IBIS2018 企画セッション 招待講演

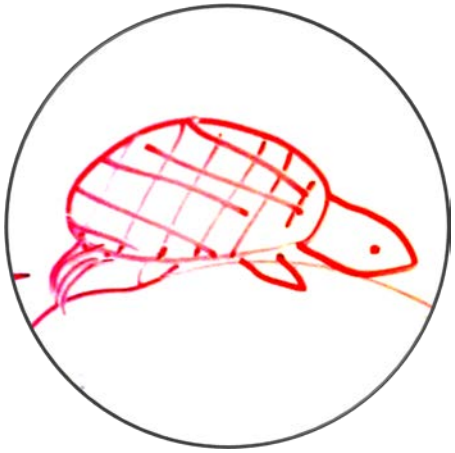
音声分野における敵対的学習の可能性と展望

高道 慎之介 (東京大学)･亀岡 弘和 (NTT)

自己紹介

高道 慎之介
Shinnosuke Takamichi

東京大学 助教
@forthshinji



亀岡 弘和
Hirokazu Kameoka

NTT 主任研究員
@kamepong

本発表のテーマ

音声分野から見た敵対的学習 (GAN)

次世代音声処理を見据えた展開

機械学習に基づく音声処理

~音声の合成・変換・強調を例にして~

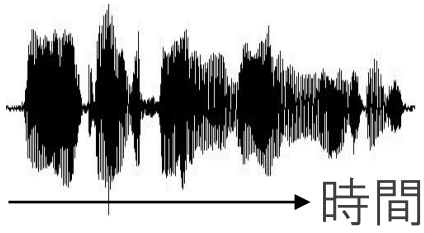
音声生成過程

音色の付与

口や舌を動かして，
音色をつける！

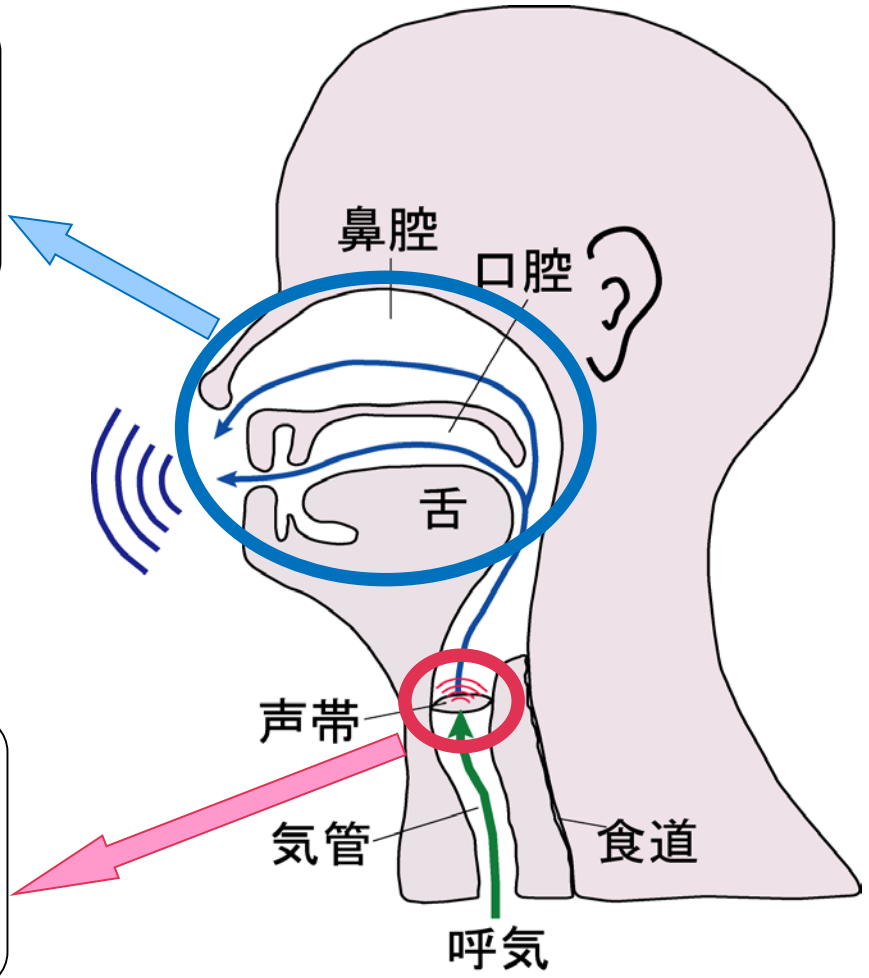
畳み込むと…

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！



音声の合成･変換･強調

音声合成 (Text-To-Speech: TTS)

– テキストなどから音声を合成する技術 (例：ドラえもんの発声機構)

音声変換 (Voice Conversion: VC)

– 入力音声を異なる音声に変える技術 (例：名探偵コナンの変声器)

音声強調 (Speech Enhancement: SE)

– 騒音下音声などの音声成分を強調する技術 (例：聖徳太子の耳)

身体･文化･時空間の違いを超えた音声コミュニケーションのために
音声を人工的に生成する技術

音声変換デモ

(Conventional)

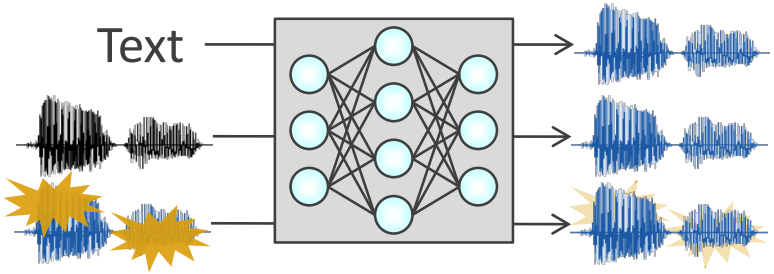
HOYA SPEECH
Pioneering Voice Technology

http://voicetext.jp/voiceactor/

音声の合成・変換・強調のための機械学習的アプローチ

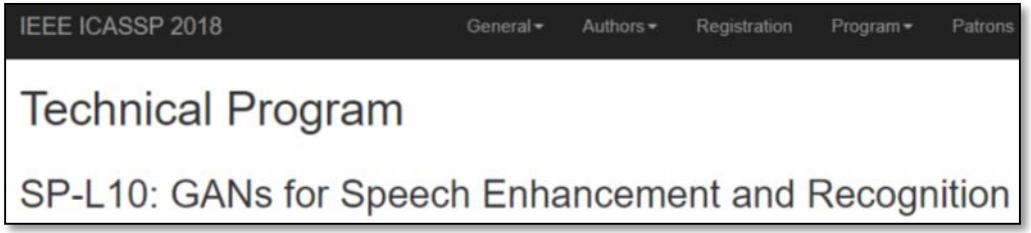
音声信号は，音信号の中でもビッグデータ性が比較的強い信号

- DNNなどによる機械学習的アプローチが有効
- (逆に，空間情報を扱う音響信号には統計的信号処理が有効)



GANの登場により，合成・変換・強調品質が飛躍的に向上

- 2017年春の初登場 (後述) から爆発的に使用されるようになり，最近では，GANを使った音声処理のセッションも組まれる



https://2018.ieeeicassp.org/Papers/PublicSessionIndex3.asp?Sessionid=1143

音声合成・変換におけるGANの登場

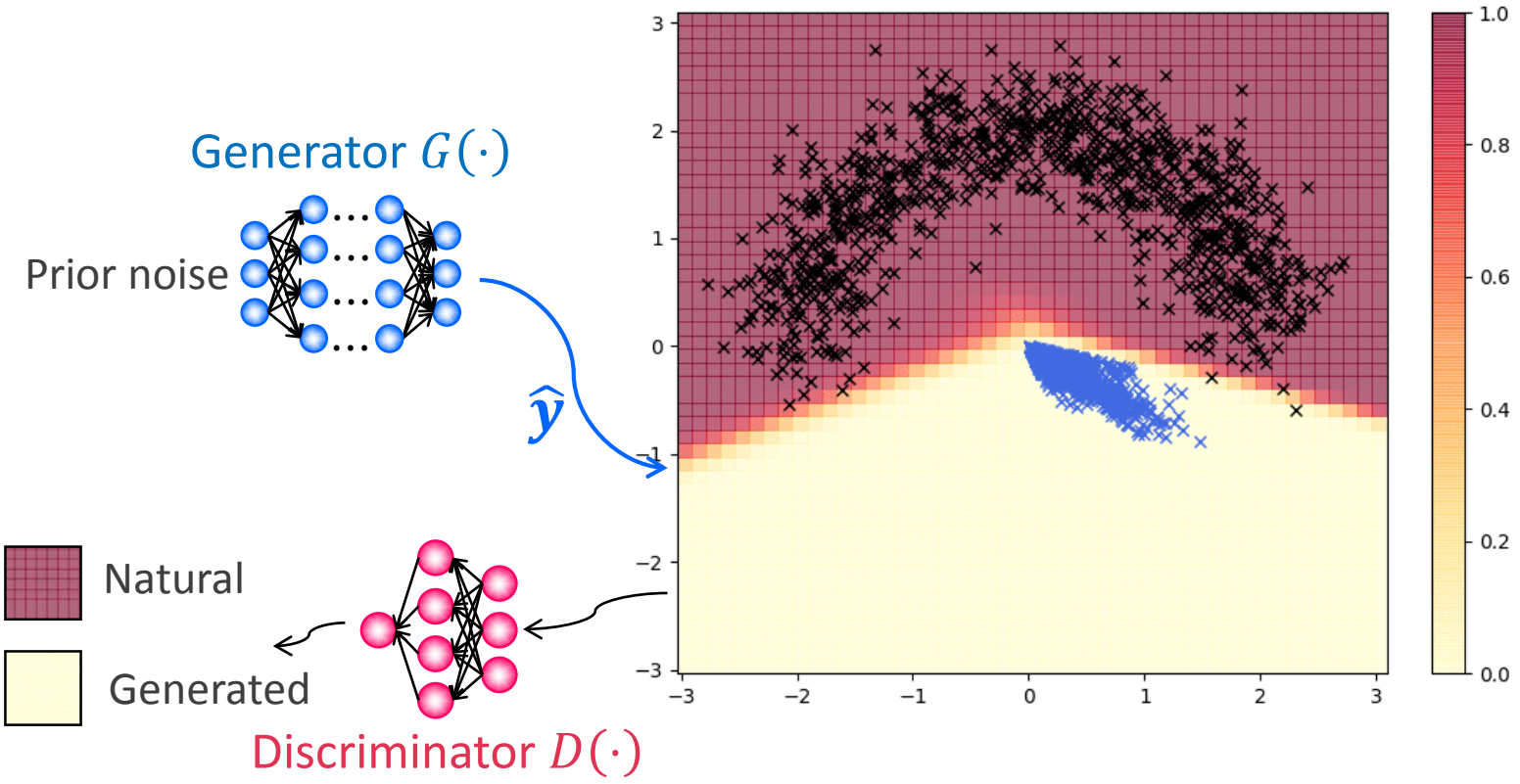
Generative Adversarial Network (GAN)

[Goodfellow14]

Generative adversarial network

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

– 合成器と，自然／合成音声を識別する識別器を敵対



GAN の数学的解釈

[Goodfellow14]

Minimaxゲームによる学習

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

分布フィッティングとしての解釈

– $V(G,D)$を最小化するD

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

– $\hat{D}$を代入した$V(G,D) = G$から見た目的関数

$$C(G) = \min_D V(G,D) =$$

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

p_{data}とp_Gの負のJensen-Shannon ダイバージェンス

音声の生成タスクにおいて GANを利用した最初の国際会議論文

GENERATIVE ADVERSARIAL NETWORK-BASED POSTFILTER FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Takuhiro Kaneko[†], *Hirokazu Kameoka*[†], *Nobukatsu Hojo*[‡]
Yusuke Ijima[‡], *Kaoru Hiramatsu*[†], *Kunio Kashino*[†]

[STFT版は2017/09]

TRAINING ALGORITHM TO DECEIVE ANTI-SPOOFING VERIFICATION FOR DNN-BASED SPEECH SYNTHESIS

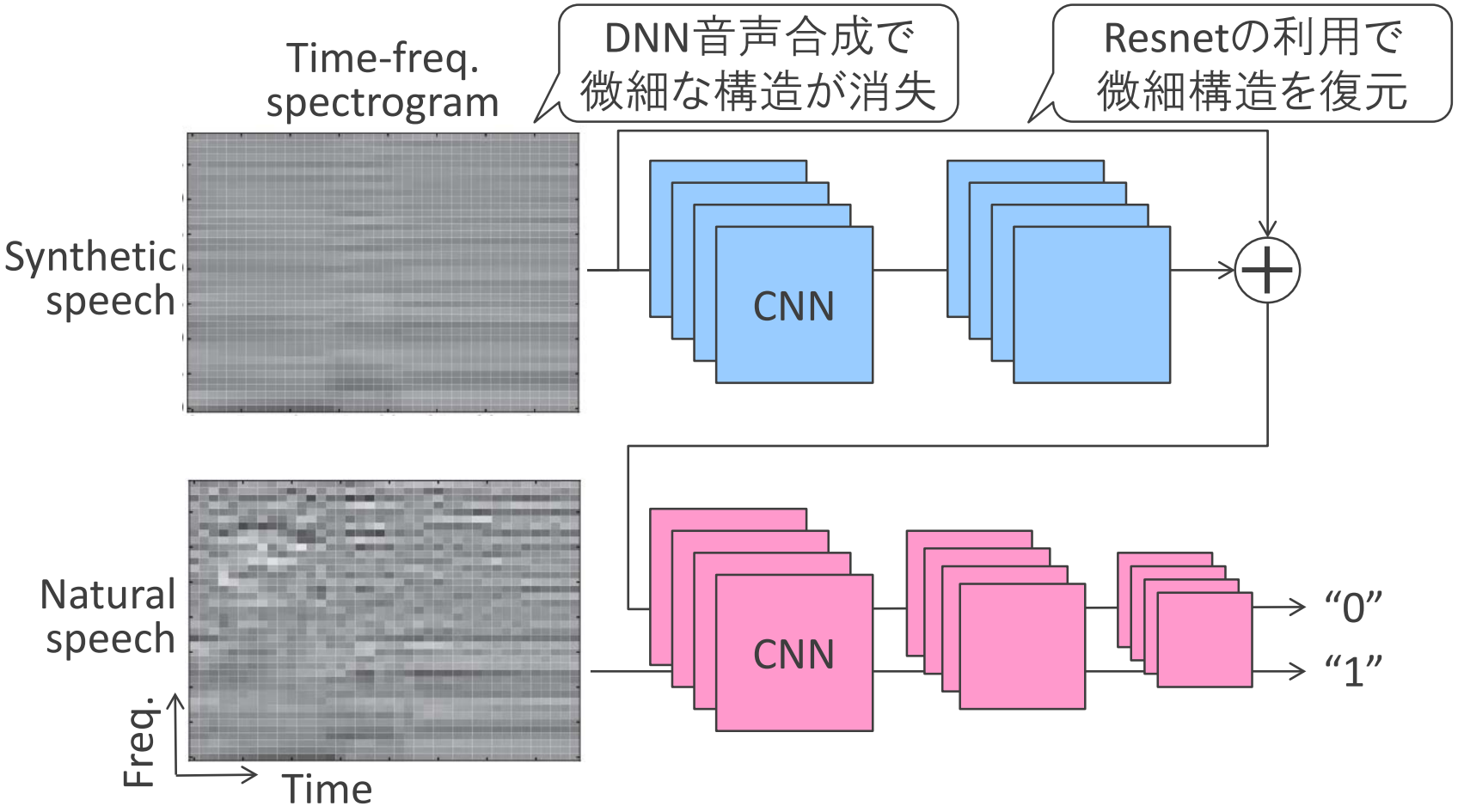
Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari

[ジャーナル版は2018/01]

同じ会議の同じセッションで同じ目的 (合成音声の高品質化) を持った2つの論文が登場. しかし philosophy は全く異なる.

GAN-based post-filter:
音声信号を画像とみなしてGANを利用

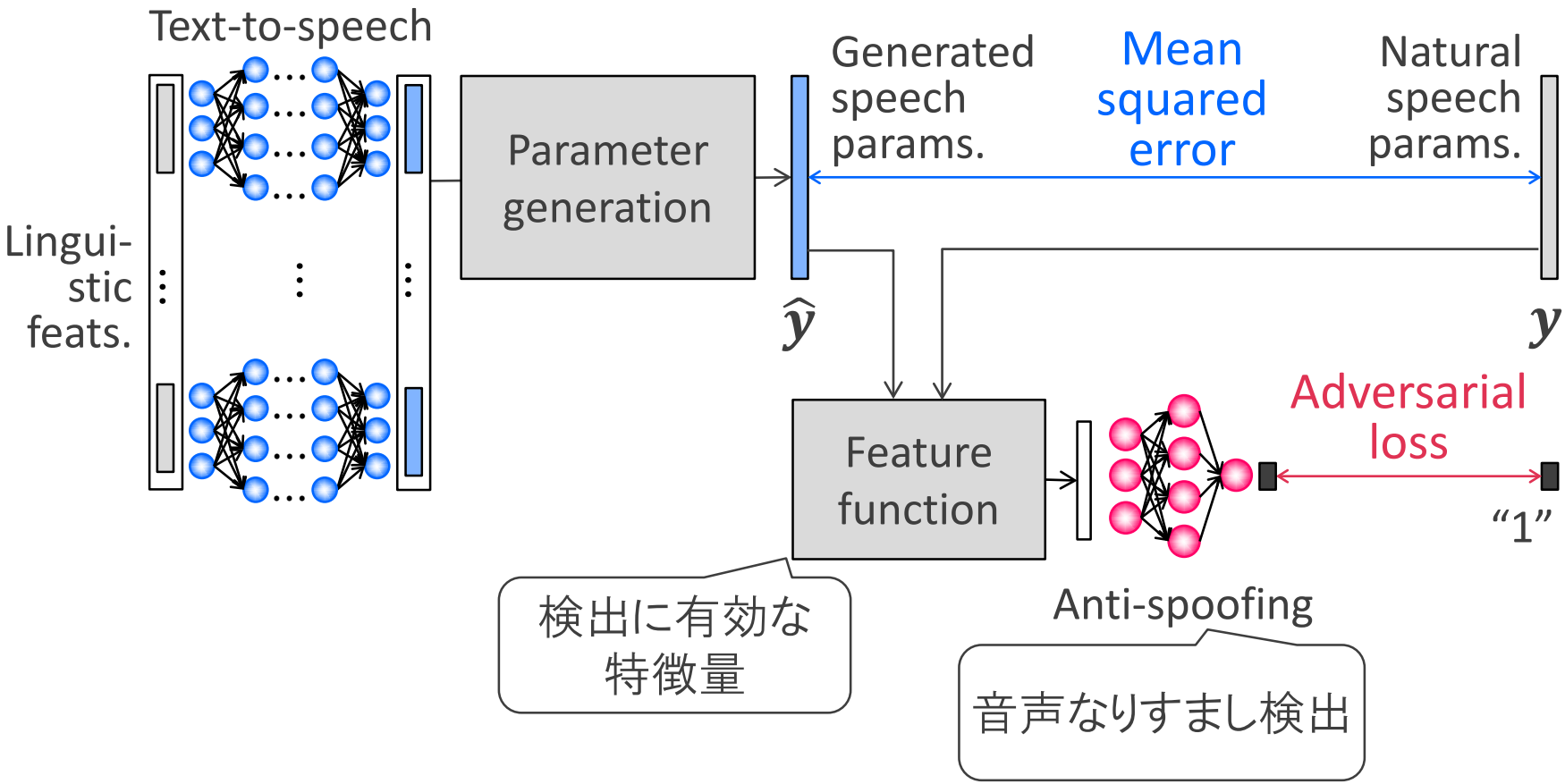
[Kaneko17][Kaneko17-2]



画像分野における知見を積極的にGANに導入して高品質化！

Training algorithm to deceive anti-spoofing: 音声なりすまし検出を騙すためにGANを利用

[Saito17][Saito18]



音声なりすまし検出セキュリティを積極的に騙して高品質化！

音声変換における
GANのさらなる応用発展

深層生成モデルにおけるGANの位置づけ

音声の生成モデル化の難しさ

– いかに長い時系列データの同時分布をモデル化するか？

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$$

深層生成モデルの例

– 自己回帰生成ネット (Autoregressive Generative Net)

- 同時分布をfactorize： $p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$
- 各条件付分布 $p(\mathbf{x}_t|\mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$ をNNでモデル化
- 音声分野では「WaveNet」[van den Oord+2016]の登場により脚光を浴びる
- 学習は効率的な一方で生成は非効率的（逐次計算が不可避のため）

– **GAN**

- $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ を直接モデル化することなくGとDの敵対的学習により $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ に従うサンプルを生成するGを得ることができる
- 学習は容易ではない一方で生成は効率的（CNNを用いれば並列計算可）
  ⇒リアルタイムシステムにおいて威力を発揮する？

音声変換 (Voice conversion; VC)

様々な重要な応用

- 音声生成機能拡張 [Toda 14]
  - 発声障碍者の声をより自然な声へ

    食道音声 → [Doi+2010]

    電気音声 → [Nakamura+2010]

  - 聞き取れないほど微弱な声を明瞭な声へ

    体内伝導音声 → [Toda+2012]

    （テレパシーのような音声コミュニケーション）

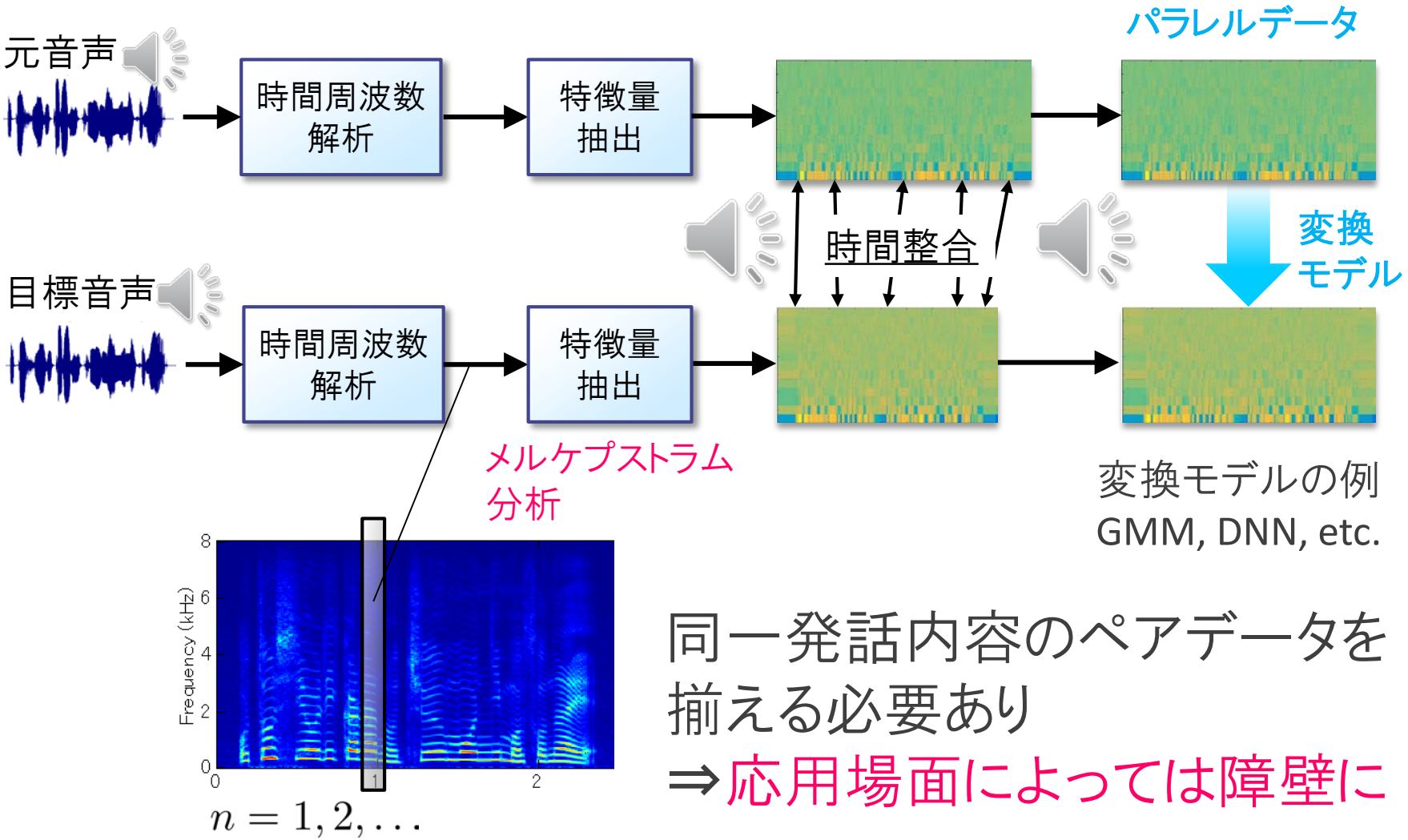
- 非母語話者の発音修正 [Kaneko&Kameoka+2017]
- アバターの話者性変換
- etc.

→いずれの応用もリアルタイム性が重要

謝辞：本ページの音声サンプルと写真は戸田智基教授（名大）にご提供いただいた。

VCの回帰問題としての定式化

「パラレルデータ」の構築，特徴量変換モデルの学習

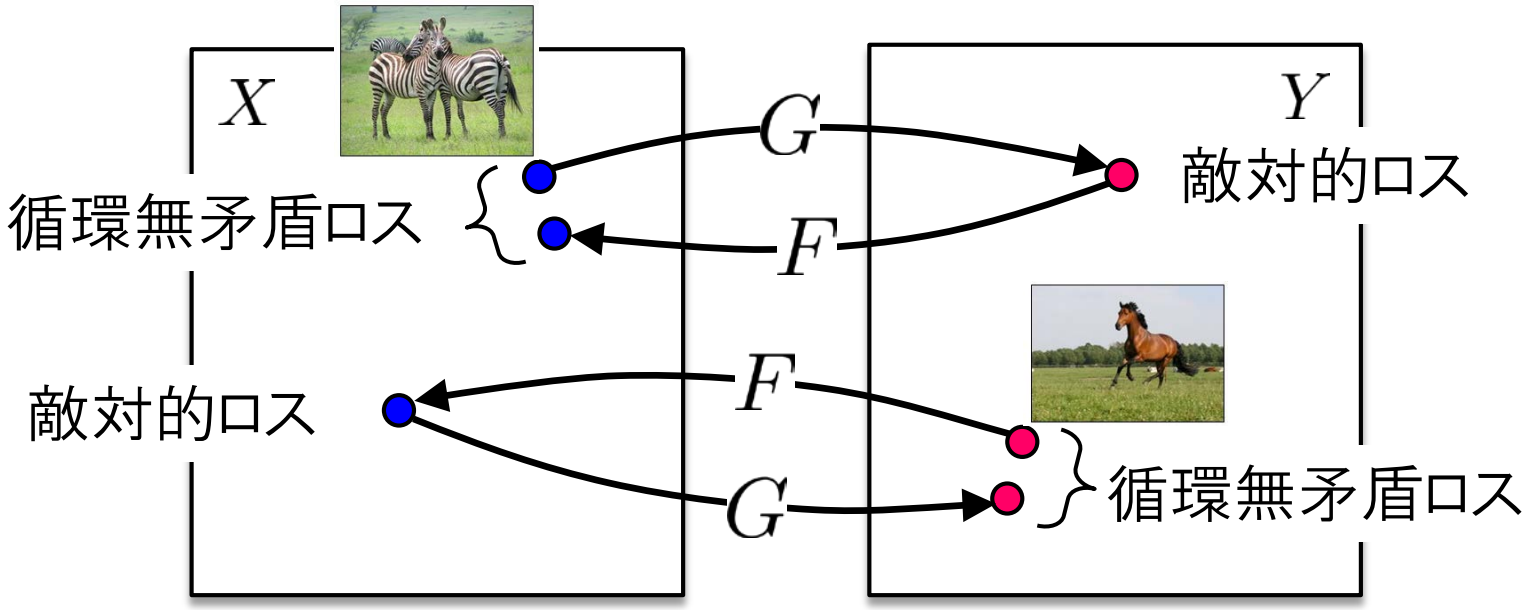


同一発話内容のペアデータを揃える必要あり

⇒応用場面によっては障壁に

CycleGAN [Zhu+2018]

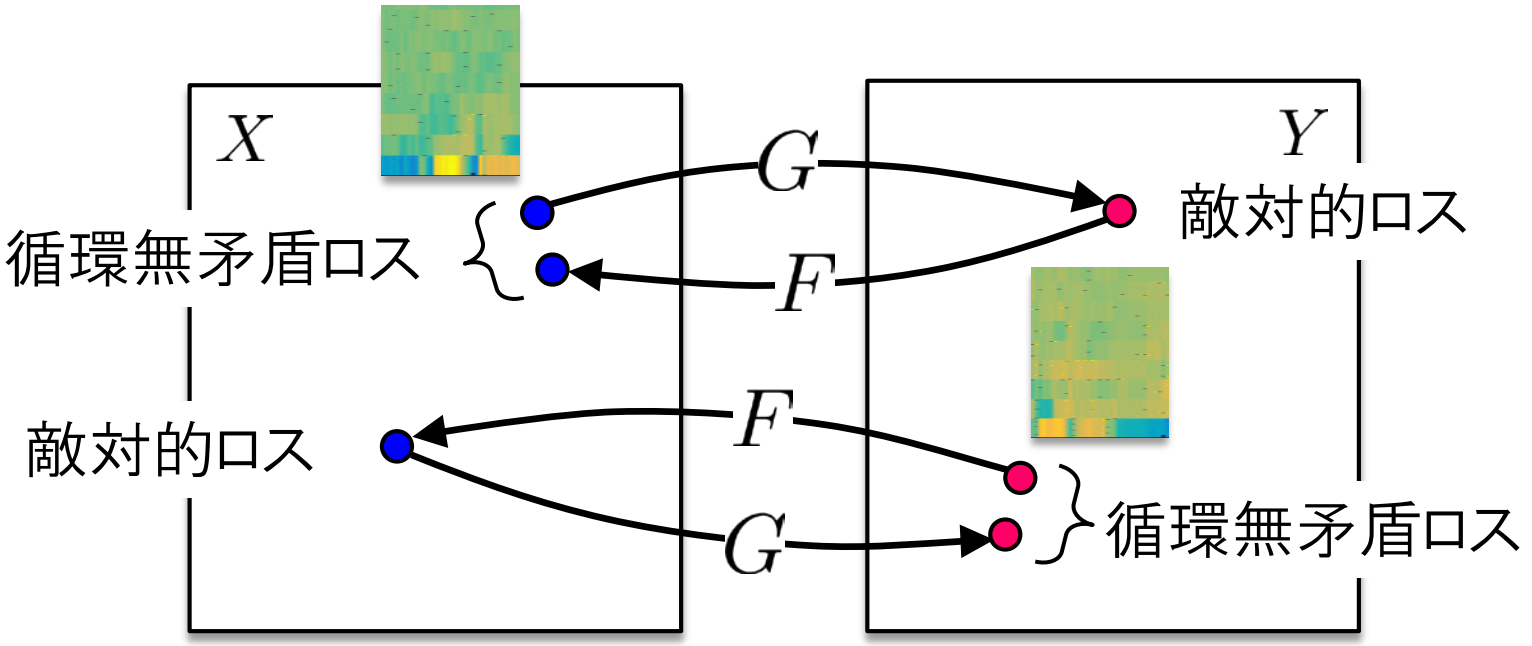
二つのドメインの**Unpairedな画像セット**を用いて画像のドメイン間の変換則を学習することが可能



- 敵対的ロス： 変換画像が変換先ドメインの画像の確率分布に従うように働く
- 循環無矛盾ロス： 一対一の変換対F, Gを見つけるよう働く

CycleGAN-VC [Kaneko&Kameoka2017]

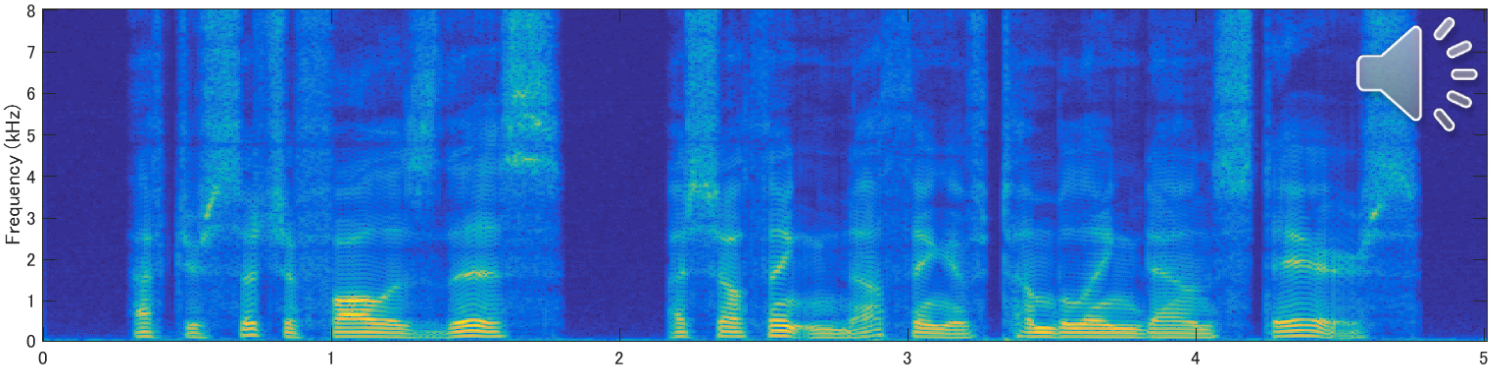
パラレルデータを必要としない音声変換



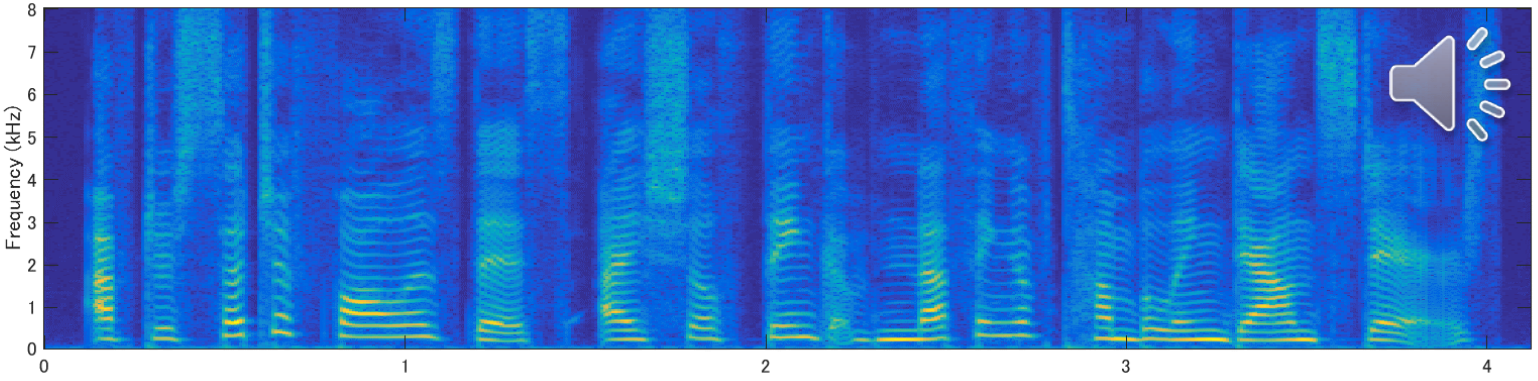
- 敵対的ロス： 変換音声の特徴量が変換先音声の特徴量の確率分布に従うように働く
- 循環無矛盾ロス： 一対一の変換対F, Gを見つけるよう働く

CycleGAN-VCによる音声変換例

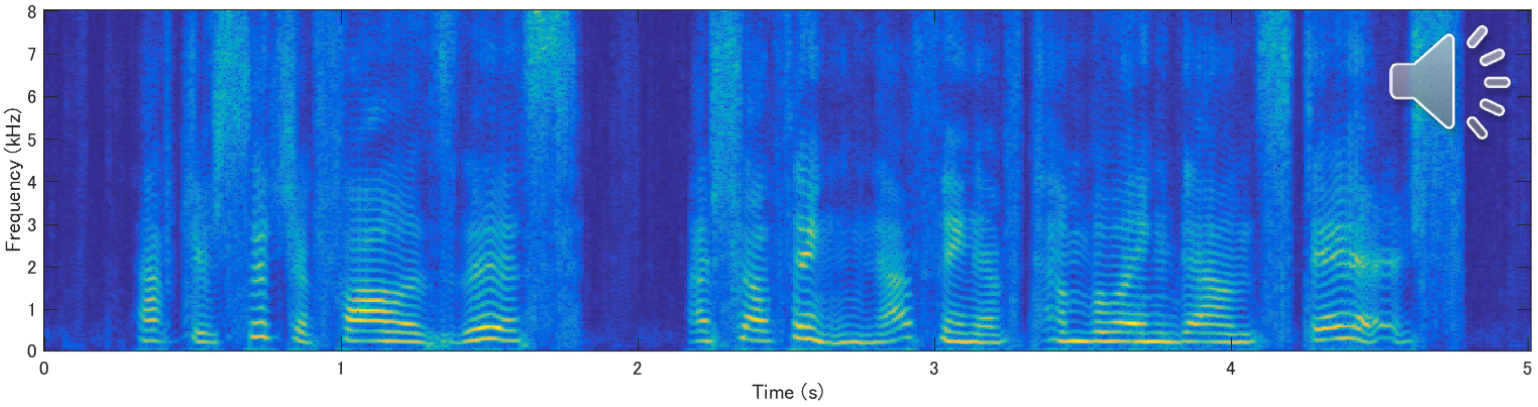
Source speech (real)



Target speech (real)



Converted speech (synthetic)

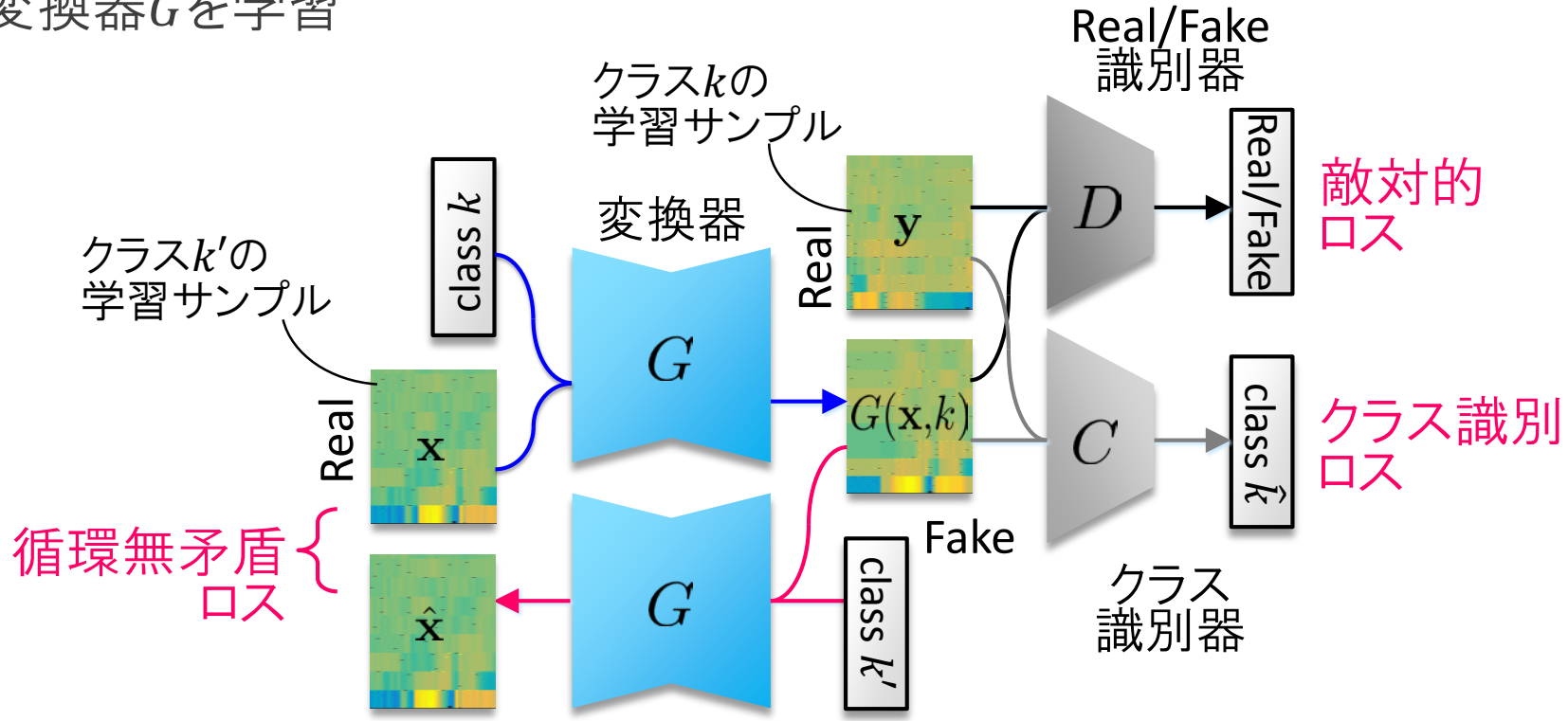


StarGAN-VC [Kameoka+2018]

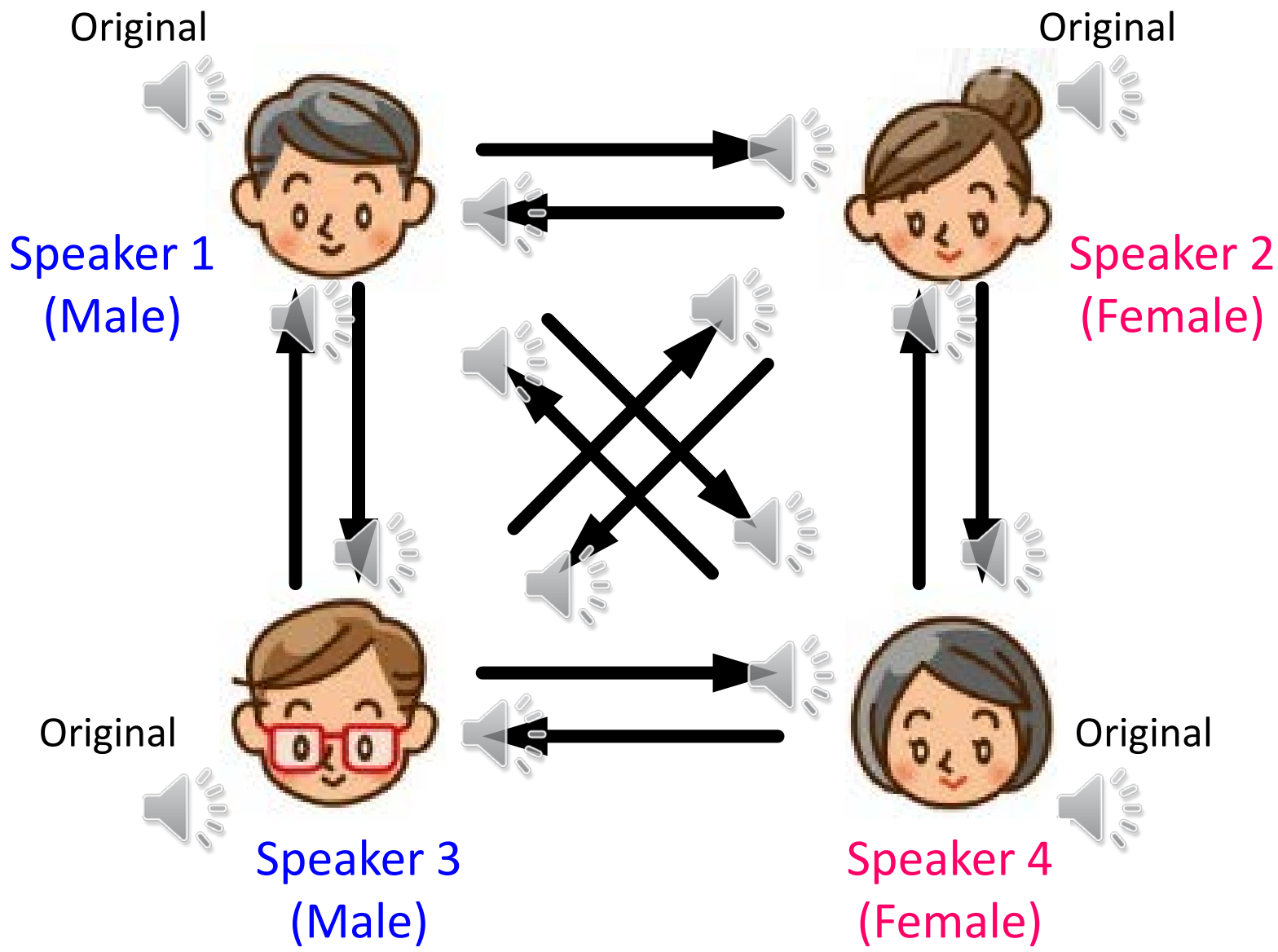
CycleGAN-VCでは学習できるのは特定の音声ペア間の変換のみ
⇒ N種類の音声間の変換には$N(N-1)$個の変換モデルが必要

StarGAN-VC: 多種音声間の変換を単一モデルで行える方式

- 画像変換法として提案されたStarGAN [Choi+2017]の応用
- 変換器出力$G(\mathbf{x}, k)$がDに「実音声」，Cにクラスkと識別されるように変換器Gを学習



StarGAN-VCによる音声変換例



次世代音声処理に向けて

“Text”-to-speechを超える音声合成とそのためのコーパス

[Takamichi18]

Jpn. end-to-end TTS
[Ueno18]

Reading-style speech (10 hrs)

JSUT

Single Japanese speaker’s voice

JSUT-vi

Vocal imitation (0.4 hrs)

JSUT-song

Singing voice (0.5 hrs)

[Release in this winter]
[new!] JSUT-book
Audiobook
[new!] JSUT-???
???

音声による抽象化･具体化を利用した多元的情報の融合へ

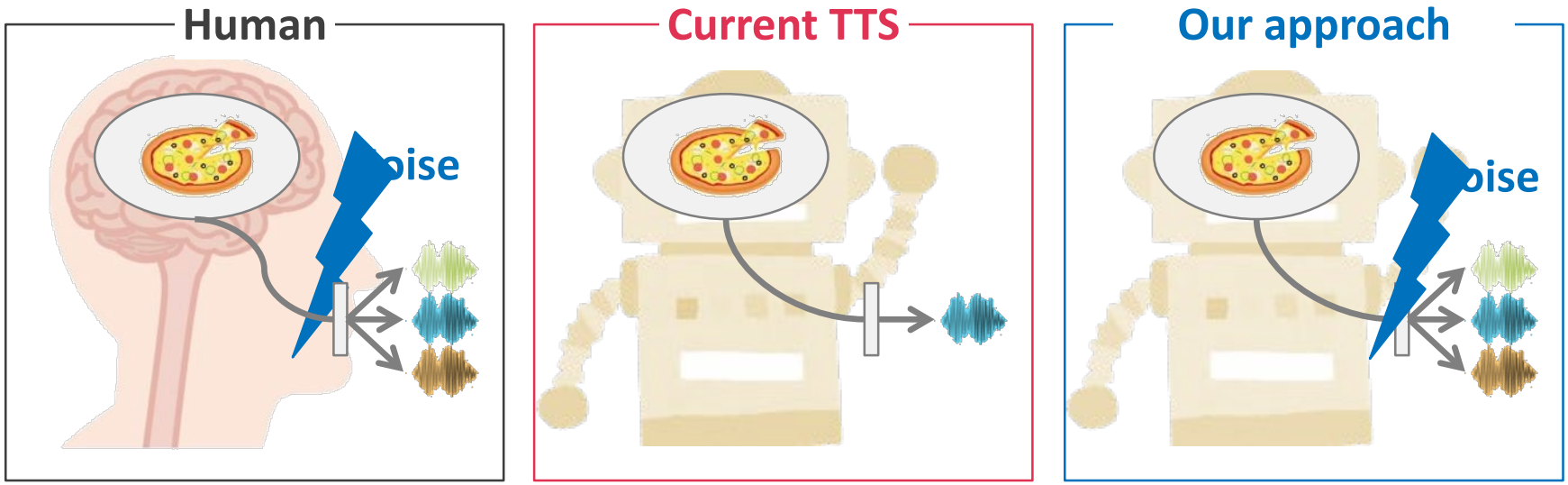
日本語End-to-end音声合成のサンプル音声は，京都大学 河原先生･上乃さまに提供して頂いた

一期一会音声合成

[Takamichi17]

今の音声合成は間違えてくれない…

- いつも同じ声・セリフ… → 人間はそうじゃない
- Moment matching network に基づく音声サンプリング [Takamichi17]
- たまにセリフを間違えたり，風邪をひいたり… (個人的希望)



「正しく喋る」から「正しく間違えて喋る」音声合成へ

感情音声合成から扇情音声合成へ

感情音声合成

- 計算機の所望した感情を合成音声に付与する技術
- 聞き手 (人間側)のことを何も考えていない

扇情音声合成

- 計算機の所望した「ユーザの感情」を起こす技術
- 人間の挙動を計算機ループに組み込んだ学習
- 計算機に気持ちよく操られたい
- (名前募集中)

人間の音声の挙動を計算機ループに組み込んだ
Human-in-the-loop 音声合成へ

これからの音声処理の所感 (高道)

音声処理の役目は，正確に合成・変換・強調することだけ？

- **答えはNo.** (もちろん正確性も大事)

音声処理の役目の1つは「音声コミュニケーションを拡張」

- 音声の芸術性を満たすには？
  - 「正しく間違える」ことは芸術的か？
- セキュリティとの関連？
  - 声の肖像権はどうあるべき？
- 人間を組み込んだ音声処理？
  - ヒューマンコンピュテーション的なアプローチ？
- IoA (Internet of Ability)としての音声処理？
  - 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？

End-to-Endモデルについて

音声認識・音声合成における旧来のアプローチでは、研究者に信号処理、言語解析、特定の確率モデル、システム設計に関する高度な知識が求められたのに対し、End-to-Endモデルの研究が進み、非専門の研究者でも比較的容易に参入できるようになってきている。

反面、End-to-Endモデルは（ほとんどの場合）学習に必要な学習データ量（音声収録、ラベリング作業にかかるコスト）が大きくなる。

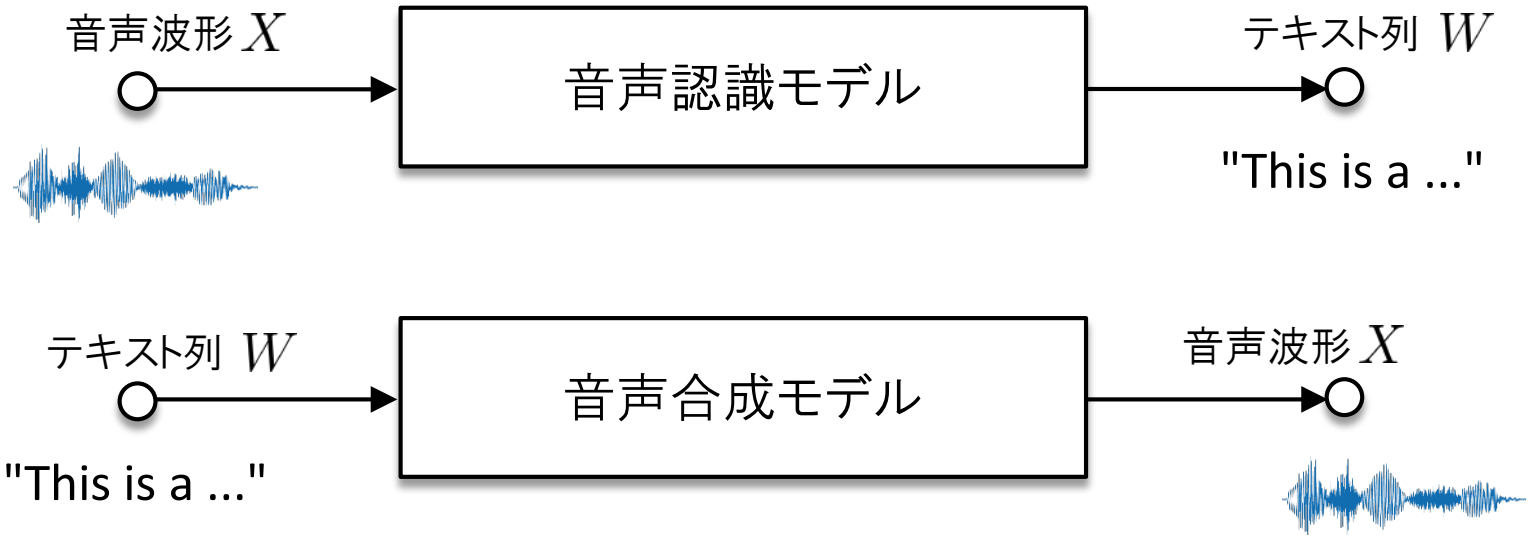
→ いかに少ないリソースで学習するか、
いかにデータ構築にかかる人手のコストを減らせるか、
がポイントになる

End-to-End･･･そしてその先は？

Endが存在する限り人手のラベリング作業はなくならない

容易に入手可能なunlabeledデータを活用した学習系

⇒EndとEndを結んだEndlessモデル

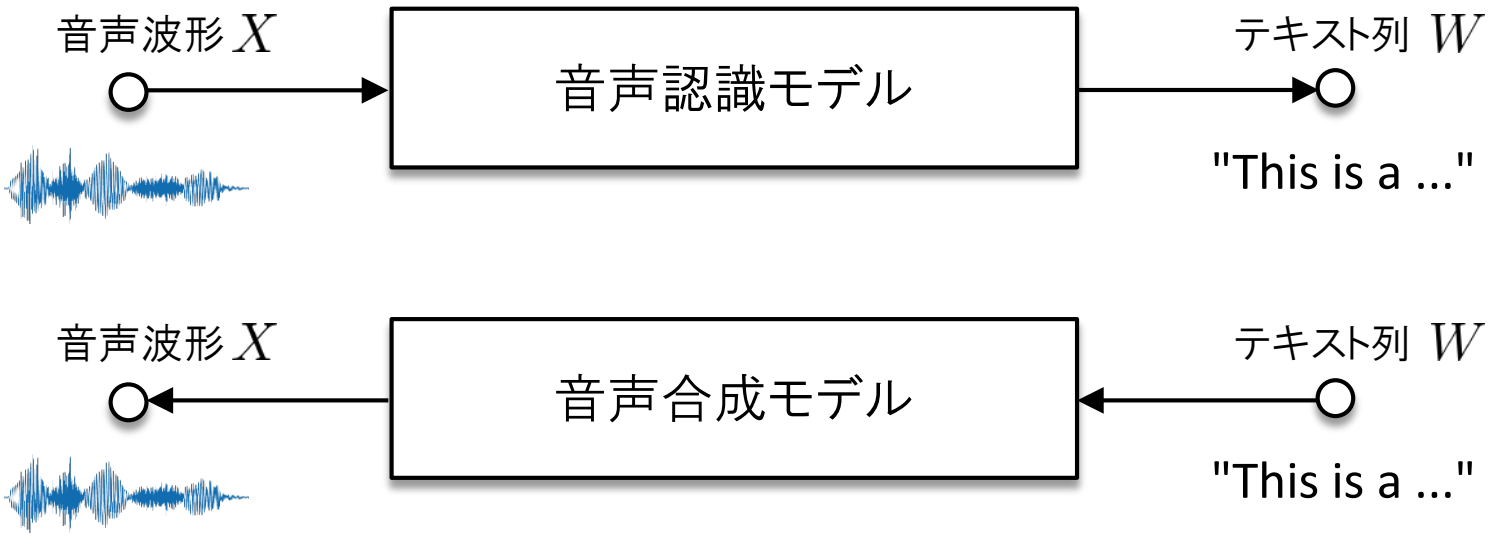


End-to-End･･･そしてその先は？

Endが存在する限り人手のラベリング作業はなくならない

容易に入手可能なunlabeledデータを活用した学習系

⇒EndとEndを結んだEndlessモデル

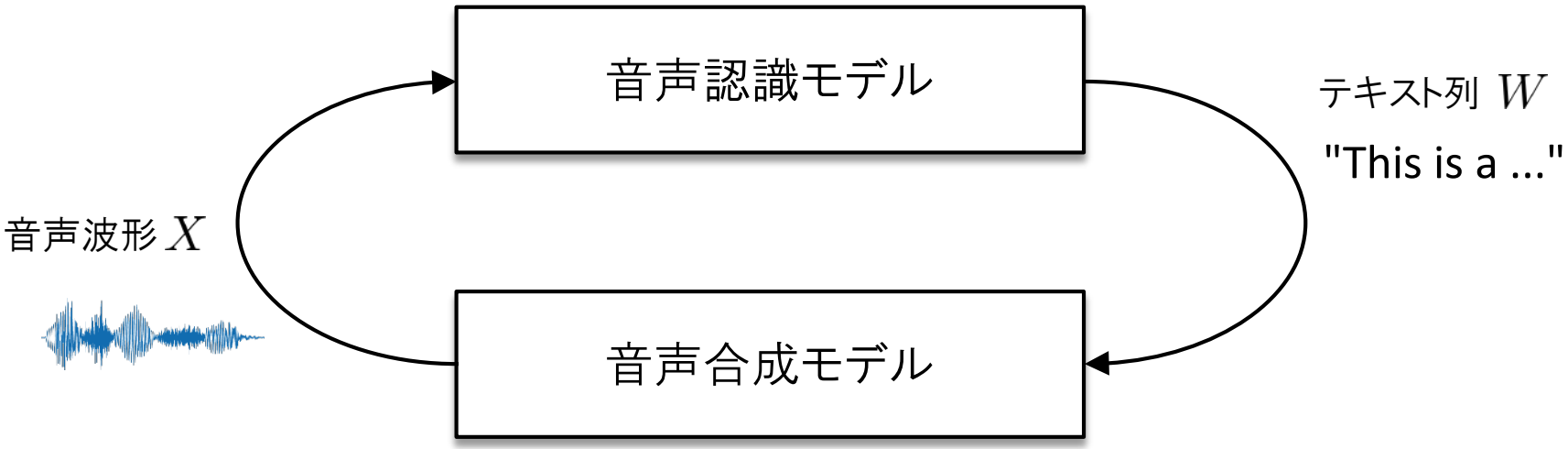


End-to-End･･･そしてその先は？

Endが存在する限り人手のラベリング作業はなくならない

容易に入手可能なunlabeledデータを活用した学習系

⇒EndとEndを結んだEndlessモデル

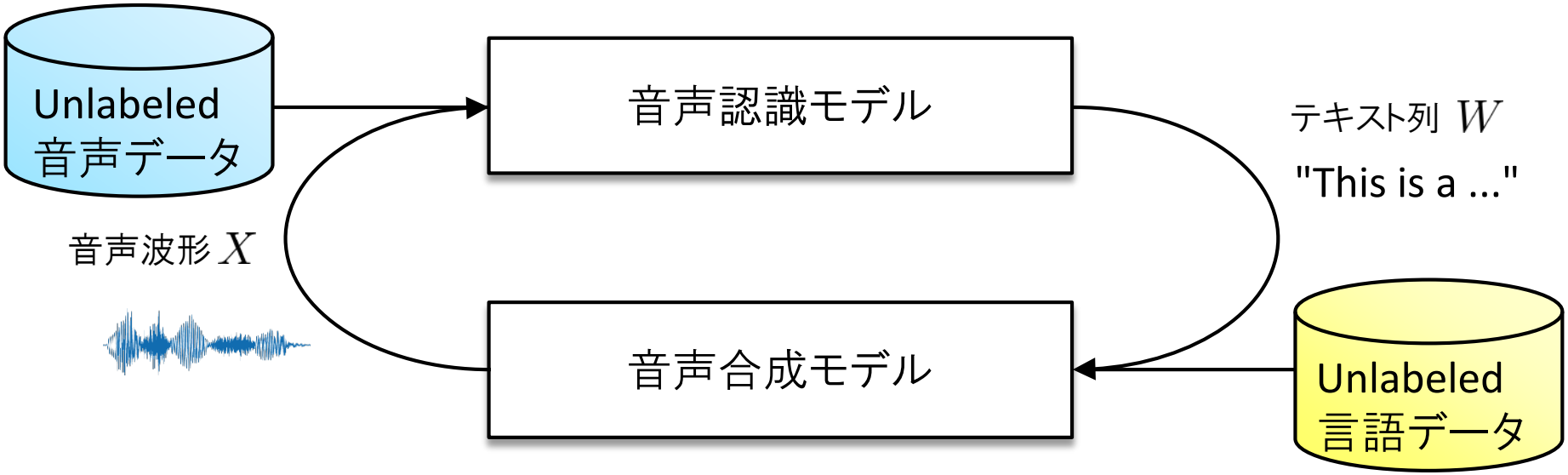


End-to-End･･･そしてその先は？

Endが存在する限り人手のラベリング作業はなくならない

容易に入手可能なunlabeledデータを活用した学習系

⇒EndとEndを結んだEndlessモデル

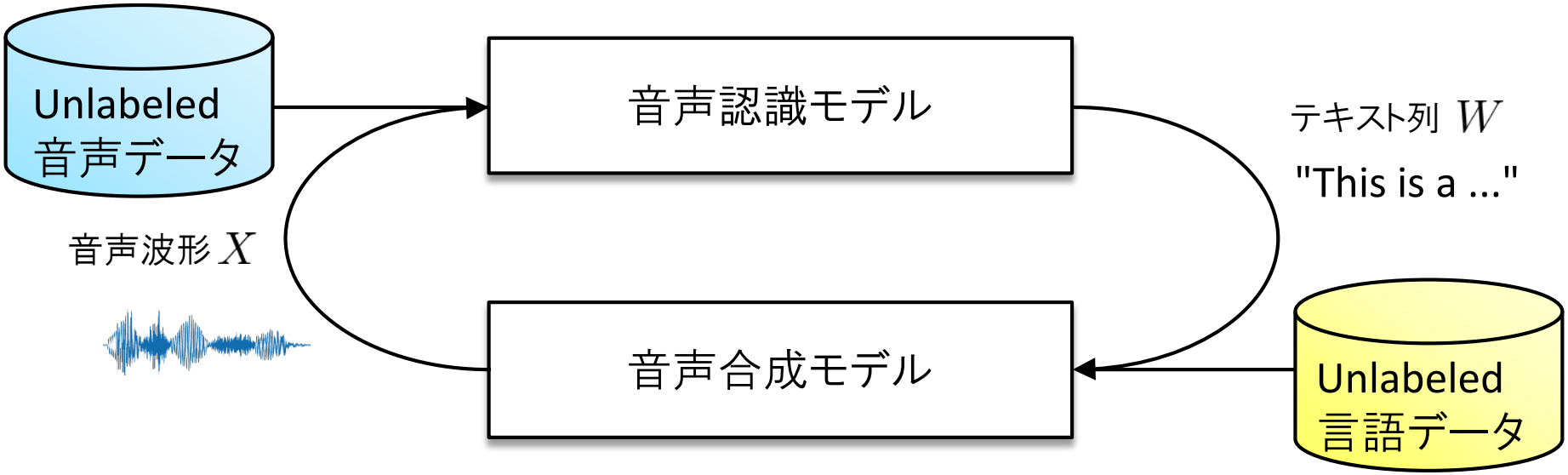


End-to-End･･･そしてその先は？

Endが存在する限り人手のラベリング作業はなくならない

容易に入手可能なunlabeledデータを活用した学習系

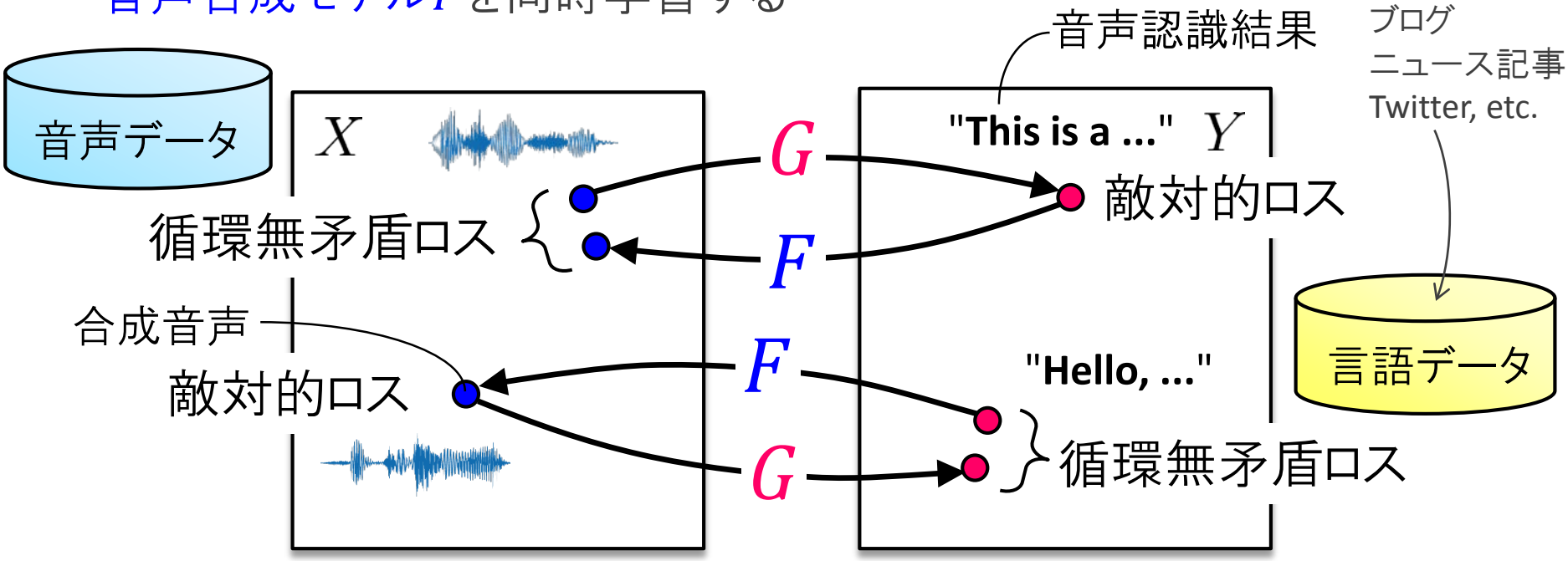
⇒EndとEndを結んだEndlessモデル



合成音声を用いて音声認識モデルの学習データを拡張する（両システムを疎結合する）試みもあるが、ここで考えたいのは一体化したモデル

CycleGANによるEndlessモデルの実現？

Unlabeledな音声／テキストデータを用いて音声認識モデルGと
音声合成モデルFを同時学習する



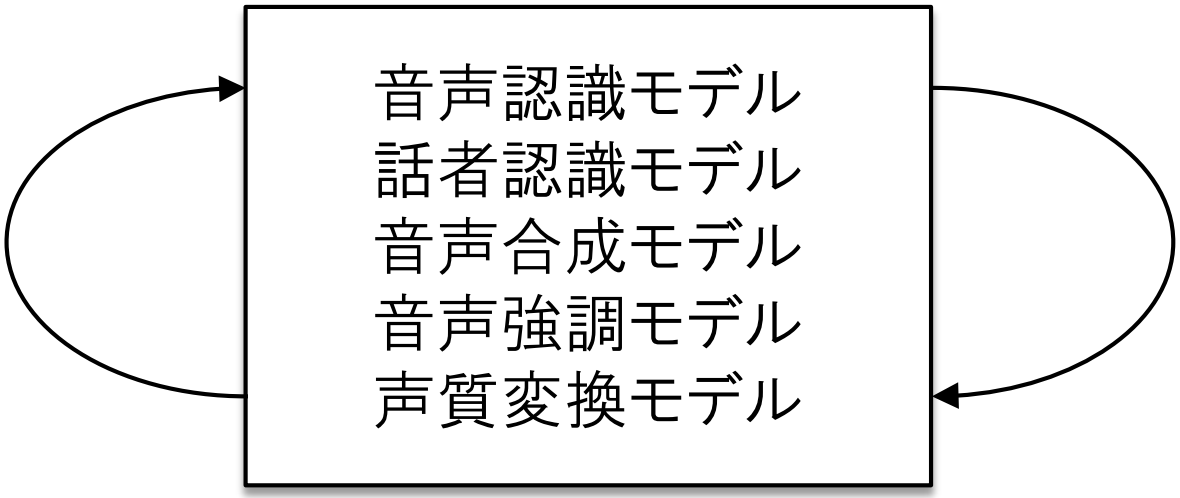
循環無矛盾ロスにより一対一の変換対F, Gを見つけられる

テキスト列⇒音声⇒テキスト列
音声⇒テキスト列⇒音声
} で元どおりになる

敵対的ロスにより、合成音声が実音声らしい音声になり、かつ、
認識結果が自然言語的に自然なテキスト列になる

End-to-Endの先は？（亀岡）

音声認識と音声合成のペアだけでなく、
音声強調、声質変換、話者認識なども含めたEndlessモデルを
考えることで、多様な音声、ノイジーな音声を扱える枠組に



究極は、人手のラベリングがまったく不要な
多目的音声処理システム

まとめ

本発表のまとめ

音声分野から見た敵対的学習 (GAN)

- 音声合成・変換におけるGANの応用事例
  - 音声を画像と見なしたGANの適用
  - 音声なりすまし検出を騙すためのGANの利用
- 音声変換におけるGANのさらなる応用発展
  - 他の深層生成モデル vs GAN
    ⇒ 学習の効率性 vs 生成の効率性
    （リアルタイムシステムにおいてGANが優位になりうる）
  - CycleGANによるパラレルデータフリー音声変換
  - StarGANによる多対多音声変換、StarGANの改良

次世代音声処理を見据えた展開

- 人間のようなランダム性を備えた「一期一会音声合成」
- 人間の挙動を意識した「扇情音声合成」
- 人手のラベリングが不要な「Endlessモデル」