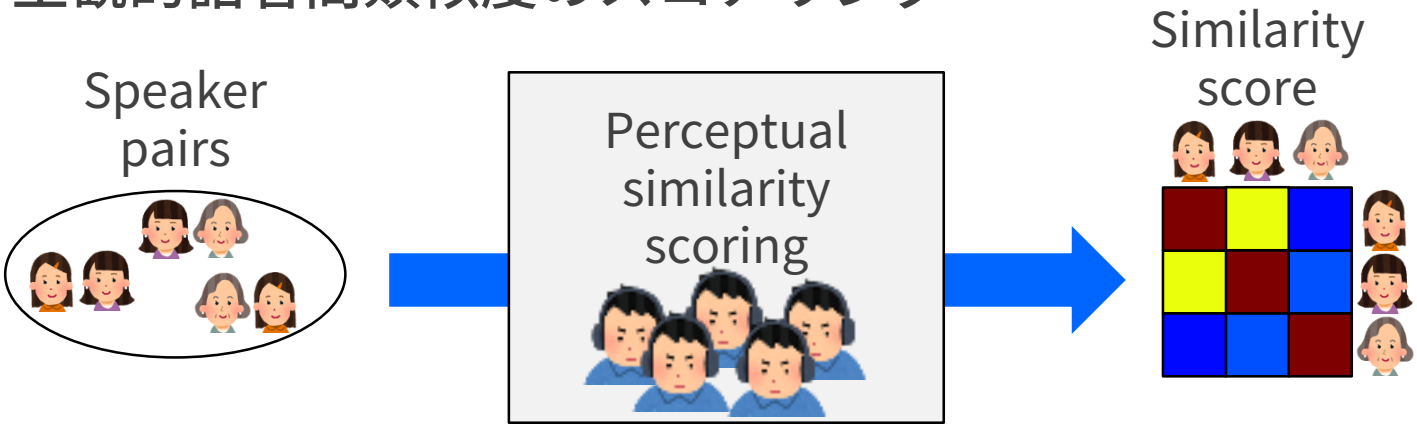
音声パラメータから話者コード [Hojo+18] を予測するような学習

話者の識別/認証に適した話者表現 = 話者知覚を無視

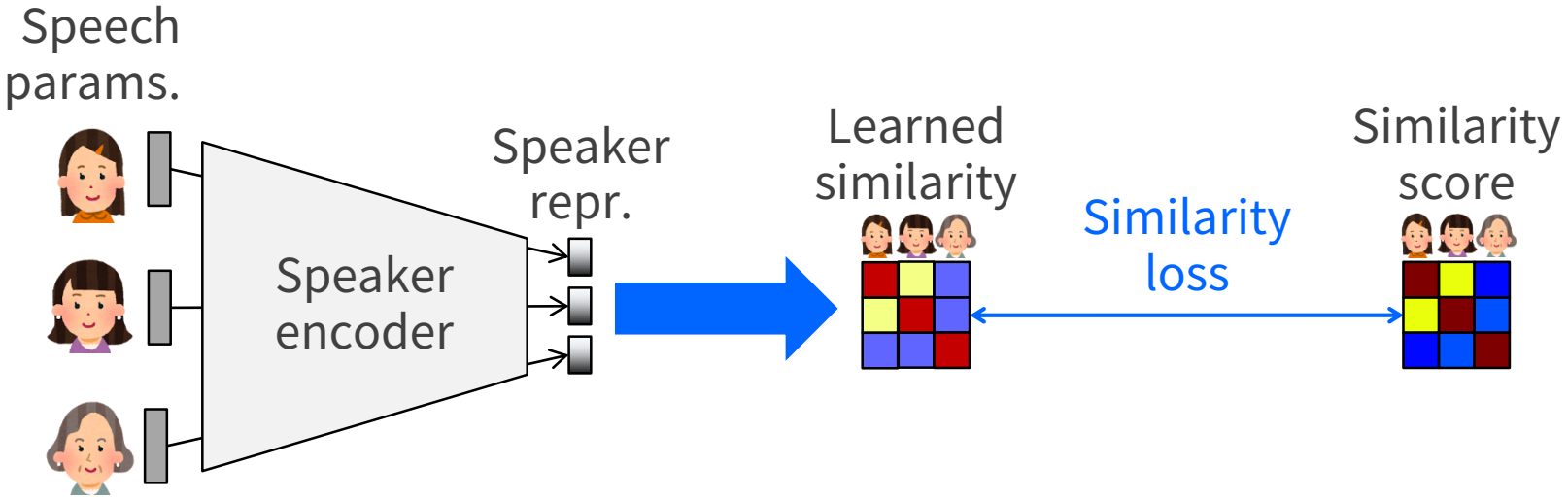


提案法: 主観的話者間類似度を考慮した話者表現学習

1. 主観的話者間類似度のスコアリング



2. 類似度スコアを用いた DNN ベースの話者表現学習



主観的話者間類似度の大規模スコアリング

クラウドソーシングで, 話者間の主観的な類似度をスコアリング

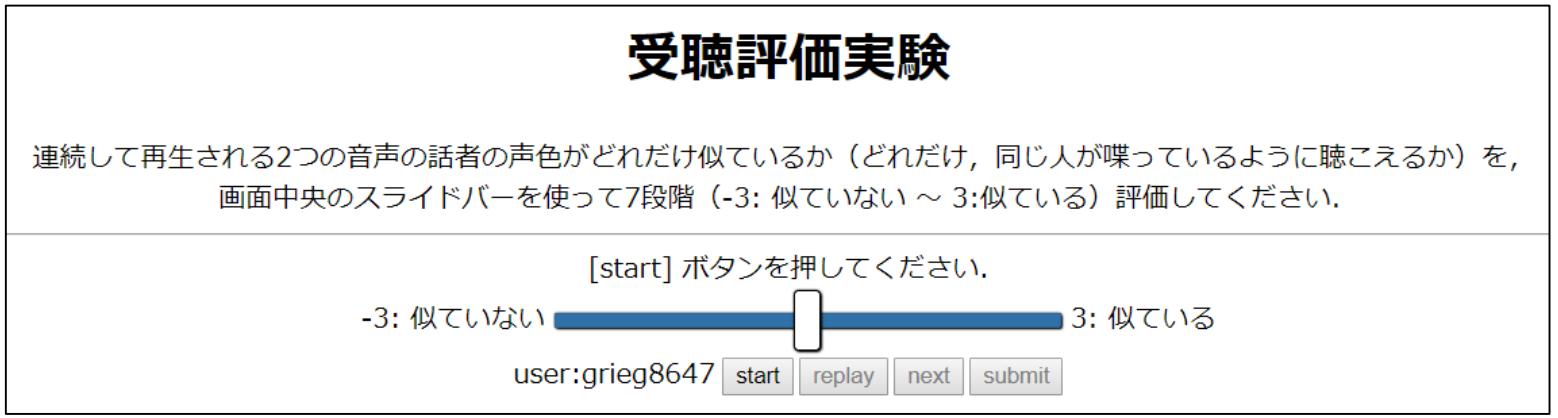
JNAS [Itou+99] コーパスに含まれる153名の女性話者の発話を使用

各話者毎に異なる発話内容 → テキスト非依存な類似度を評価

合計の評価者数: **4,060** 名 (ランダムに選ばれた34話者対/評価者)

スコアリングの評価値: −3 (似ていない) 〜 +3 (似ている)

1つの話者対を少なくとも異なる10名以上が評価



話者対のサンプル

主観的話者間類似度の行列表現

類似度スコア行列 $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_i, \cdots, \boldsymbol{s}_{N_S}]$

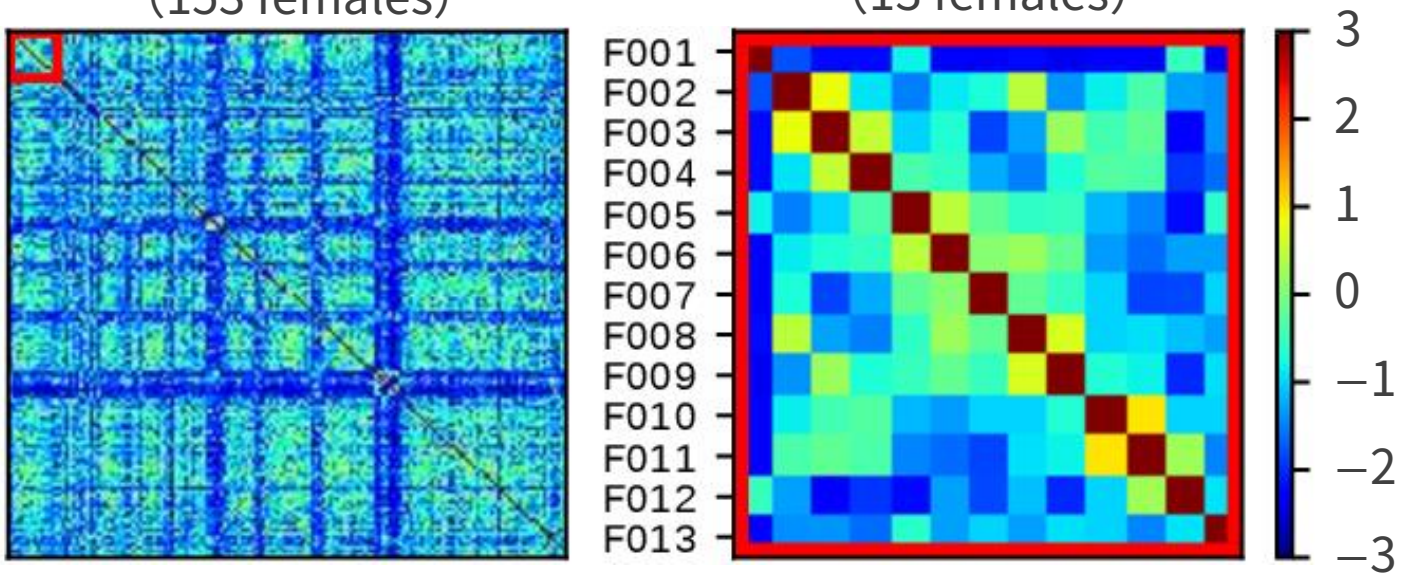
N_S: スコアリングに用いられた話者の数

$\boldsymbol{s}_i = [s_{i,1}, \cdots, s_{i,j}, \cdots, s_{i,N_S}]^\top$: i番目の話者の類似度スコアベクトル

$s_{i,j}$: i番目の話者とj番目の話者の類似度スコア $(-v \leq s_{i,j} \leq v)$

(a) Full score matrix
(153 females)

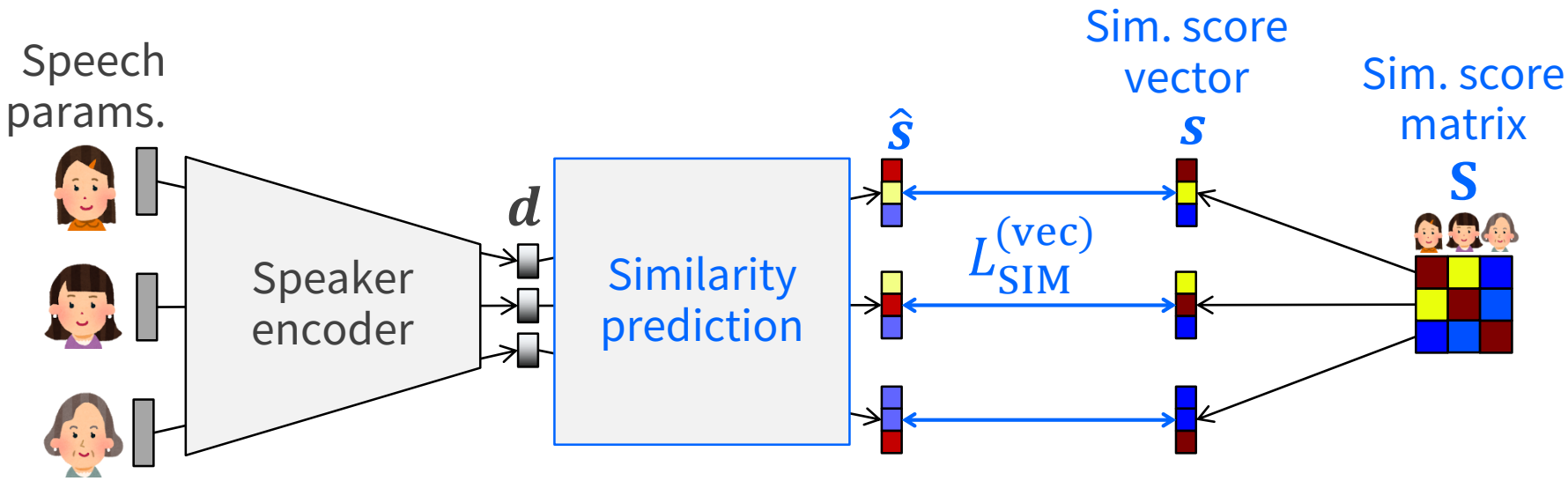
(b) Sub-matrix of (a)
(13 females)



本論文では, 類似度スコアを用いた話者表現学習法を3つ提案

学習法1: 類似度ベクトル埋め込み

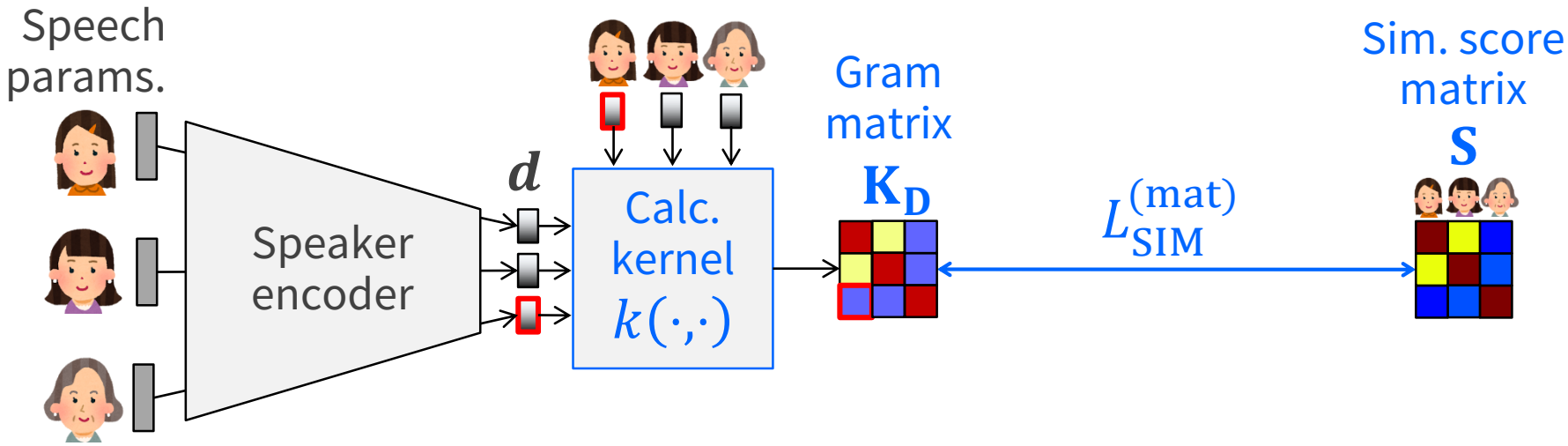
音声パラメータから類似度スコアベクトルを予測するように学習



$$L_{\mathrm{SIM}}^{(\mathrm{vec})}(\boldsymbol{s},\hat{\boldsymbol{s}}) = \frac{1}{N_s}(\hat{\boldsymbol{s}}-\boldsymbol{s})^{\mathsf{T}}(\hat{\boldsymbol{s}}-\boldsymbol{s})$$

学習法2: 類似度行列埋め込み

話者表現の Gram 行列を類似度スコア行列に近づけるように学習

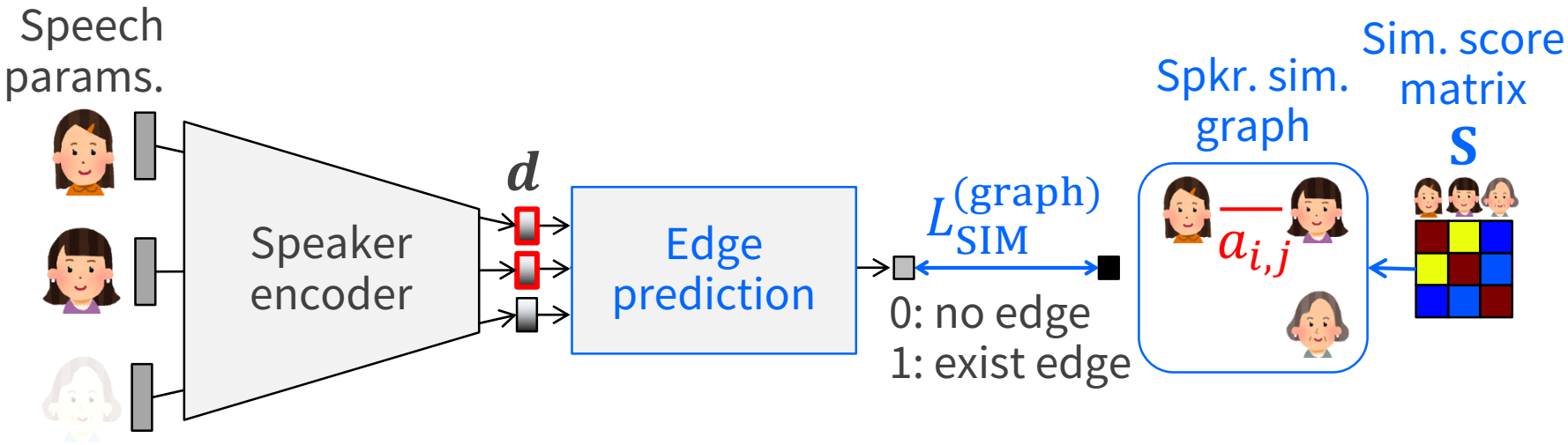


$$L_{\mathrm{SIM}}^{(\mathrm{mat})}(\mathbf{D},\mathbf{S}) = \frac{1}{Z_{\mathrm{S}}} \left\| \tilde{\mathbf{K}}_{\mathbf{D}} - \tilde{\mathbf{S}} \right\|_F^2$$

Z_{S}: 話者数の影響を正規化するための係数 ($\tilde{\cdot}$ は, 当該行列の対角成分を除いた行列)

学習法3: 類似度グラフ埋め込み

話者表現の対から類似度グラフの辺の有無を予測するように学習

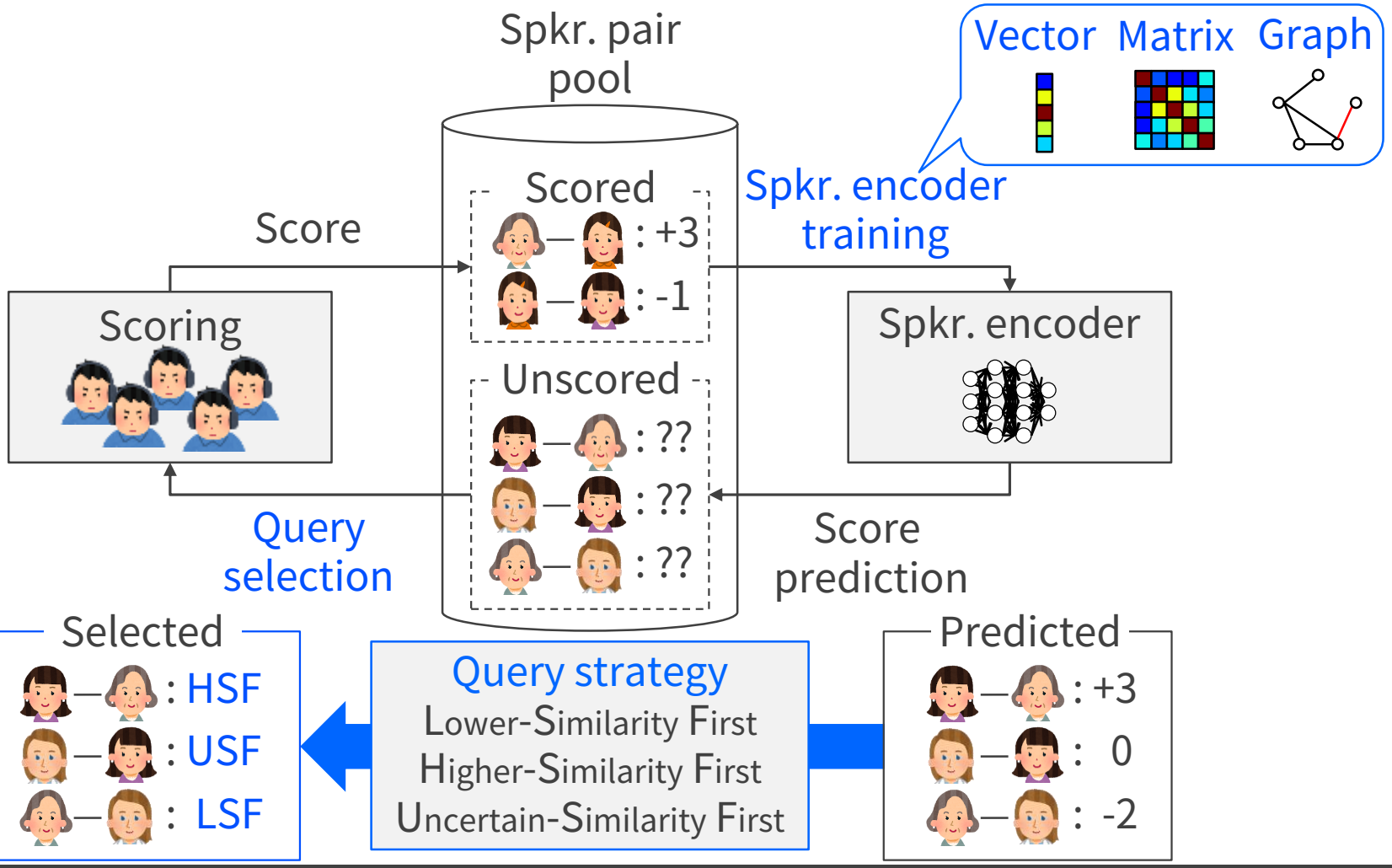


$$L_{\mathrm{SIM}}^{(\mathrm{graph})}(\boldsymbol{d}_i, \boldsymbol{d}_j) = -a_{i,j} \log p_{i,j} - (1 - a_{i,j}) \log(1 - p_{i,j})$$

$p_{i,j} = \exp\left(-\left\|\boldsymbol{d}_i - \boldsymbol{d}_j\right\|_2^2\right)$: 辺の生起確率 ([Li+18] を参照に定義)

主観的話者間類似度を考慮した話者表現の active learning

スコアリングと話者表現学習を反復し, スコアリングコストを削減



主観的話者間類似度を考慮した話者表現学習に関する考察

提案法: 話者間の印象 (主観的類似度) を考慮した話者表現学習

- 単一話者の印象を考慮した従来研究の拡張
  - HMM-TTS [Tachibana+06], GMM-VC [Ohta+07]
- 話者と聴者間の関係もモデル化可能 (例: 感情表現と知覚の差) [Lorenzo-Trueba+18]

提案する話者表現学習法の比較

- 類似度 { ベクトル, 行列 } 埋め込み: 回帰問題に基づく最適化
  - 類似度スコアの値を直接用いる学習法
- 類似度グラフ埋め込み: 識別問題に基づく最適化
  - 類似度スコアの値から間接的に定まる関係性に着目する学習法
  - グラフ信号処理 [Shuman+13] やグラフ NN [Scarselli+08] の知見も導入可能

提案する active learning: Human-in-the-loop 型話者表現学習

- 人間が計算ループに内在するような音声合成研究の先駆け
  - 例) discriminator を人間で置換した GAN ベース音響モデリング [Fujii+20]

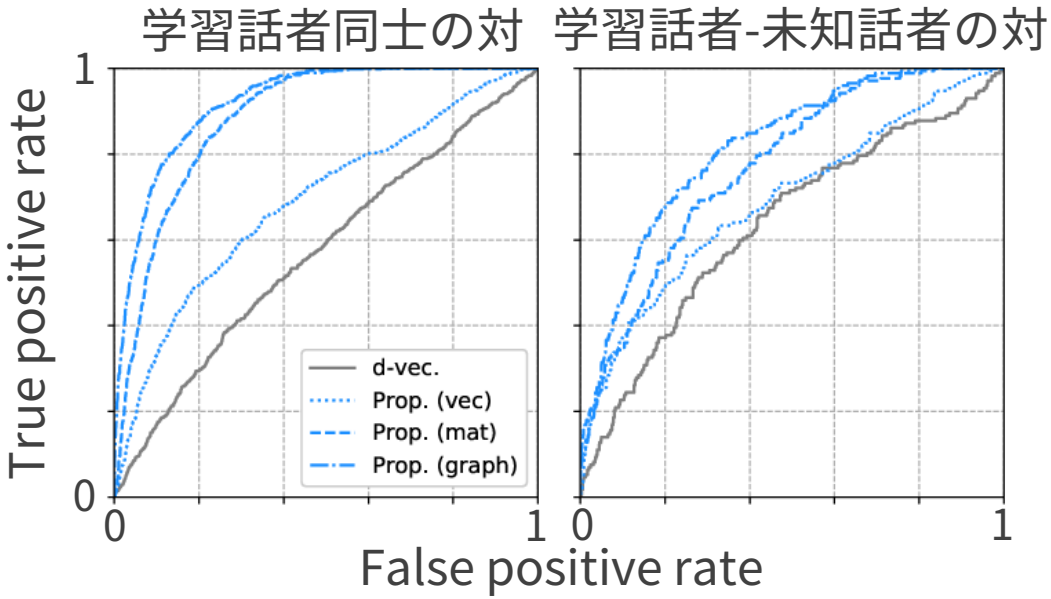
実験条件
(主観的話者間類似度を考慮した話者表現学習)

Q. 提案法により, 話者表現の解釈性は向上するか？

学習/評価データ (16 kHz sampling)	JNAS [Itou+99] の女性話者153名 主観スコアリング用: 5発話 話者表現学習/評価用: 約130発話/約15発話 **(F001 ~ F013 の13名は, 学習データから除外 = 未知話者)**
主観スコアリングの値	−3 (似ていない) ~ ＋3 (似ている) の整数 (話者表現学習時には [−1, ＋1] か [0, 1] に正規化)
音声パラメータ	40次メルケプストラム, F0, 非周期性指標
DNN アーキテクチャ	すべて Feed-Forward 型ネットワーク
手法	**(1) d-vec. : 話者認識に基づく学習** [Variani+14] **(2) Prop. (vec): 類似度ベクトル埋め込み** **(3) Prop. (mat): 類似度行列埋め込み** **(4) Prop. (graph): 類似度グラフ埋め込み**

話者表現を用いた類似話者対検出の ROC 曲線と AUC

Receiver Operating Characteristic (ROC) 曲線



Area Under the ROC Curve (AUC): ROC 曲線の下面積 (0.5 ~ 1.0)

	d-vec.	Prop. (vec)	Prop. (mat)	Prop. (graph)
学習話者-学習話者	0.5711	0.6909	0.8847	**0.9204**
学習話者-未知話者	0.6347	0.6909	0.7712	**0.8157**

提案法は, 話者間の主観的な類似度を考慮した話者表現を学習！

合成音声の話者類似性に関する主観評価 (話者補間)

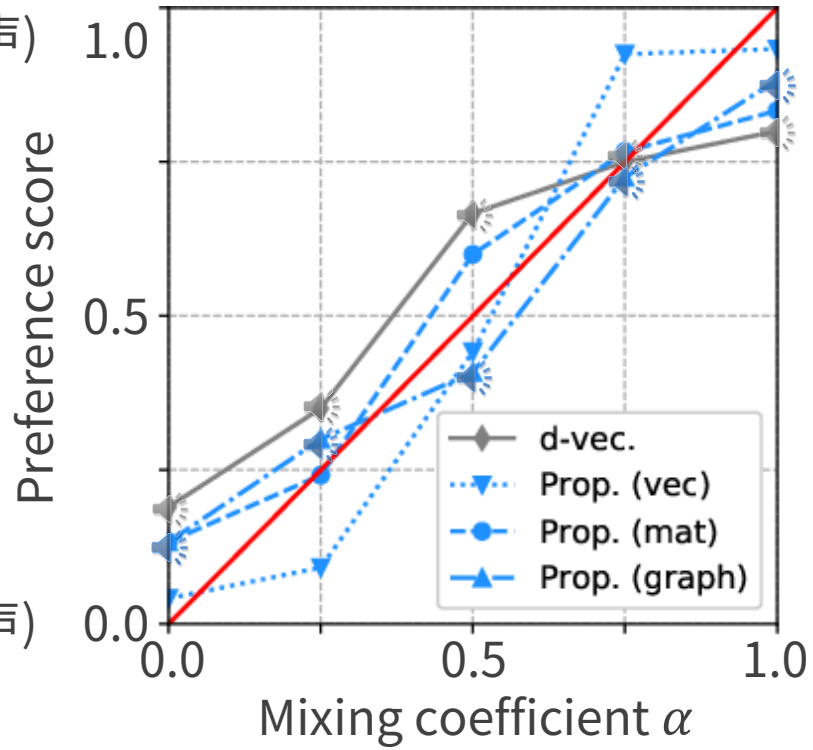
話者補間: 複数話者の話者表現を混合し, 新たな話者性を生成

混合係数 $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ の音声 (X) の話者類似性を評価

スコア曲線が赤線に近いほど, 直感的な話者性制御が実現可能

B ($\alpha = 1$ の合成音声)

A ($\alpha = 0$ の合成音声)



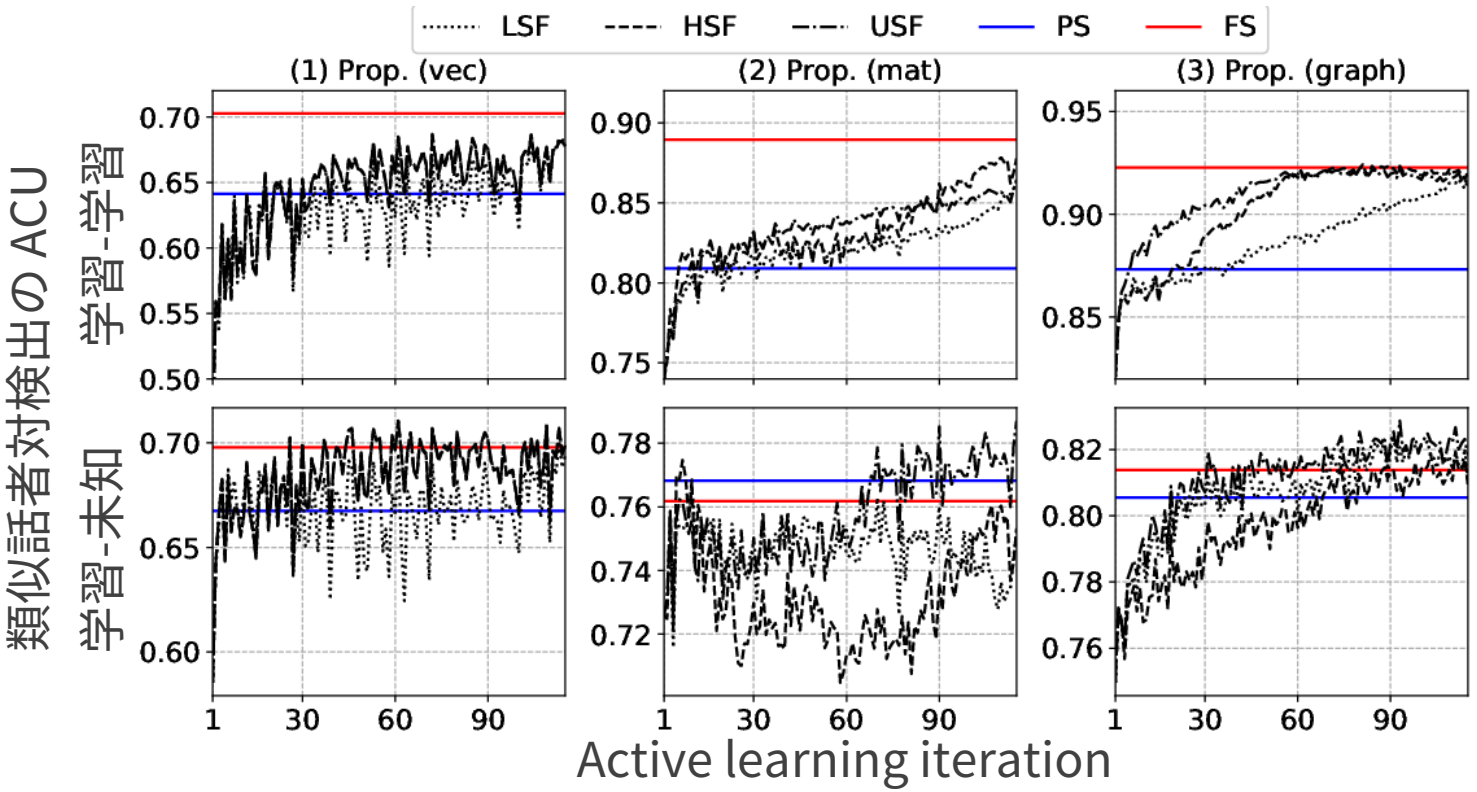
提案法は, より直感的に話者性を制御可能な話者補間を実現！

(20 answers/listener, total 30 × 2 listeners, 手法ごとに独立して評価を実施)

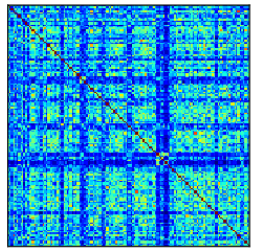
Active learning の反復による AUC の改善

各 Prop. (*) 毎に, active learning とクエリ戦略の影響を調査

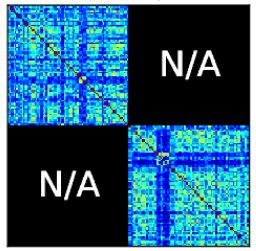
反復により, PS よりもどれだけ改善するか & FS にどれだけ近づくか



Fully Scored (FS)



Partially Scored (PS)



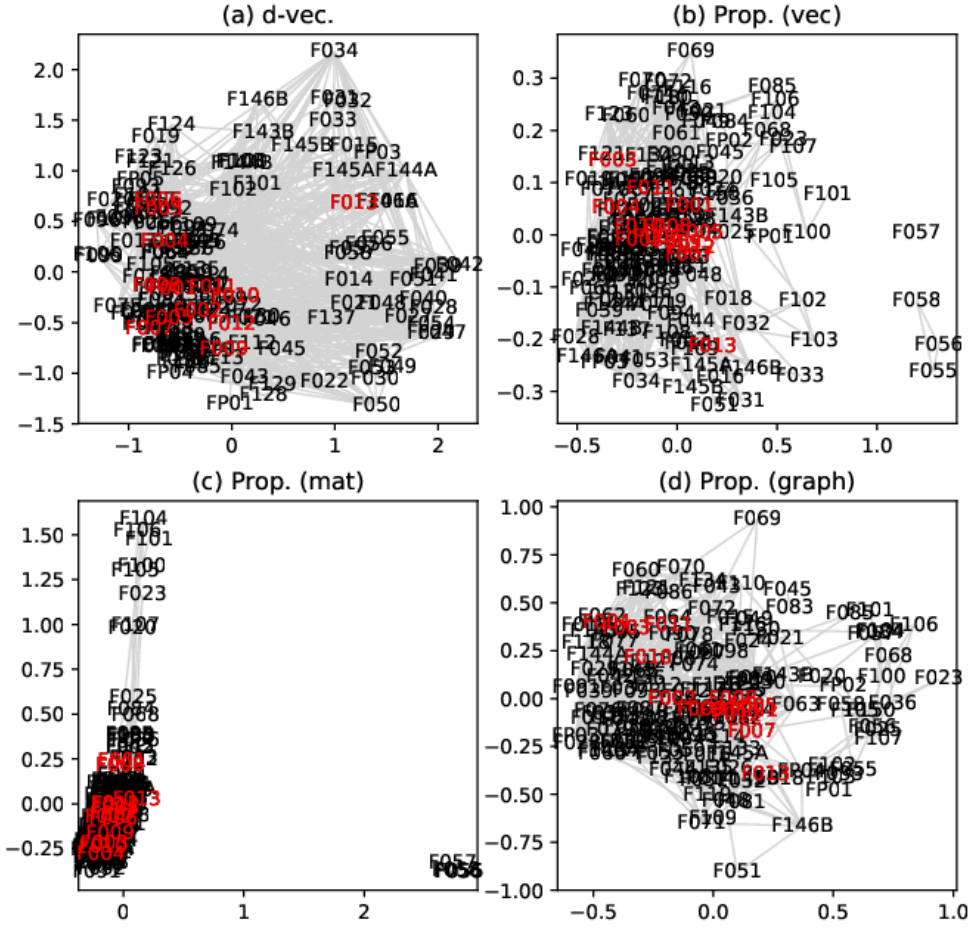
USF に基づく active learning は,
最終的な FS と同程度の AUC をより少ない反復で達成！

話者空間の可視化 (PCA プロット)

[修正]

Prop. (mat) は, disjoint な話者クラスタを過度に分離

→ 学習に用いるスコアによって大きく影響を受けやすい傾向?



結論

本論文: 人間の音声情報処理能力を統合した音声合成理論

人間からヒントを得て, 音声合成の品質・多様性・制御性を改善

本論文での提案手法

音声敵対過程を統合した高品質な音声合成 (第3章 & 第4章)
音声認識過程を統合した多様な音声合成 (第5章)
音声知覚過程を考慮した解釈性の高い話者表現学習 (第6章)

主な学術的貢献

第3章 & 第4章:

世界初の GAN ベース音声合成法 & anti-spoofing との関連性を議論

第5章:

音声/話者認識の知見導入による VAE ベース音声合成法の理論拡張

第6章:

世界初の human-in-the-loop 型話者表現学習法

関連する研究業績など

博士論文に関連する国際論文: 国際会議原稿×4 + 原著論文×6

- [Saito+ICASSP17][Saito+TASLP18] ... 第3章
- [Saito+ICASSP18-TTS][Saito+CSL19] ... 第4章
- [Saito+ICASSP18-VC][Saito+AST20] ... 第5章 & Appendix C
- [Saito+SSW19][Saito+TASLP20(submitted)] ... 第6章
- [Saito+IEICE17] ... Appendix A
- [Saito+IEICE20] ... Appendix B

博士論文に関連する受賞: 10件

- **2020 IEEE SPS Young Author Best Paper Award**
- Spoken Language Processing Student Grant Award of ICASSP2017
- 第34回 電気通信普及財団 テレコムシステム技術学生賞
- 2018年度 C&C 若手優秀論文賞
- etc...