第21回音声言語シンポジウム
音声研究会（SP） 音声言語情報処理研究会（SLP）2研究会連立開催研究会
@NHK放送技術研究所　2019年12月6日(金) 16:55～17:55

画像変換／系列変換アプローチを用いた音声変換

NTTコミュニケーション科学基礎研究所
亀岡弘和　金子卓弘　田中宏　北条伸克

背景

- コミュニケーションは社会生活の最も重要な要素の１つ
- コミュニケーションには、発声・聴覚機能の障がいや加齢による衰え、不慣れな言語での会話能力、心理的な緊張状態など、**物理的・能力的・心理的な状態に起因する様々な形のストレスやバリア**が存在

- **音声変換(Voice Conversion; VC)技術**によりこのようなストレスやバリアをどこまで克服できるか？ cf) 無喉頭音声から通常音声への変換 [Nakamura+2012][Doi+2014]

リアルタイム音声変換(VC)システム

- リアルタイムVCシステムにより
  効率的かつ円滑なコミュニケーションを実現

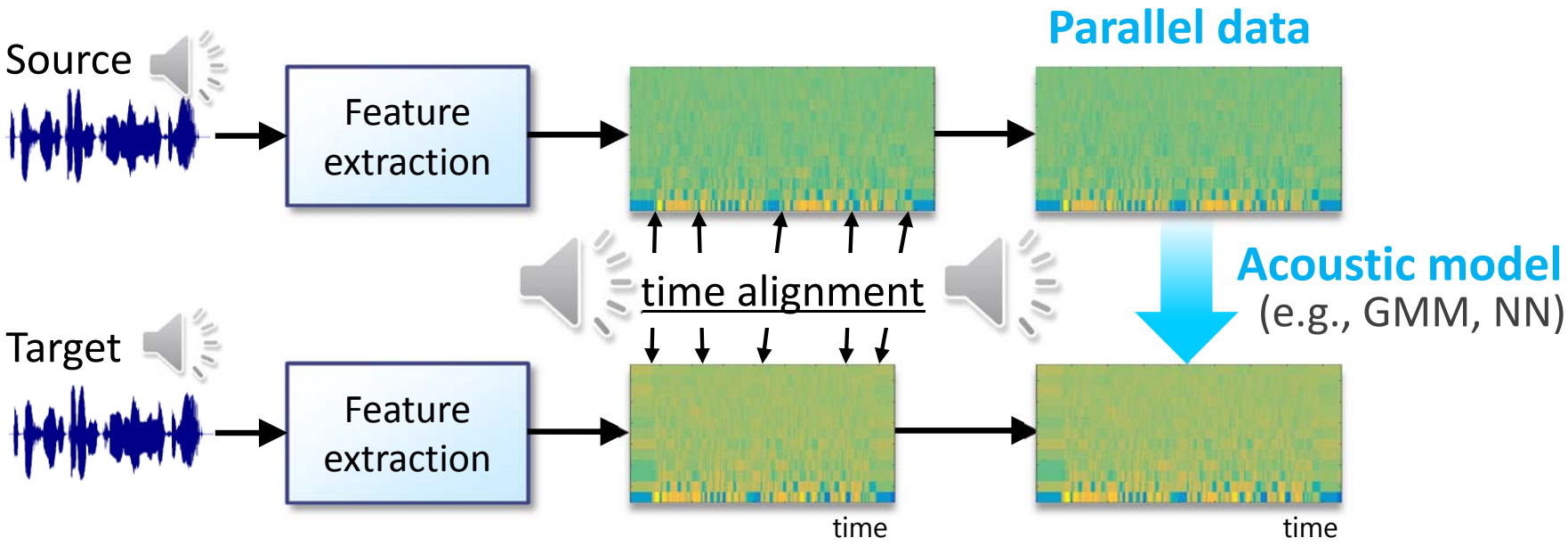
タスク例

- 通常音声 → 別人の声
- 訛りのあるアクセント → 聞き取りやすいアクセント
- 発声障がい者の声 → 健常者の声

パラレル音声変換(VC)方式

- ソース音声とターゲット音声のパラレルコーパスの用意
- 両音声の時間整合処理
- 時間整合された特徴量ペアを用いて音響モデルを学習
  - 音響モデルの例

    Gaussian mixture model (GMM) [Toda+2007], Neural net (NN) [Desai+2010]



パラレルVC方式の強みと限界

強み

- 軽量な音響モデル(GMMなど)で動作するためリアルタイム（低遅延）システムを実現可能

限界

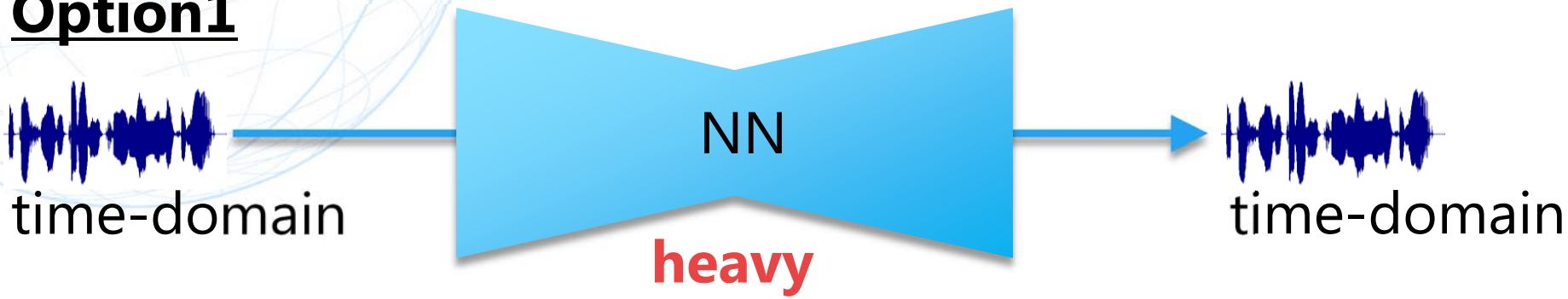
- パラレルデータが必要
  - パラレルデータの準備は高コスト
- 声質の変換が主目的
  - 超分節的特徴（アクセント、F0パターン、発話リズムなど）の変換が困難
- 音質
  - 古典的なボコーダ（メルケプストラムボコーダなど）で合成される音声と実音声は容易に聞き分け可能

→ 深層学習アプローチにより克服できるか？

VCシステムの構成

- End-to-End構成？手動設計処理部を含む構成？

Option1



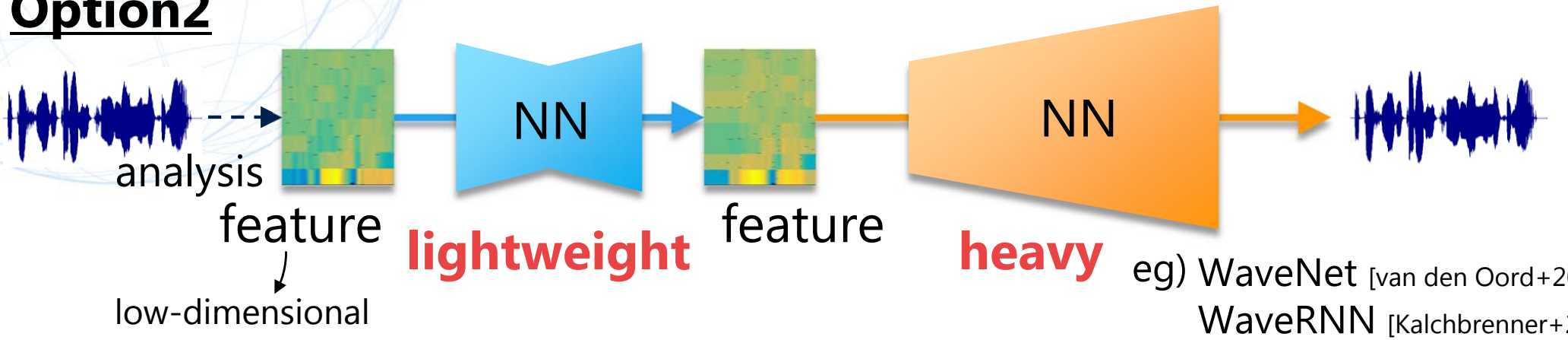
eg) VQ-VAE [van den Oord+2017]

- 完全なend-to-end
  - 大量の学習データで学習できれば高品質な音声変換が実現可能（ななず）
  - リアルタイム化するには不向き（？）

VCシステムの構成

- End-to-End構成？手動設計処理部を含む構成？

Option2

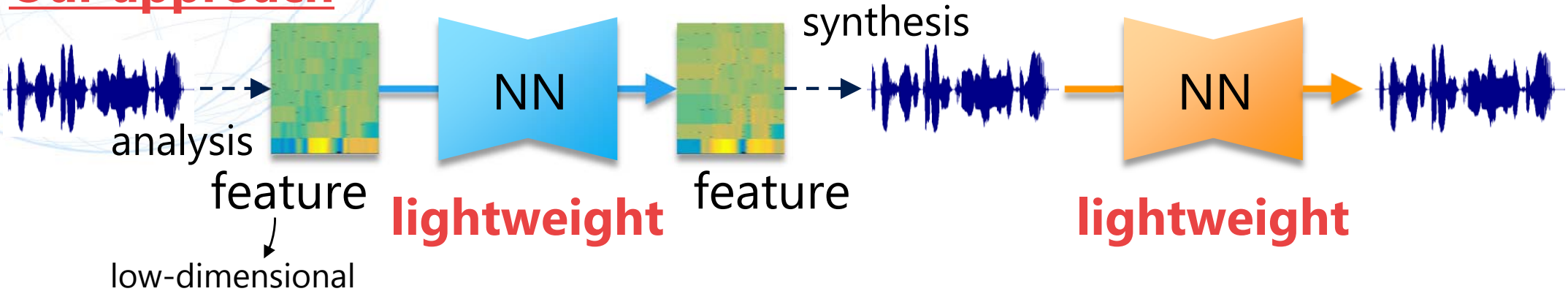


- 特徴量レベルの変換の後に波形生成
  - 特徴量変換モデルは軽量化されるため効率的な変換が可能
  - 波形生成の品質と効率性の両立はいまだ課題あり（？）
    （ただし最近は低遅延で波形生成可能なニューラルボコーダも多数出現している）

VCシステムの構成

- End-to-End構成？手動設計処理部を含む構成？

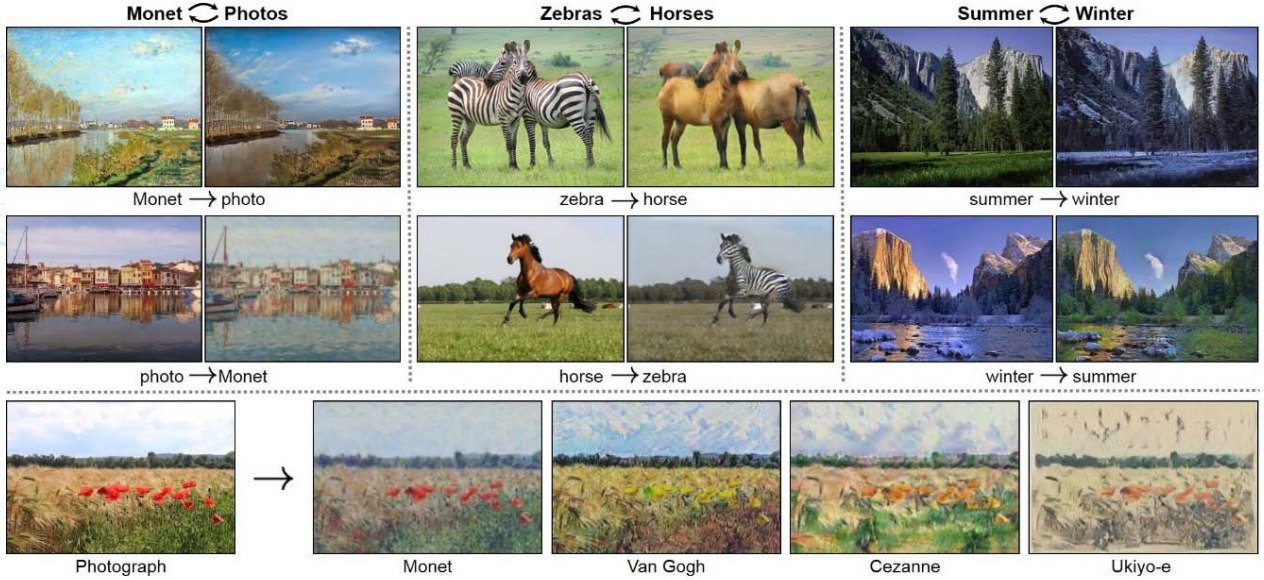
Our approach



- 特徴量変換⇒ボコーダによる音声生成⇒波形ポストフィルタリング
  - 特徴量変換モデルは軽量化されるため効率的な変換が可能
  - 波形ポストフィルタリングのためのモデルは、ゼロベースで波形生成するモデルに比べてはるかに軽量にすることが可能

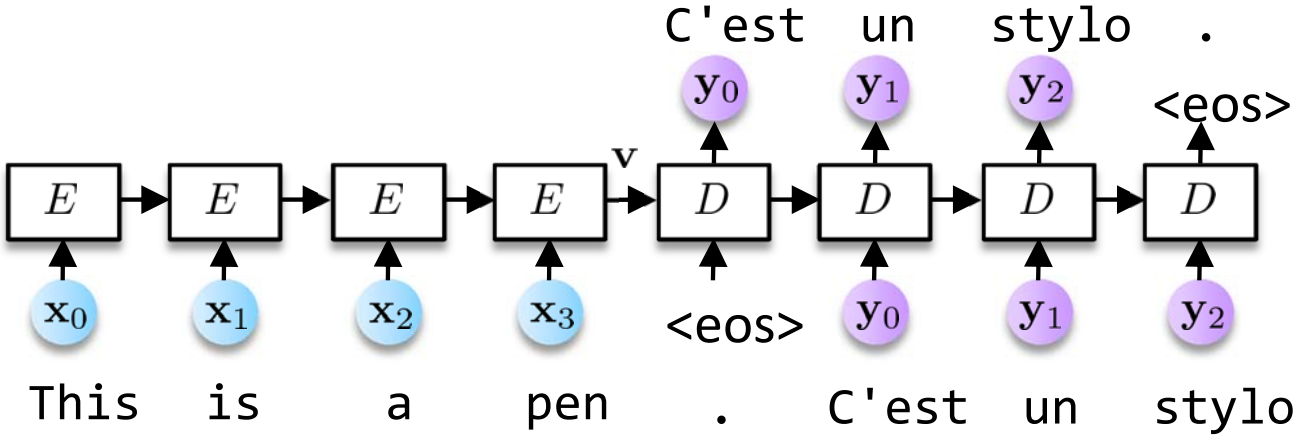
我々のアプローチ

- 画像変換アプローチ → 音響特徴量変換
  波形ポストフィルタリング



[Zhu+2017]

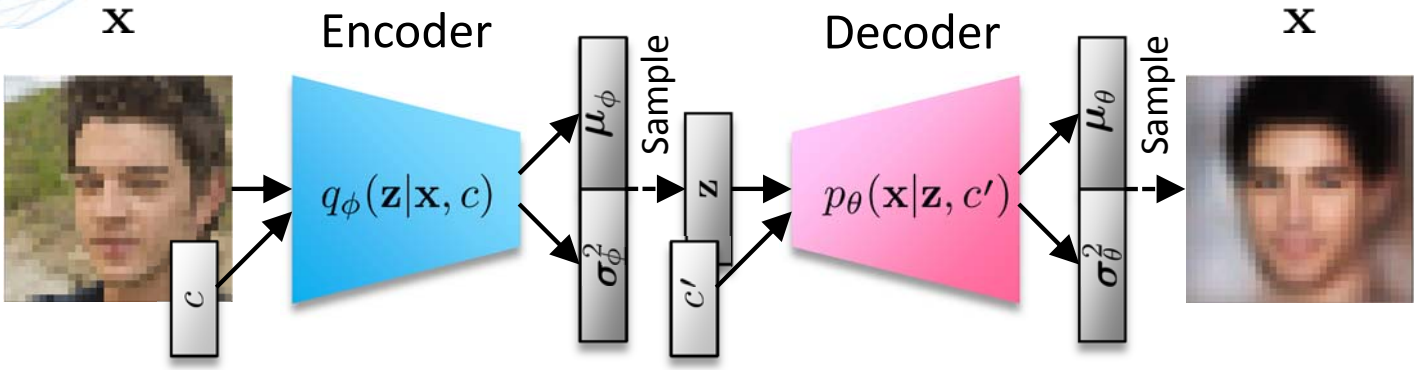
- 系列変換(Sequence-to-sequence)アプローチ → 音響特徴量変換



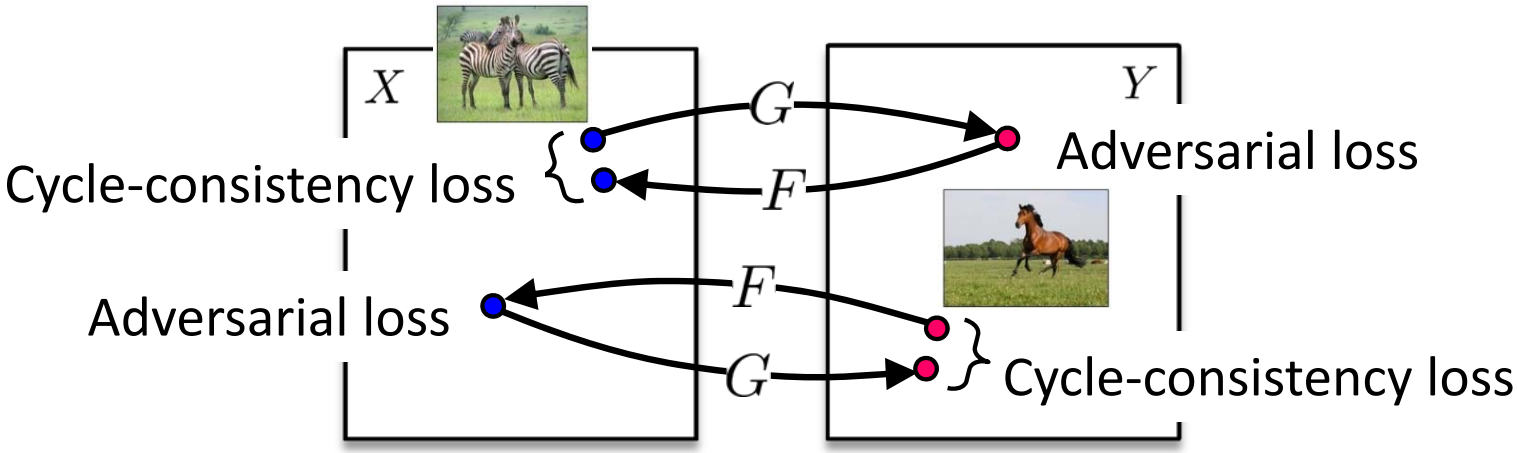
非パラレル画像変換

ソース画像とターゲット画像のペアデータを用いずとも画像のドメイン間変換を可能にする深層生成モデル

- Variational Autoencoder (VAE) [Kingma+2014, Yan+2016]



- Cycle-consistent Adversarial Network (CycleGAN) [Zhu+/Kim+/Yi+2017]



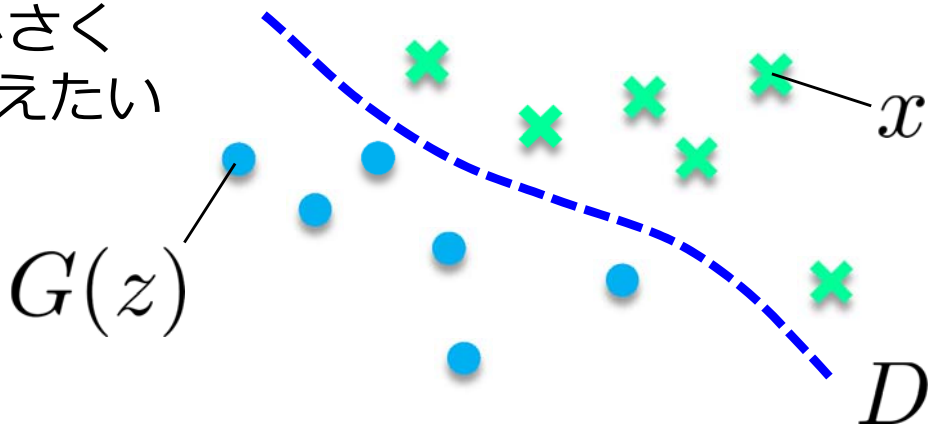
Generative Adversarial Network (GAN) [Goodfellow+2014]

- 分布形を仮定することなく学習サンプルの分布に従う擬似サンプルを生成する生成器を学習する枠組
- 実サンプルか生成器$G(z)$が生成した擬似サンプルかを識別する識別器 $p = D(x) \in [0,1]$ をだますように$G(z)$を学習
- $D(x)$もだまされないように学習

敵対ロス
（Dから見た識別スコア）

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]$$

$G(z)$は識別スコアを小さくしたい=識別境界を越えたい



Dは識別スコアが大きくなるように学習される

⇒音声合成への応用も [Kaneko+2016][Saito+2016]

Generative Adversarial Network (GAN) [Goodfellow+2014]

NTT

- 分布フィッティングとしての解釈
  - Dがどういう時に敵対ロスは最大になるか？

$$V(G,D) = \int p_{\text{data}}(x)\log D(x)dx + \int p_z(z)\log(1-D(G(z)))dz$$

$$= \int p_{\text{data}}(x)\log D(x)dx + \int p_G(x)\log(1-D(x))dx$$

$$\frac{\partial V(G,D)}{\partial D(x)} = \frac{p_{\text{data}}(x)}{D(x)} - \frac{p_G(x)}{1-D(x)} = 0 \longrightarrow \hat{D}(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_G(x)}$$

  - よってGのみにとってのロスは

$$C(G) = \max_D V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_G(x)}\right]$$

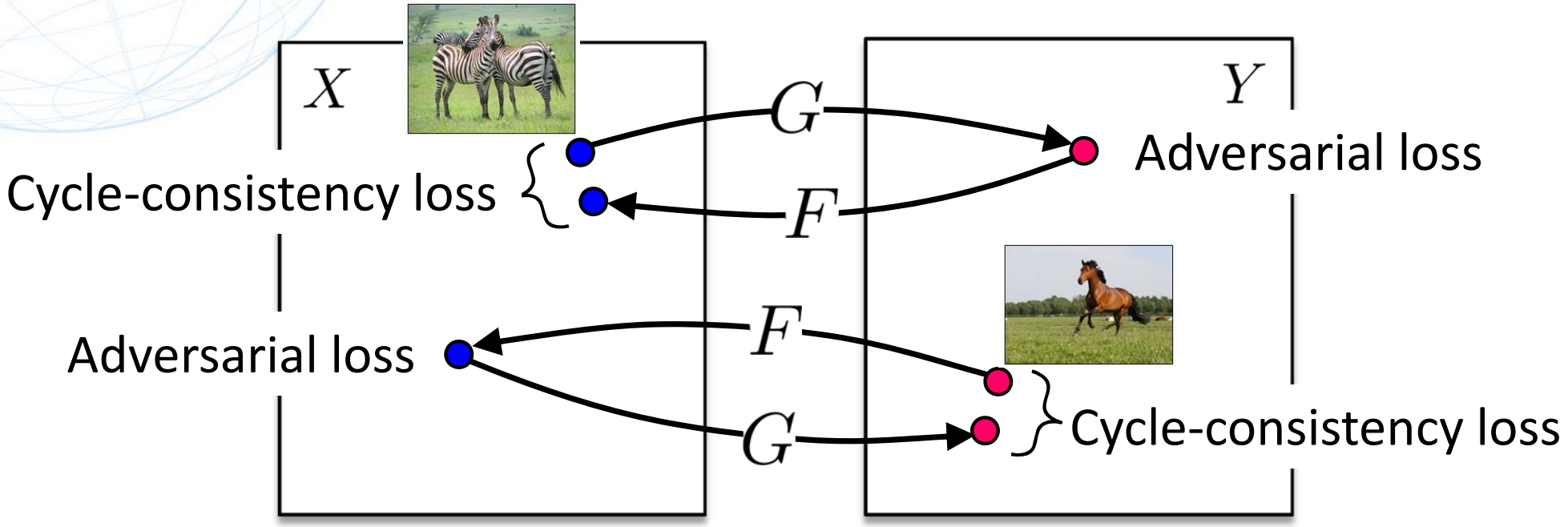
$$+\mathbb{E}_{x\sim p_G(x)}\left[\log \frac{p_G(x)}{p_{\text{data}}(x)+p_G(x)}\right]$$

$$= -\log(4) + \mathrm{KL}\left(p_{\text{data}} \middle\| \frac{p_{\text{data}}+p_G}{2}\right) + \mathrm{KL}\left(p_G \middle\| \frac{p_{\text{data}}+p_G}{2}\right)$$

p_{data}とp_GのJensen-Shanon (JS) ダイバージェンス

CycleGAN [Zhu+/Kim+/Yi+2017]

- ソース画像とターゲット画像のペアデータを用いずとも画像のドメイン間変換を可能にする深層生成モデルの１つ

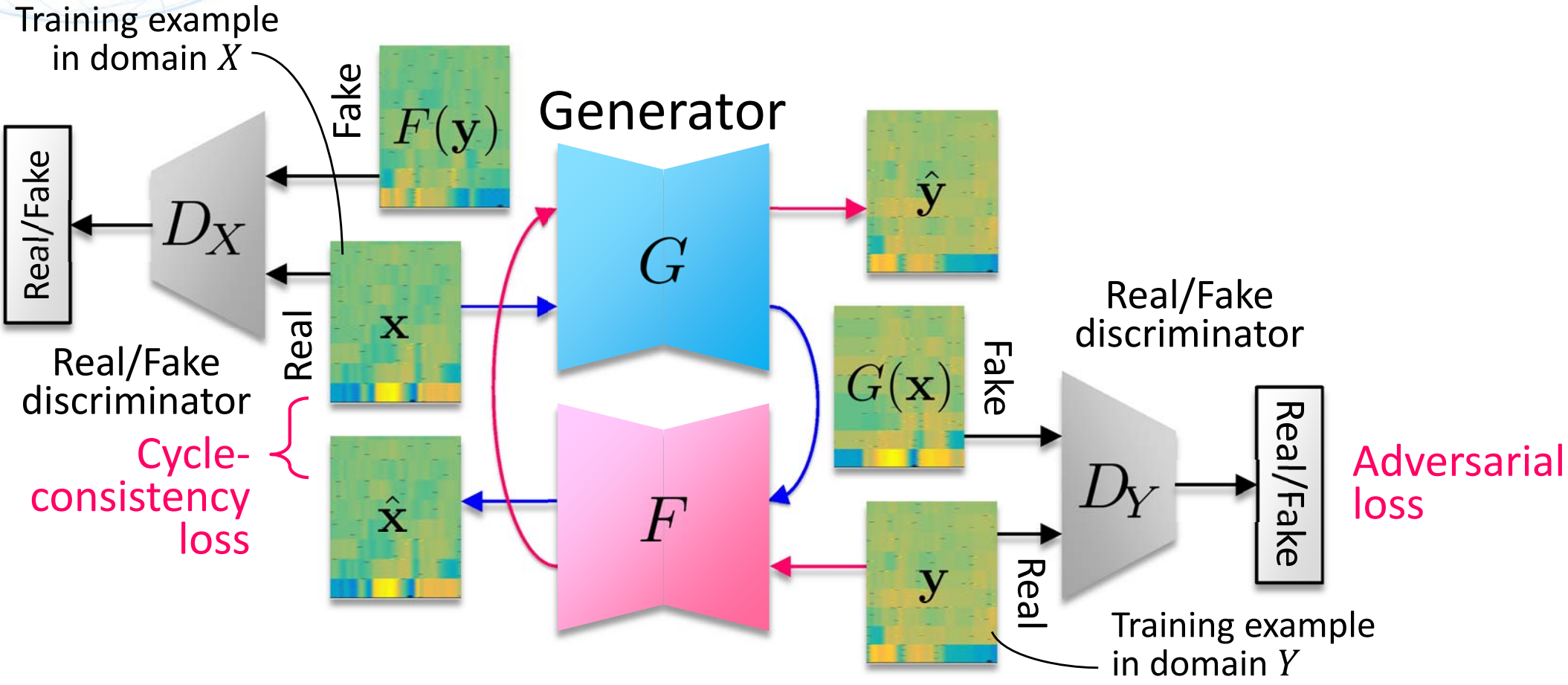


– 敵対ロス：実画像と生成画像を識別する識別器による学習ロス
  - 識別器D_X, D_Yはこれを小さくするように学習し、逆に変換器G, Fは大きくするように学習することでG, Fによる生成画像の分布が各ドメインの実画像の分布に近づく

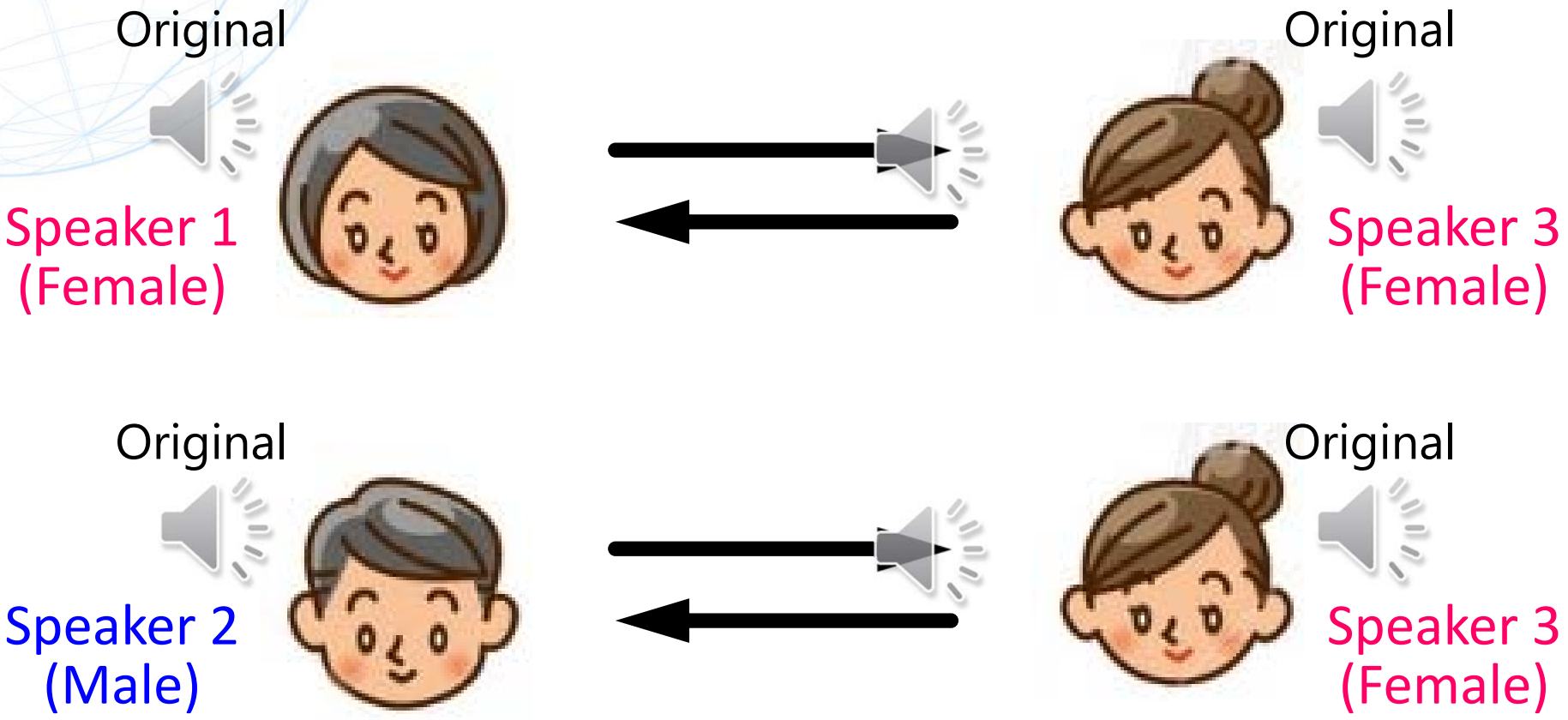
– 循環無矛盾ロス：$F\big(G(\mathbf{x})\big) \approx \mathbf{x}, G(F(\mathbf{y})) \approx \mathbf{y}$となるように誘導するロス
  - これを考慮することで変換前後でコンテンツが保持される傾向に

CycleGAN-VC [Kaneko&Kameoka2017]

- CycleGAN [Zhu+/Kim+/Yi+2017] を用いた非パラレルVC
  - 音響特徴量系列を画像（時間軸を空間座標軸）と見なしCycleGANを適用
  - 循環無矛盾ロスを考慮することで変換前後で言語情報が保持されるようにする



CycleGAN-VCによる音声変換例



多ドメインVCへの拡張

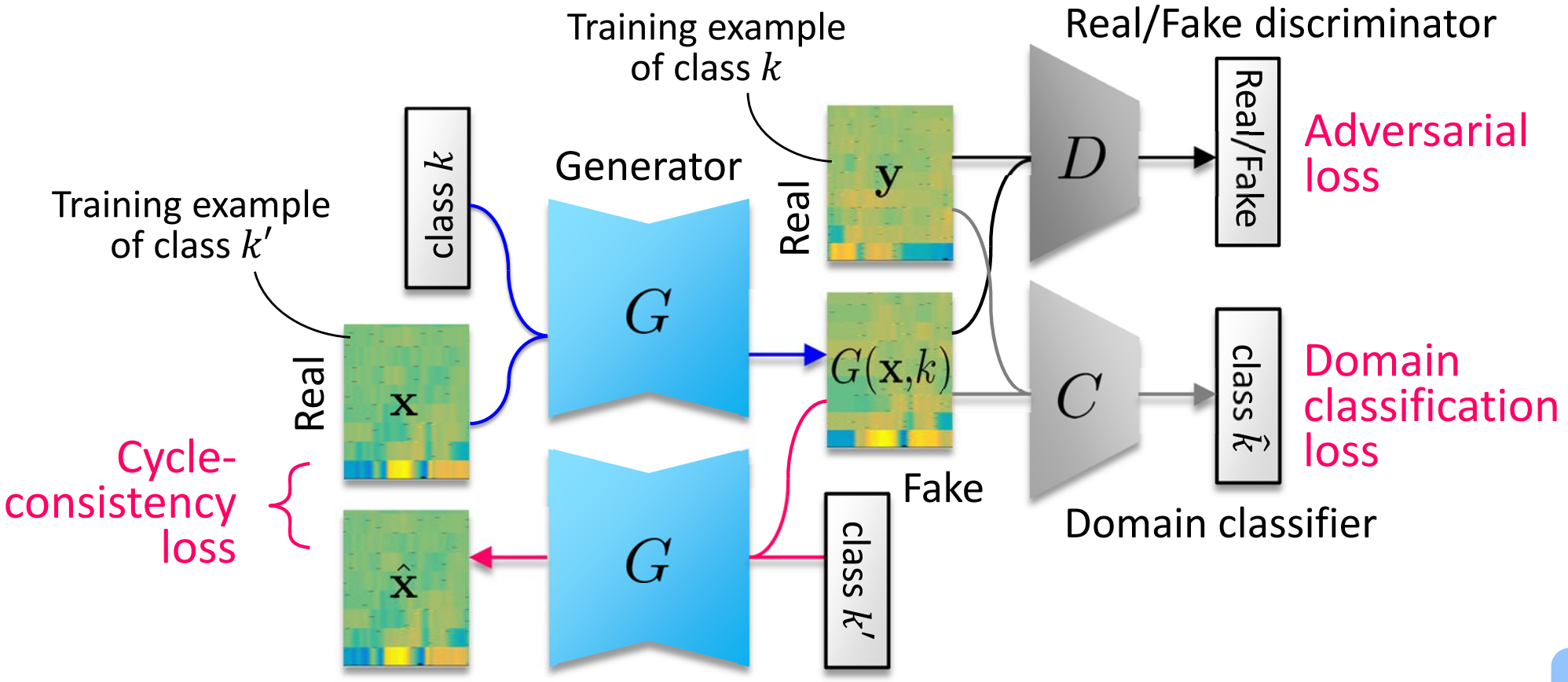
- CycleGANは２ドメイン間の変換しか学習できない
- Kドメイン間の変換を可能にするには$K(K-1)$個の変換器を学習する必要あり
- しかしこれは明らかに非効率
  （全ドメインで共有可能な特徴量が存在しうるため）

↓

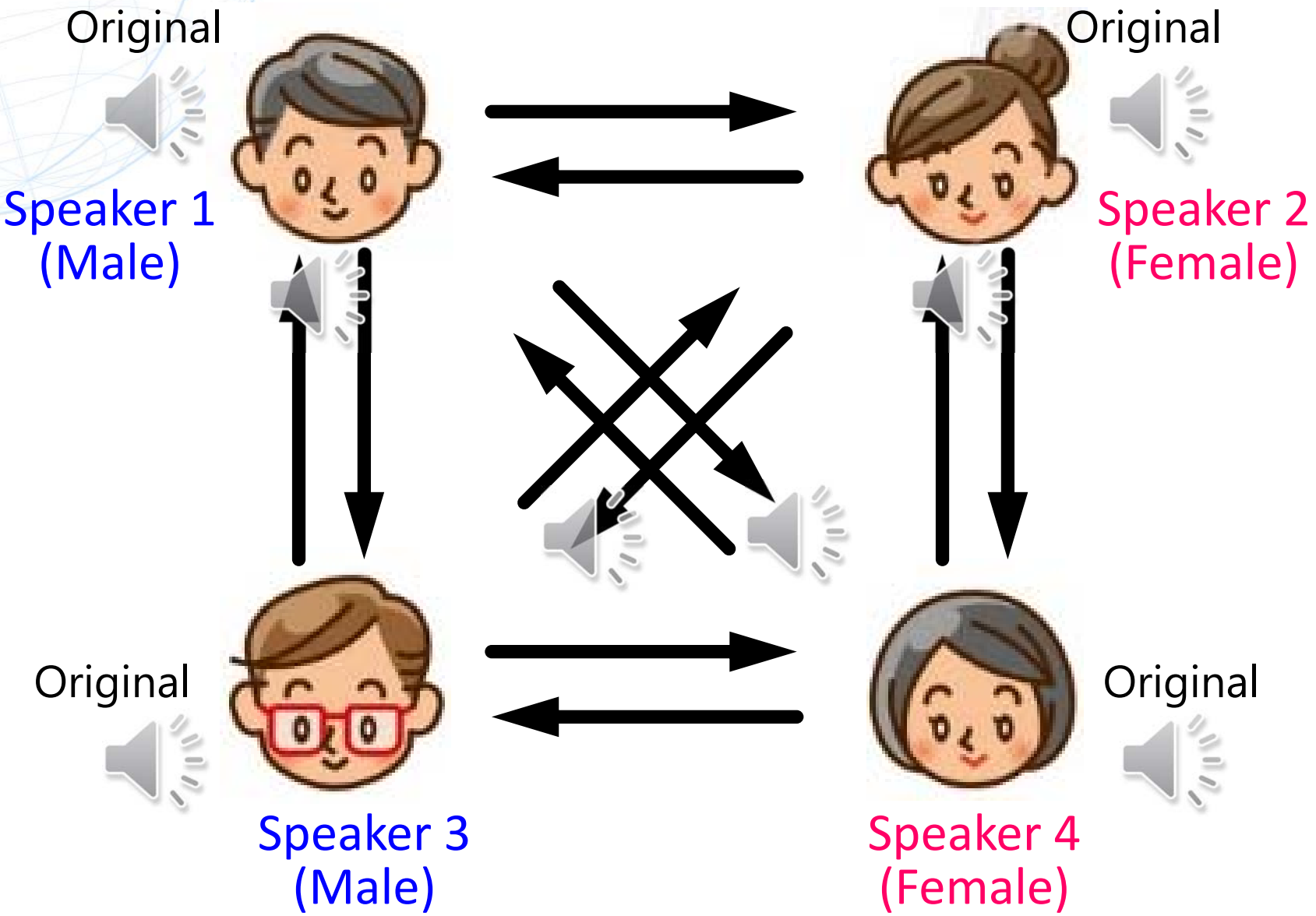
単一の生成器のみで多ドメイン間の変換を学習できないか？

StarGAN-VC [Kameoka+2018, Kaneko+2019]

- "StarGAN" [Choi+2018] を用いた非パラレル多ドメインVC
  - 単一の生成器Gを用いて多ドメイン間の変換を同時学習可能
  - $G(\mathbf{x}, k)$がDに"real"と識別され、Cにクラスkと識別されるようにGを学習
  - 循環無矛盾ロスを考慮することで変換前後で言語情報が保持されるようにする



StarGAN-VCによる音声変換例

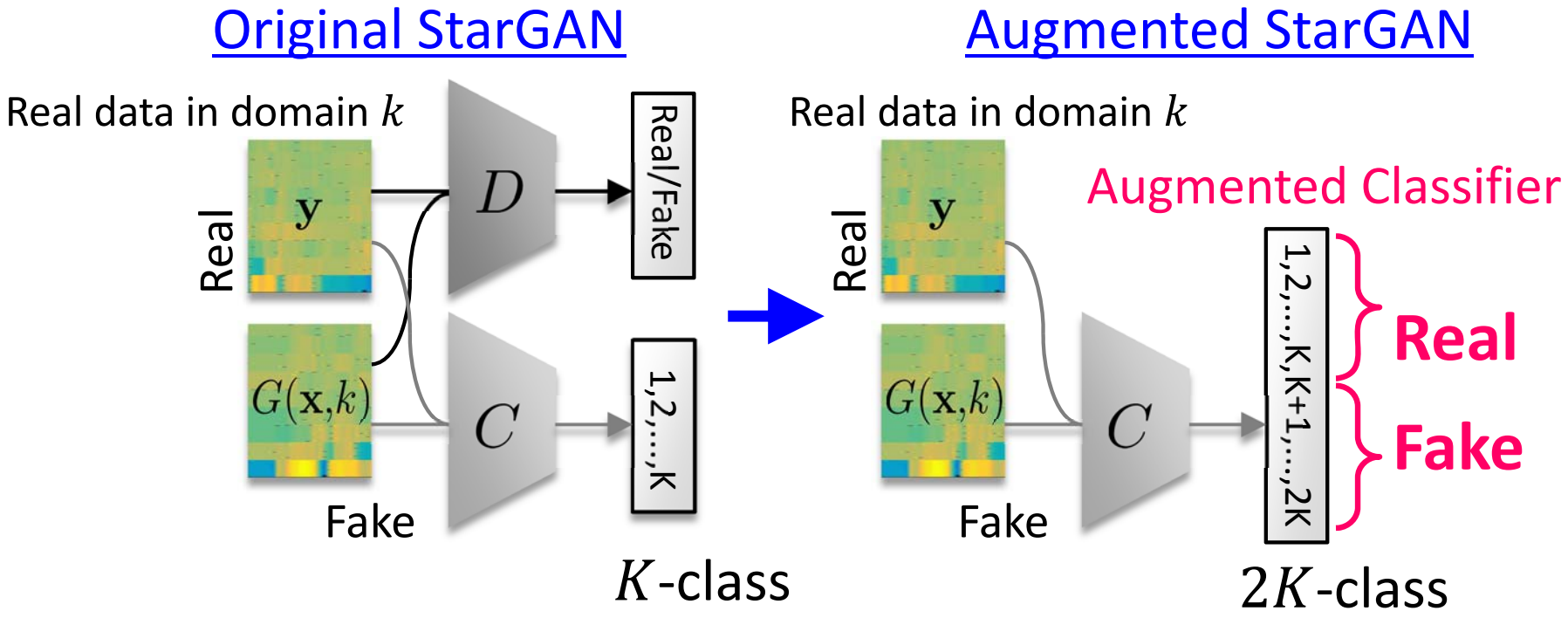


Augmented StarGAN [Kameoka+, to appear]

- ただしStarGANには改良の余地あり
  - 変換データが変換先ドメインの分布に従うことが保証されない？
    （分布フィッティングとしての解釈が不明瞭）
- 改良のための２つのアイディア

Augmented StarGAN [Kameoka+, to appear]

- ただしStarGANには改良の余地あり
  - –変換データが変換先ドメインの分布に従うことが保証されない？
    （分布フィッティングとしての解釈が不明瞭）
- 改良のための２つのアイディア
  - ① Real/Fake識別器とクラス識別器を一つに統合
    （クラス$1,\cdots,K$を実データ、クラス$K+1,\ldots,2K$を偽データと見なす）



Augmented StarGAN [Kameoka+, to appear]

- ただしStarGANには改良の余地あり
  - –変換データが変換先ドメインの分布に従うことが保証されない？
    （分布フィッティングとしての解釈が不明瞭）
- 改良のための２つのアイディア
  - ① Real/Fake識別器とクラス識別器を一つに統合
    （クラス$1,\cdots,K$を実データ、クラス$K+1,\ldots,2K$を偽データと見なす）
  - ② DとGを下記の目的関数が小さくなるように学習

$$V(D) = -\mathbb{E}_{k\sim p(k),\mathbf{y}\sim p(\mathbf{y}|k)}[\log D_k(\mathbf{y})] - \mathbb{E}_{k\sim p(k),\mathbf{x}\sim p(\mathbf{x})}[\log D_{K+k}(G(\mathbf{x},k))]$$
$$I(G) = -\mathbb{E}_{k\sim p(k),\mathbf{x}\sim p(\mathbf{x})}[\log D_k(G(\mathbf{x},k))] + \mathbb{E}_{k\sim p(k),\mathbf{x}\sim p(\mathbf{x})}[\log D_{K+k}(G(\mathbf{x},k))]$$

分布フィッティングとしての解釈

$$\underset{D_k(\mathbf{y})}{\operatorname{argmin}}\, V(D) = \frac{p(k)p_{\text{data}}(\mathbf{y}|k)}{\sum_k p(k)p_{\text{data}}(\mathbf{y}|k) + \sum_k p(k)p_G(\mathbf{y}|k)}$$

Substitution

$$I(G) = \mathbb{E}_{k\sim p(k)}\left[\mathrm{KL}\left(p_G(\mathbf{y}|k)\|p_{\text{data}}(\mathbf{y}|k)\right)\right]$$

Kullback-Leibler divergence

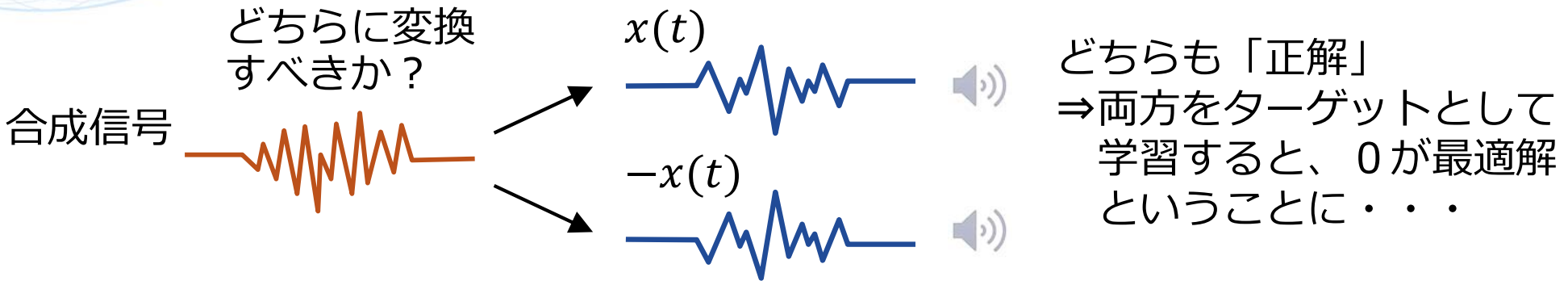
客観評価

- VCC2018データセットを用いた非パラレル音声（話者性）変換実験
  - 学習データ：81文（5分）、テストデータ：35文（2分）
- メルケプストラム歪み(MCD) [dB]

Speakers		Methods			
Source	Target	VAE [Hsu+2016]	VAE-WGAN [Hsu+2017]	**StarGAN**	**A-StarGAN**
SF1 (female)	SM1	7.66±0.123	7.70±0.122	**7.26±0.116**	7.27±0.129
	SF2	7.53±0.118	7.43±0.124	7.16±0.126	**6.98±0.147**
	SM2	8.06±0.143	8.04±0.145	7.67±0.120	**7.58±0.115**
SM1 (male)	SF1	8.25±0.104	8.20±0.128	7.69±0.099	**7.45±0.096**
	SF2	7.43±0.111	7.23±0.117	6.95±0.103	**6.66±0.114**
	SM2	7.92±0.106	7.82±0.103	7.24±0.089	**7.08±0.102**
SF2 (female)	SF1	7.97±0.127	7.83±0.121	7.59±0.102	**7.40±0.107**
	SM1	7.38±0.108	7.37±0.097	6.91±0.121	**6.83±0.126**
	SM2	7.92±0.122	7.78±0.109	7.49±0.116	**7.48±0.103**
SM2 (male)	SF1	8.33±0.148	8.20±0.158	7.83±0.152	**7.67±0.139**
	SM1	7.73±0.138	7.66±0.142	7.07±0.122	**6.97±0.127**
	SF2	7.74±0.135	7.65±0.137	7.27±0.143	**6.98±0.154**
All pairs		7.83±0.045	7.74±0.046	7.35±0.044	**7.19±0.046**

波形ポストフィルタリング

- 合成音声を実音声に変換
- 単純な回帰分析問題として扱って良いか？
  - 合成信号と実信号は１対１対応しない（位相はその時々で変化するため）

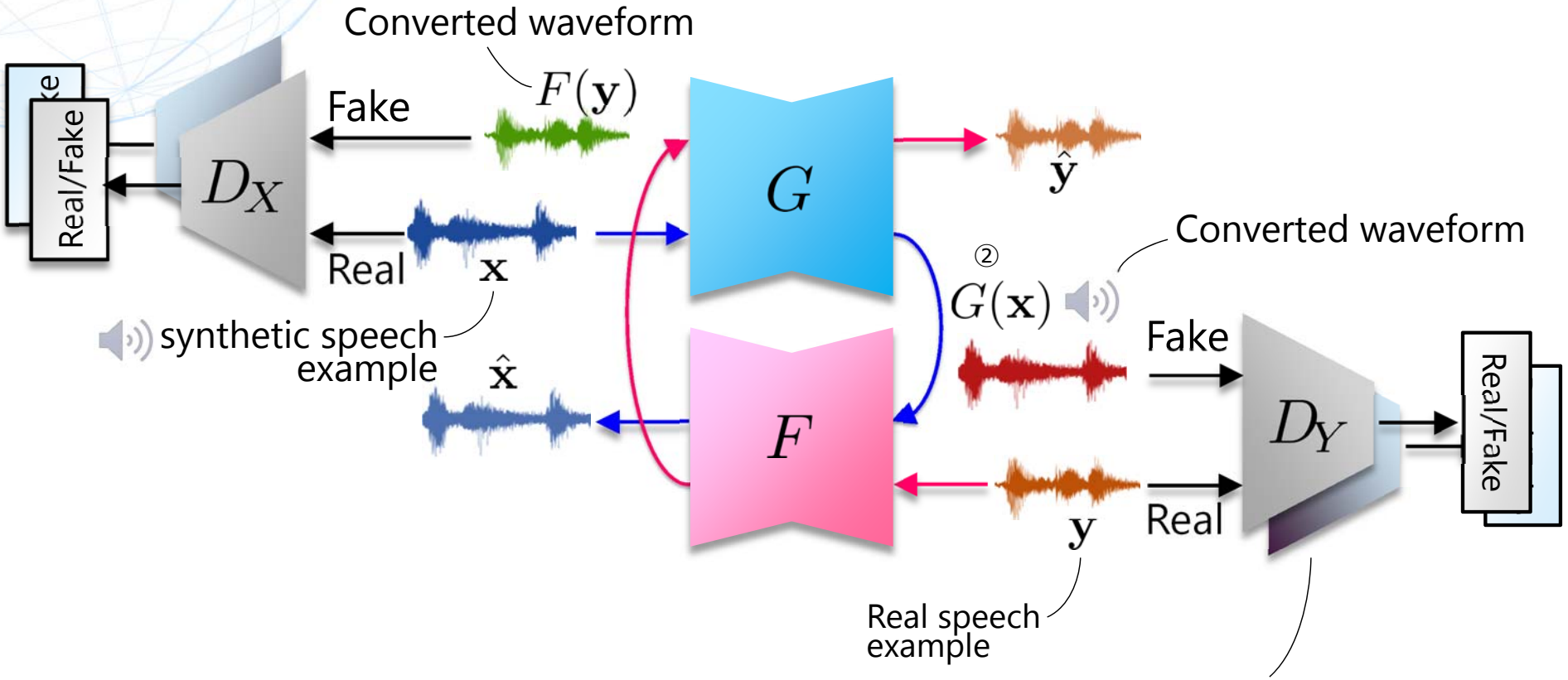


  - よって波形ポストフィルタリング問題を普通の回帰分析問題としては扱えない

  - どうアプローチするか？

WaveCycleGAN [Tanaka+2018]

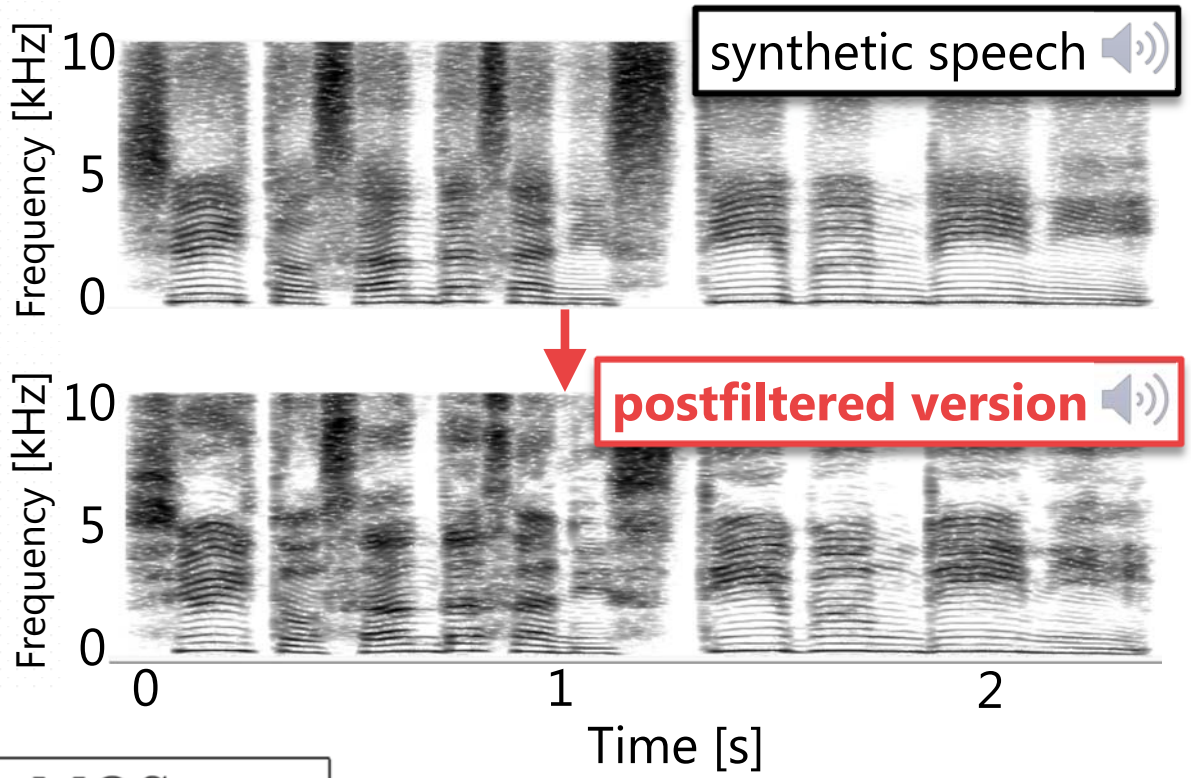
- 合成音声と実音声の信号を非パラレル学習データと見なし、CycleGANを適用



メルスペクトログラム領域やMFCC領域において
入力波形が実音声らしいかをチェックする識別器

WaveCycleGANによる音声変換例

- DNN音声合成[Zen+2013]の合成音声に適用した例



- 他の波形生成法とのMOS値の比較

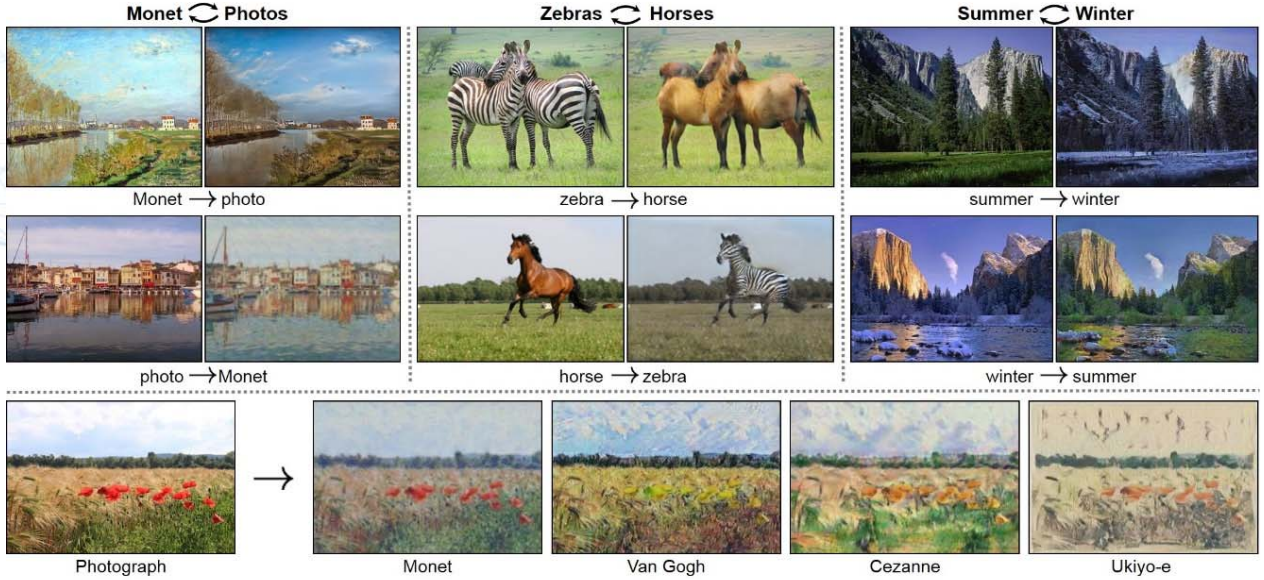
System	MOS
Recorded	4.590 ± 0.082
WORLD (from mel-cepstrum)	3.124 ± 0.150
Griffin-Lim [Griffin&Lim1984]	3.300 ± 0.143
open WaveNet [van den Oord+2016]	3.657 ± 0.162
WaveGlow [Prenger+2018]	3.443 ± 0.164
WaveCycleGAN (Ours)	$\mathbf{4.023} \pm 0.124$

unofficial implementation available on GitHub

official implementation provided by the authors

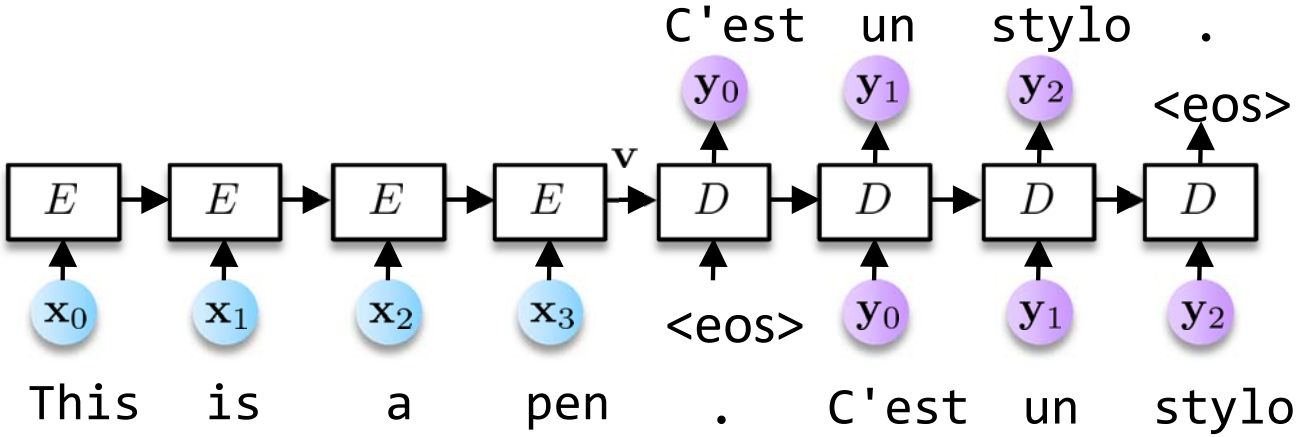
我々のアプローチ

- 画像変換アプローチ → 音響特徴量変換
  波形ポストフィルタリング



[Zhu+2017]

- 系列変換(Sequence-to-sequence)アプローチ → 音響特徴量変換



従来のパラレル／非パラレルVCにおける限界

- 超分節的特徴の変換が困難
  - –従来方式による英語アクセントの変換効果

入力 → sprocket (GMM-VC) [Kobayashi+2018] → 変換 ターゲット

Indian English (male) → US English (female)

  - –従来方式によるF_0パターンの予測性能
    - GMM-VCによる無喉頭音声強調

入力（食道音声） → 変換 [Doi+2010]

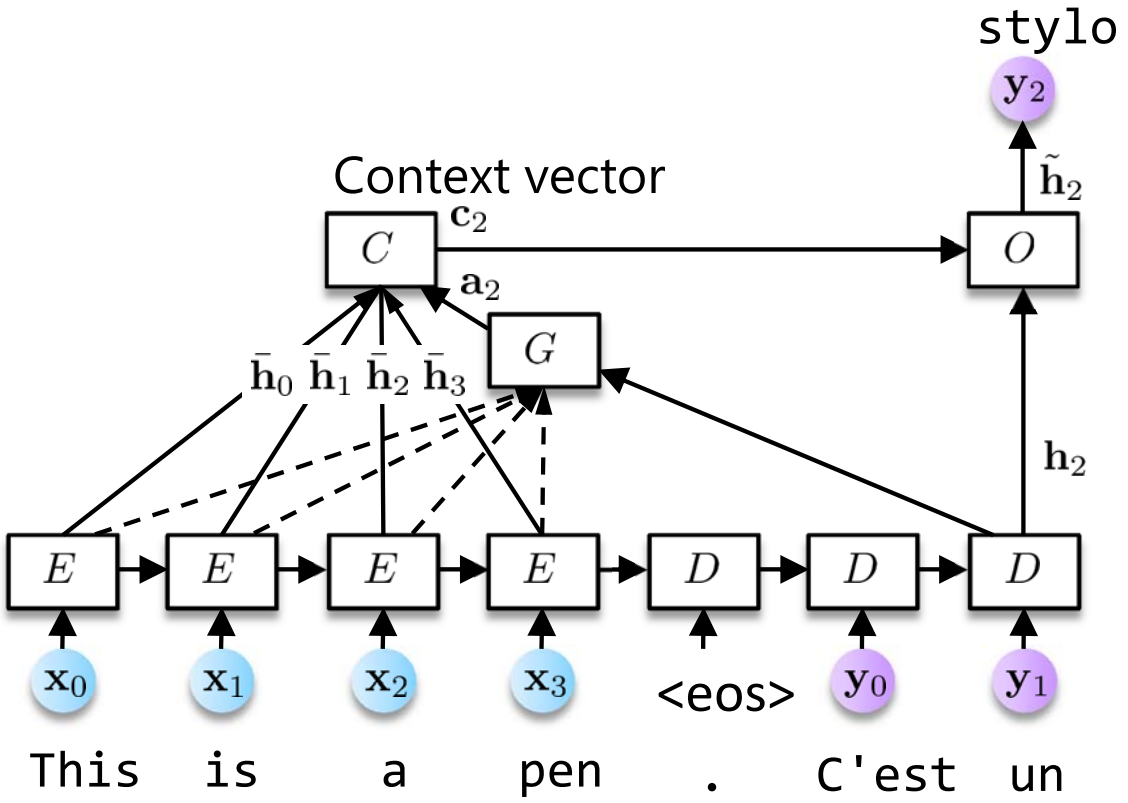
入力（電気喉頭音声） → 変換 [Nakamura+2010]

電気喉頭

Acknowledgement: We thank Prof. Tomoki Toda (Nagoya University) for providing us with the voice samples and the photo in this slide.

系列変換(S2S)学習 [Sutskever+2014]

- 系列から系列への変換を行う枠組
- 機械翻訳、音声認識(ASR)、テキスト音声合成(TTS)などの幅広いタスクにおける有効性が既に示されている
- エンコーダ／デコーダ構造
  - エンコーダは入力系列を内部表現に変換
  - デコーダは内部表現をもとに出力系列を逐次生成
- 注意機構 [Luong+2015]
  - デコードの際、入力系列のどの要素に注目すべきかを学習する機構



S2Sアプローチによる VC

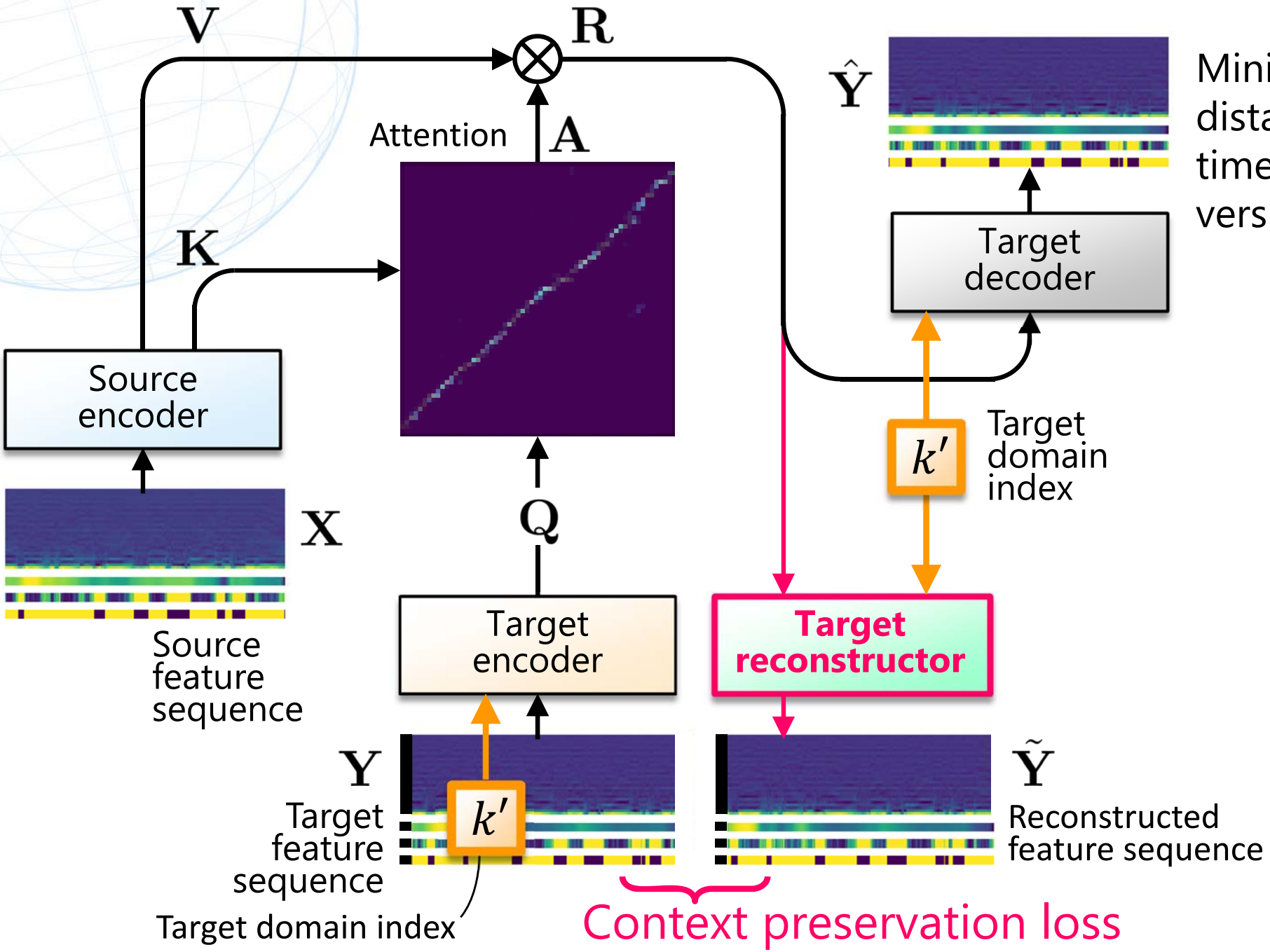
- 関連研究
  - 音素事後確率予測(ASR)⇒音素事後確率系列間のS2S⇒波形生成 [Miyoshi+2017]
  - ASRを援用した特徴量系列間のS2S（ASRシステムで得られるボトルネック特徴量を入力ベクトルに追加） [Zhang+2018]
  - 特徴量系列間のS2SモデルとASRモデルのマルチタスク学習 [Biadsy+2019]

↓

いずれもASRシステムまたはテキストラベルを必要とする

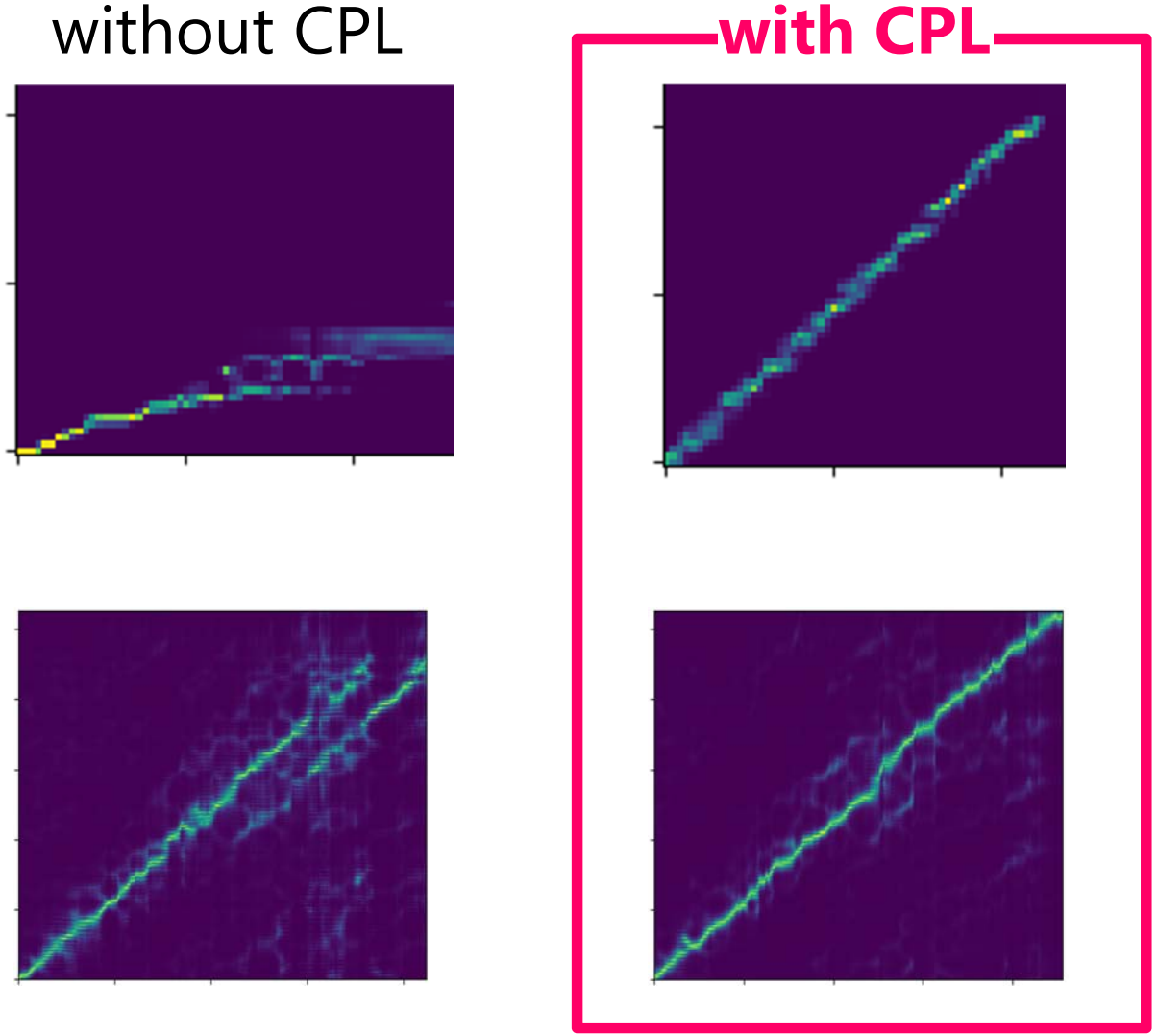
- 我々のアプローチ [Tanaka+2018, Kameoka+2018]
  - コンテキスト保持機構の導入により、ASRシステムやテキストラベルを用いずに学習可能
  - StarGAN-VC同様、多ドメイン変換が可能
  - Tacotronベースのネットワーク[Tanaka+2018]と全層畳み込みネットワーク[Kameoka+2018]による実装

S2S-VCアーキテクチャ



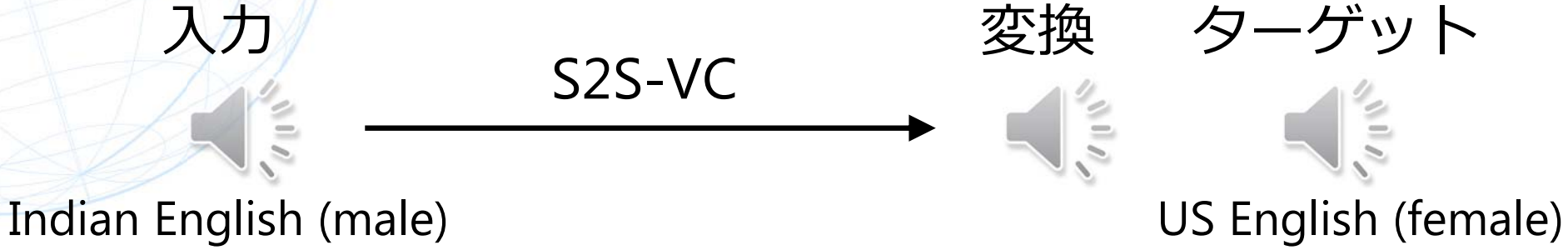
コンテキスト保持機構の効果

- コンテキスト保持機構を考慮せず学習したモデルと考慮して学習したモデルで予測された注意(Attention)行列の例

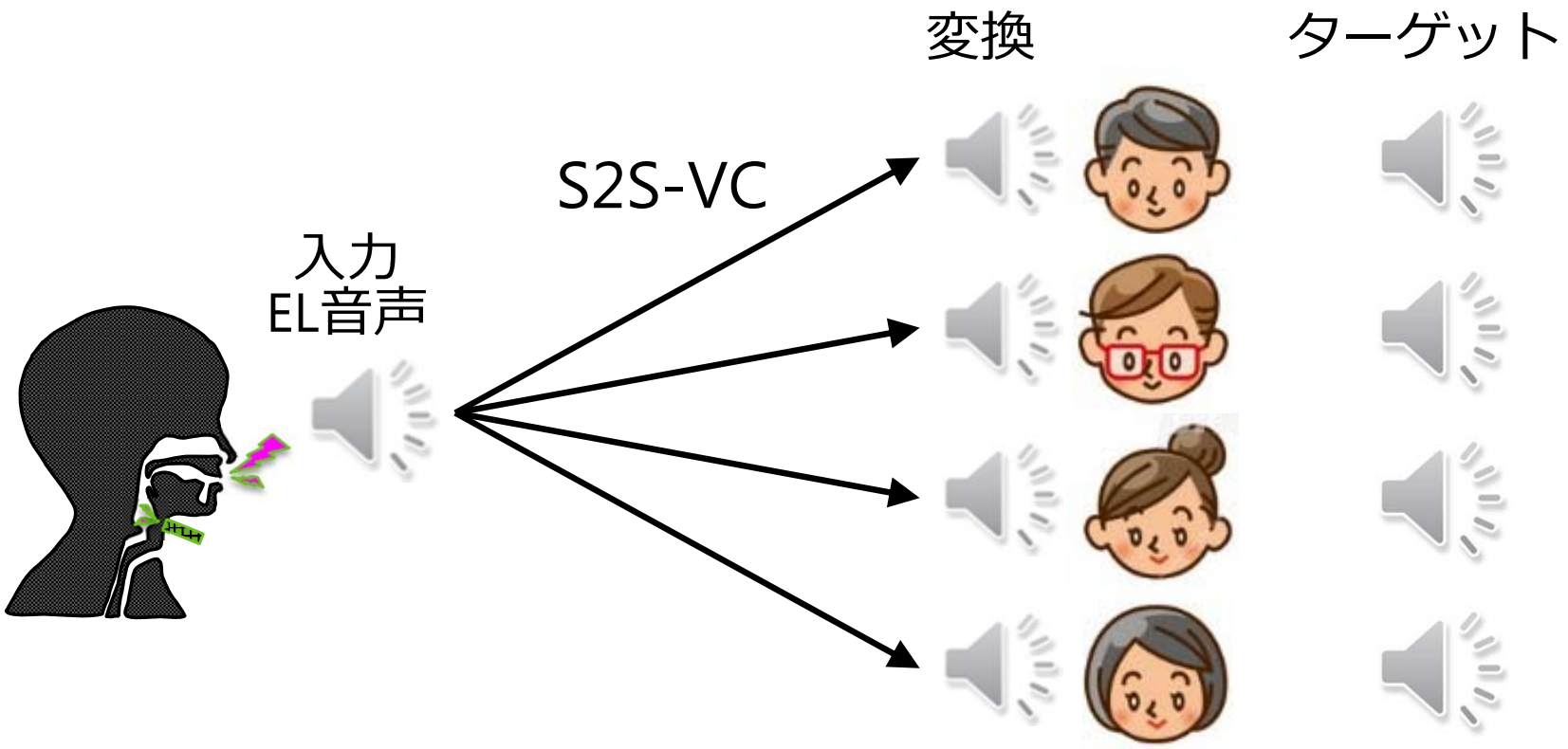


英語アクセント変換と電気喉頭音声変換の例

- 英語アクセント変換例

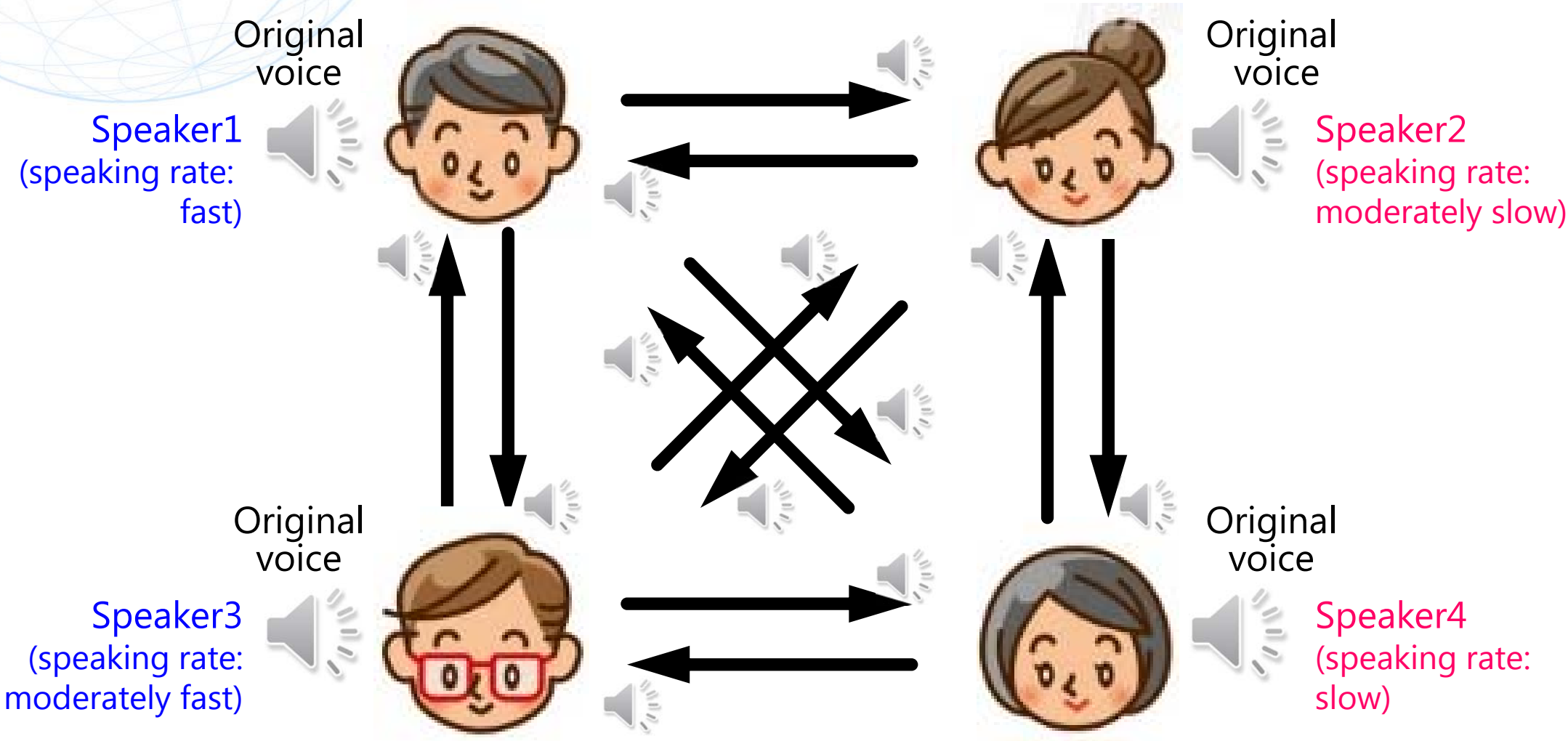


- 電気喉頭音声変換例



Multi-speaker identity conversion

- Audio examples obtained with multi-domain conversion model, trained on 450 sentences per speaker in ATR Japanese database



Real-time VC demo

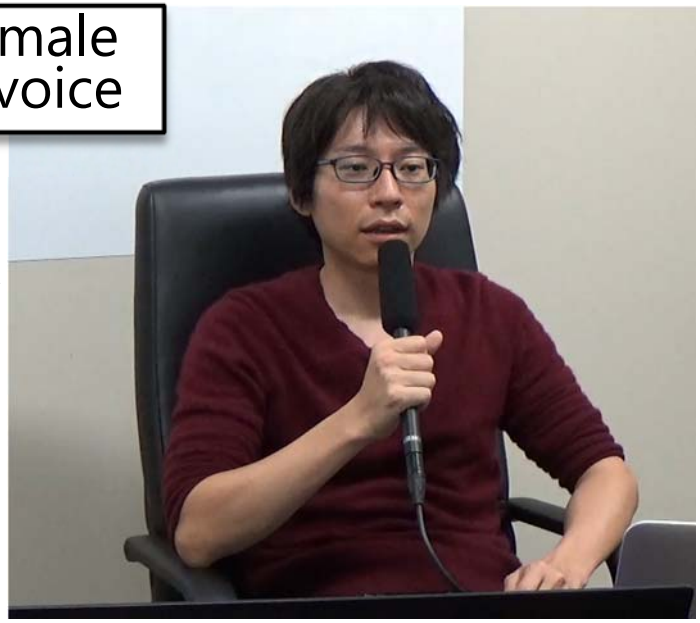
- S2S-VC followed by WaveCycleGAN

Converted

female voice

Original voice

male voice



まとめ

- 音声変換(VC)は音声コミュニケーションにおけるバリアやストレスを解消できる可能性を秘めた重要技術
- 特に下記要件が重要
  - 高品質であること
  - 学習が効率的であること（少量データ，非パラレルデータで動く）
  - リアルタイムに動作すること
  - 声質だけでなく超分節的特徴などの柔軟な変換が可能であること
- 以上の要件を意識しながらいくつかの手法 (CycleGAN-VC, StarGAN-VC, WaveCycleGAN, S2S-VC, etc.) を提案

- 謝辞
  - 本研究は右記の有能な同僚と共に進めました
  - Chainerには大変お世話になりました

金子卓弘

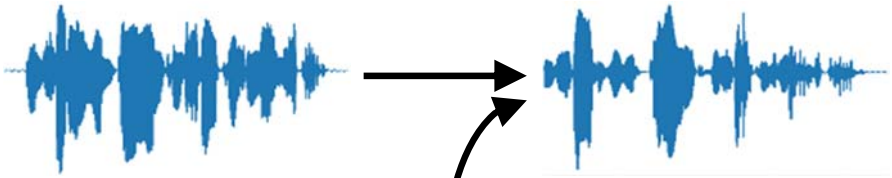
田中宏

北条伸克

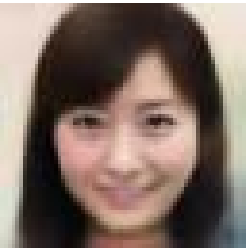
おまけ：クロスモーダルVC [Kameoka+2019]

- 声と顔には相関がある。この相関を手がかりに、（１）与えられた顔画像の印象に合った声質に入力音声を変換したり、（２）入力音声に合った印象の顔画像を生成したりできないか？

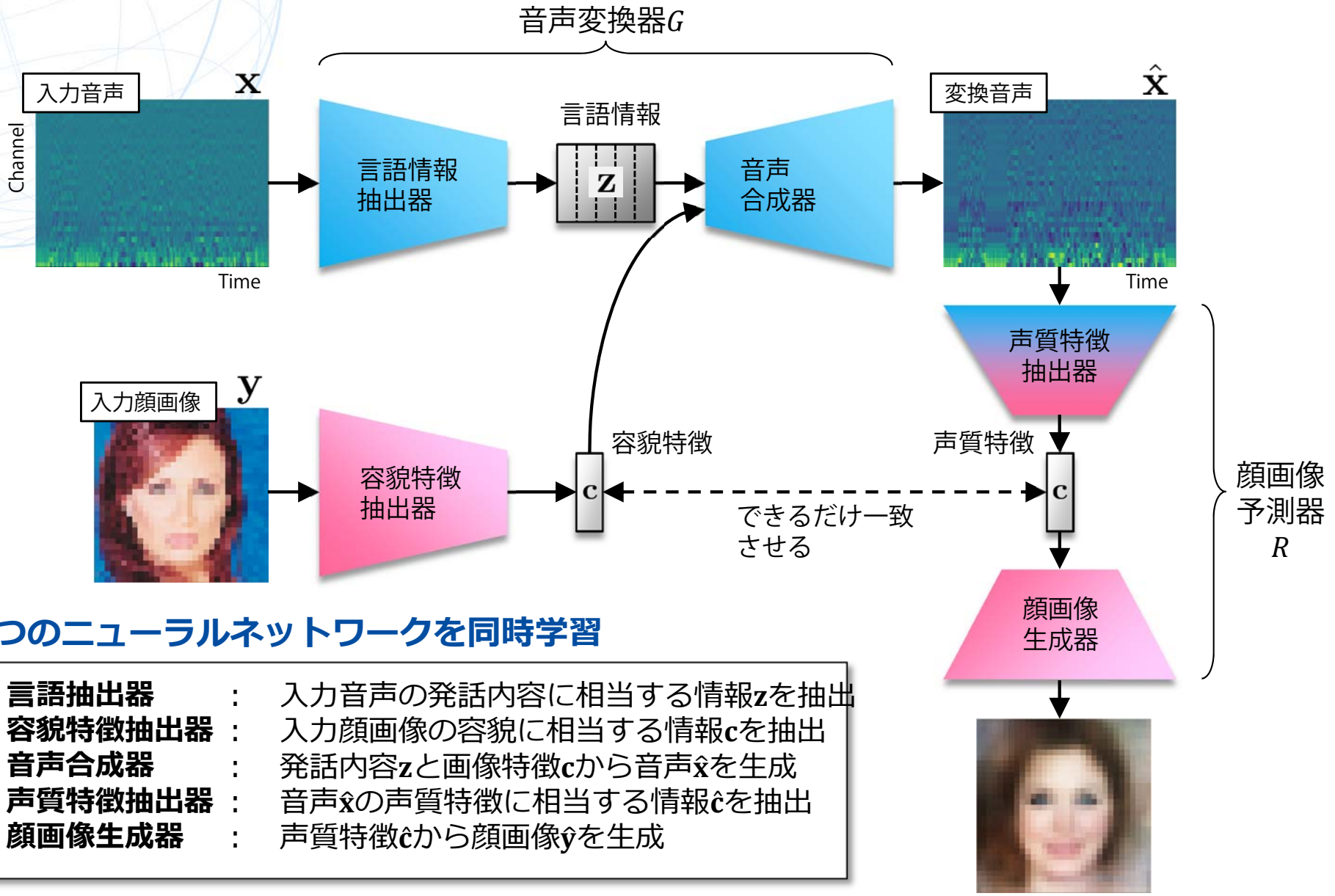
（１）入力音声の声質を、
入力顔画像に合わせて変換する



（２）入力音声のみから
話者の顔を予測する



おまけ：クロスモーダルVC [Kameoka+2019]

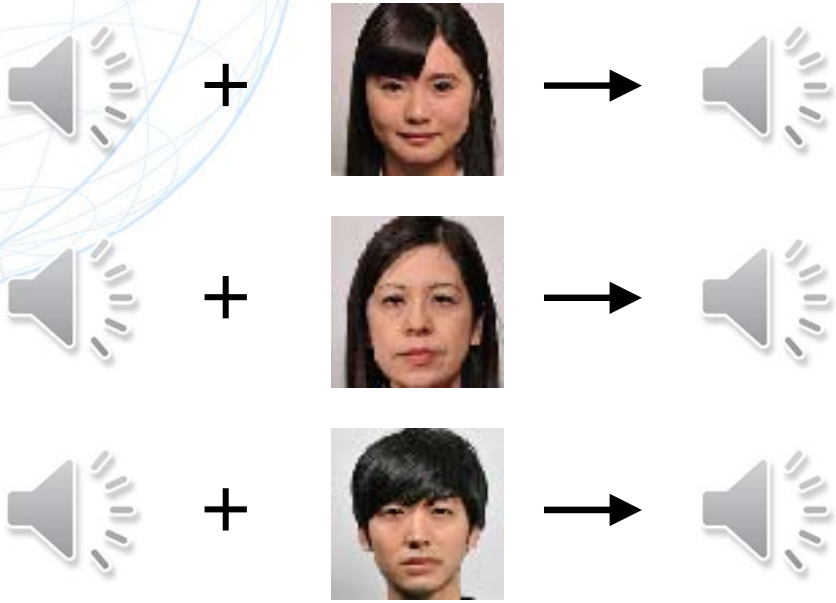


5つのニューラルネットワークを同時学習

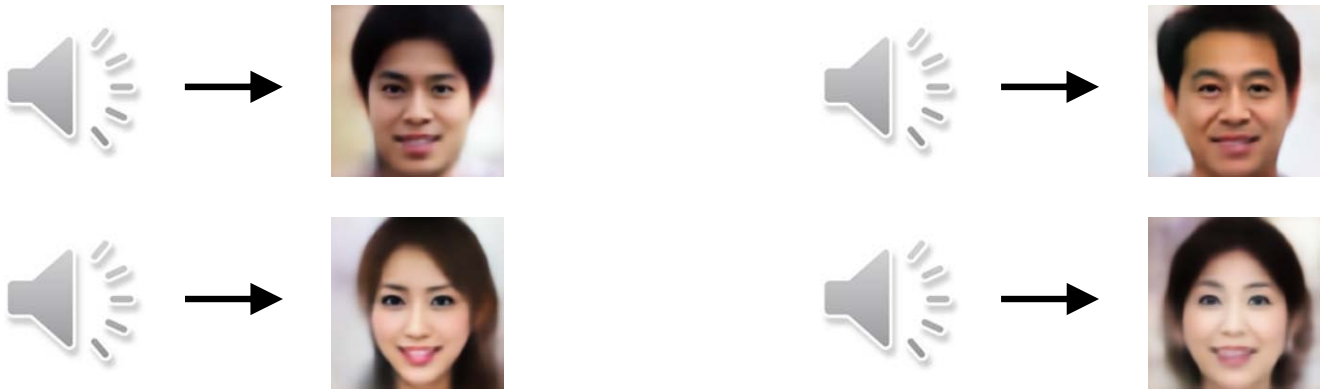
言語抽出器	：	入力音声の発話内容に相当する情報$\mathbf{z}$を抽出
容貌特徴抽出器	：	入力顔画像の容貌に相当する情報$\mathbf{c}$を抽出
音声合成器	：	発話内容$\mathbf{z}$と画像特徴$\mathbf{c}$から音声$\hat{\mathbf{x}}$を生成
声質特徴抽出器	：	音声$\hat{\mathbf{x}}$の声質特徴に相当する情報$\hat{\mathbf{c}}$を抽出
顔画像生成器	：	声質特徴$\hat{\mathbf{c}}$から顔画像$\hat{\mathbf{y}}$を生成

音声予測例、顔画像予測例

- 音声変換



- 顔画像予測



まとめ

- 音声変換(VC)は音声コミュニケーションにおけるバリアやストレスを解消できる可能性を秘めた重要技術
- 特に下記要件が重要
  - 高品質であること
  - 学習が効率的であること（少量データ，非パラレルデータで動く）
  - リアルタイムに動作すること
  - 声質だけでなく超分節的特徴などの柔軟な変換が可能であること
- 以上の要件を意識しながらいくつかの手法 (CycleGAN-VC, StarGAN-VC, WaveCycleGAN, S2S-VC, etc.) を提案
- 謝辞
  - 本研究は、金子卓弘氏、田中宏氏、北条伸克氏らと共に行いました
  - Chainerには大変お世話になりました

金子卓弘

田中宏

北条伸克