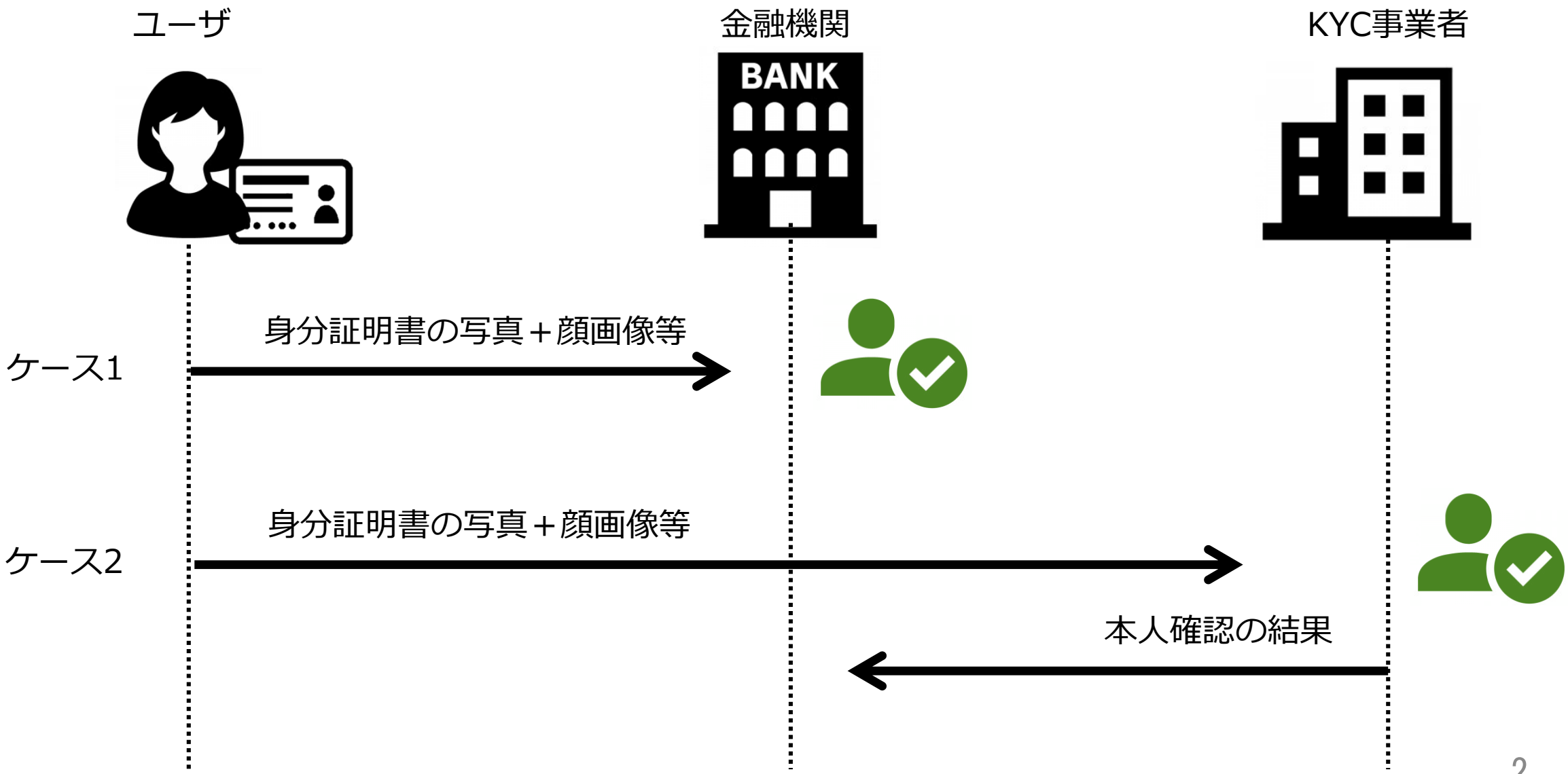
ディープ・フェイクによるなりすましに関する最新の研究動向：USENIX2022の研究紹介

2022年11月22日

日本銀行 金融研究所 情報技術研究センター
菅 和聖

※本発表の内容は、発表者個人の見解であり、
日本銀行の公式見解を示すものではありません。

金融機関におけるeKYCの利用



ディープ・フェイクとは

深層学習で作成された人物に関する精巧な合成メディア（画像／動画／音声）を指す。学術用語ではなく意味は曖昧。

(例) ゼレンスキー大統領が兵士に投降を呼びかける偽動画が拡散（2022年3月）

(例) オンライン面接などで他人になりすまして米国のITエンジニア職に応募するケースの増加をFBIが警告。

- リモート・ワークの普及が背景。企業の機密情報へのアクセスや違法な米国での就業が目的か。

（資料）1. https://www.bbc.com/news/technology-60780142
2. https://www.businessinsider.jp/post-256041

本日ご紹介する論文(1) USENIX 2022より

- Changjiang Li *et al.*, "Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era."

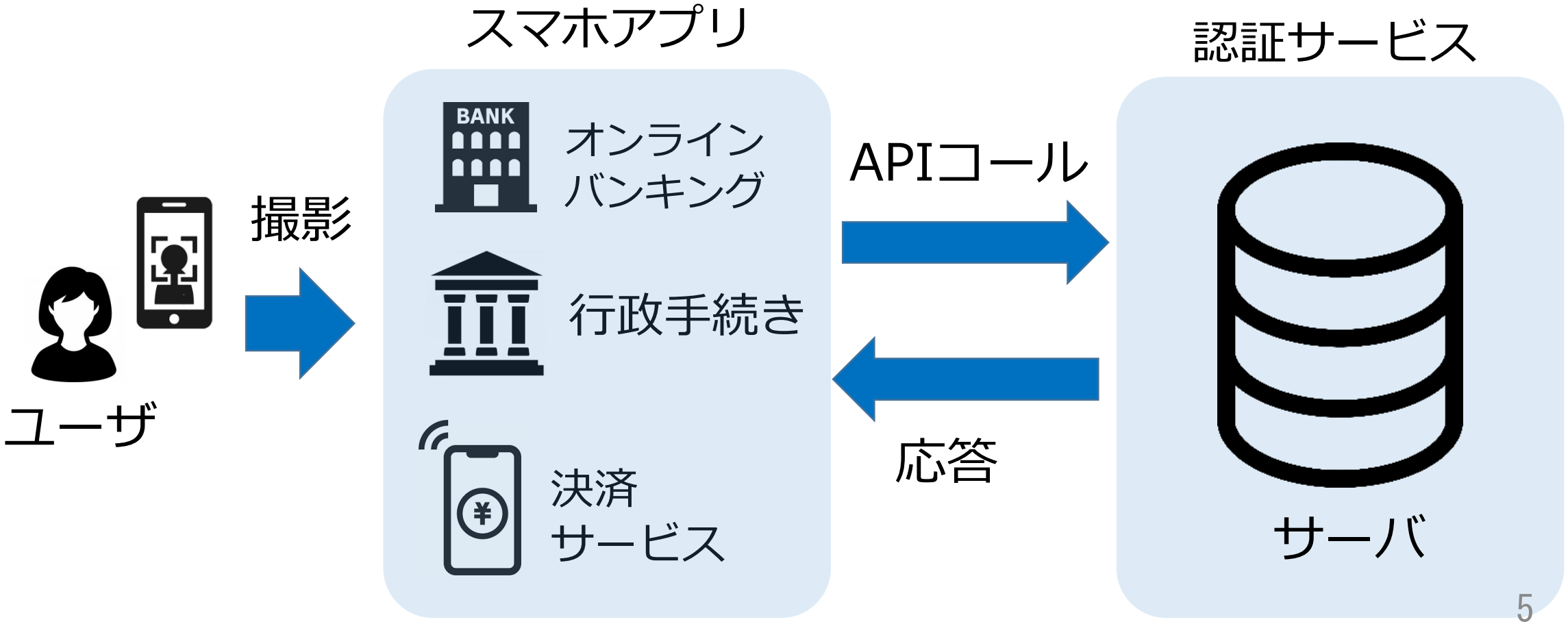
  実運用されている顔画像認証のAPIサービスを対象に、ディープ・フェイクを見抜けないことを検証

- Logan Blue *et al.*, "Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction."

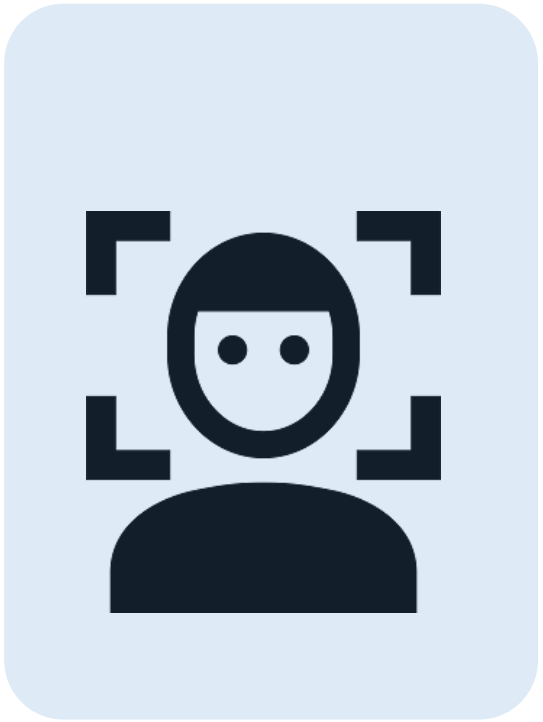
  人間の発声メカニズムをシミュレートして、ディープ・フェイクによる音声を高精度で見破る手法を提案

顔画像認証 (facial liveness verification: FLV)

- eKYC等で利用される顔の動画像で認証を行う方式
- クラウド基盤で動作する認証サービスのAPIが提供される

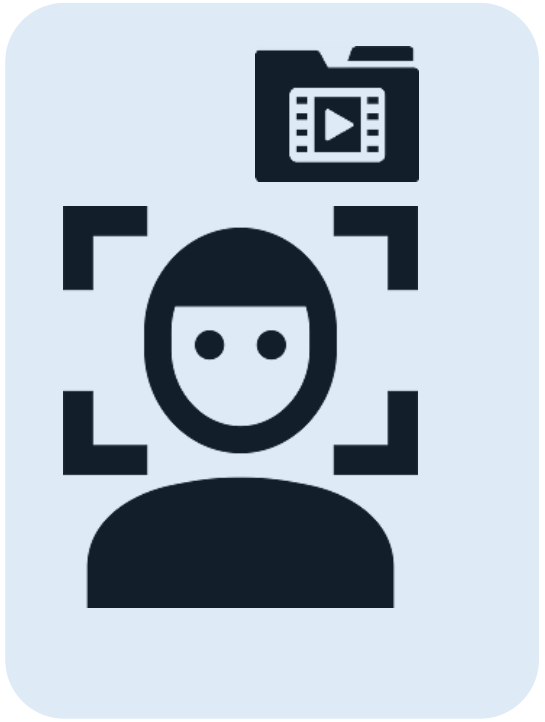


顔画像認証のタイプ



静止画

Image-based FLV

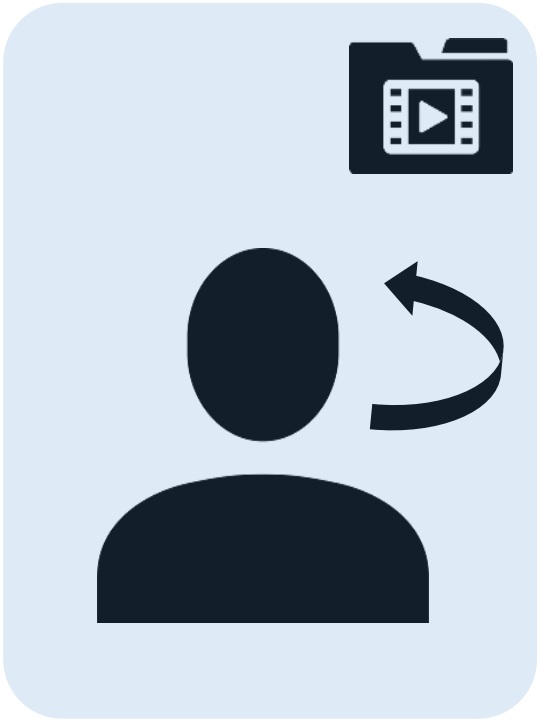


音声なし動画

Silence-based FLV

音声あり動画

Voice-based FLV



動作あり動画

Action-based FLV

顔画像認証の処理

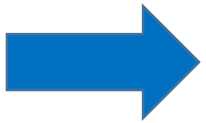
(1) 生体検知（liveness detection）

・入力データの偽造（presentation attack）を防ぐ
（例）過去動画を再生するリプレイ攻撃（replay attack）の防止

(2) ディープ・フェイク検知（deepfake spoofing detection）

・ディープ・フェイクによるなりすましを防止

(3) 顔の照合（face matching）

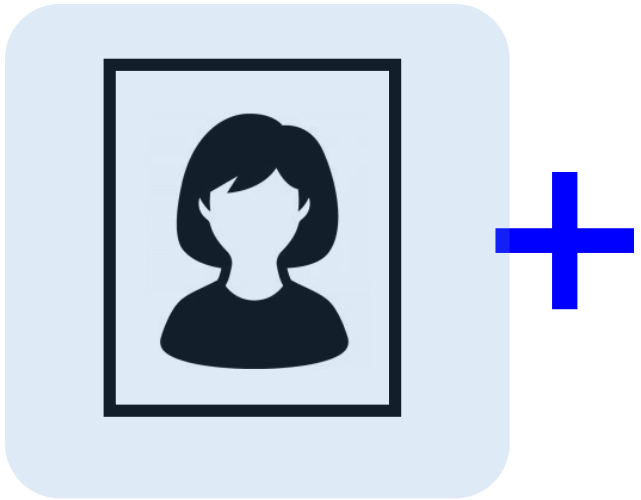


すべて突破すると
なりすまし成功！

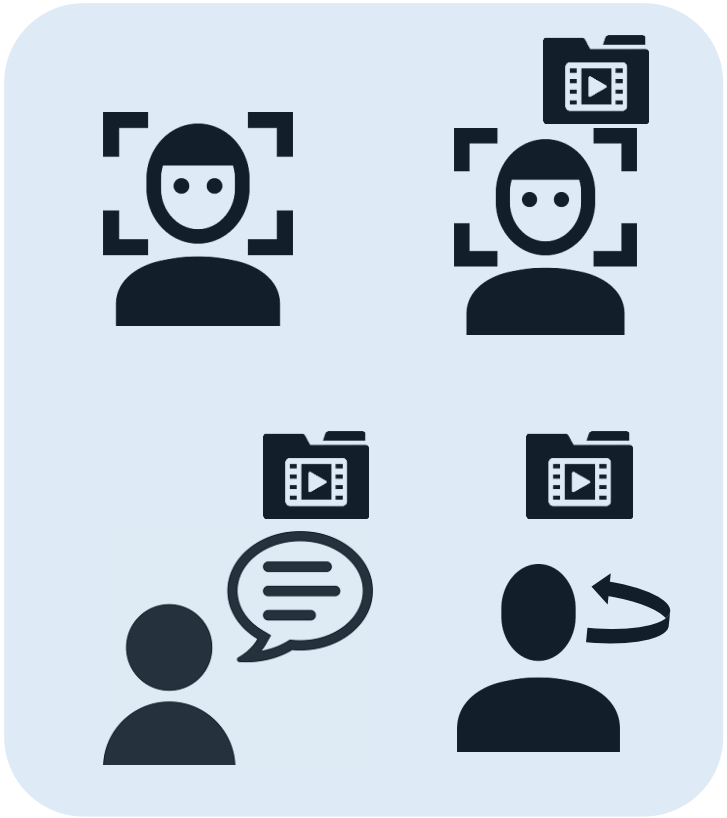
攻撃モデル

- 攻撃者は標的（victim）の顔画像1枚からFLVの突破を試みる
- APIサービスの内部実装は知らない

標的の静止画像



様々な駆動データ



様々な合成データ

ディープ・フェイクの作成手法(1/2)

- 標的画像が動画として駆動可能 → 動画認証への攻撃に悪用
- 標的の顔画像の表情を変えることができる特徴（背景は不変）

標的画像
(target image)

駆動画像
(driving image/video)

表情の再現
(face reenactment)

（資料）Changjiang Li *et al*. (2022) Figure 2

ディープ・フェイクの作成手法(2/2)

- 標的の顔画像の背景などを変えることができる特徴
- 標的画像の背景や髪形などが認証突破の妨げになる場合に有効と期待される

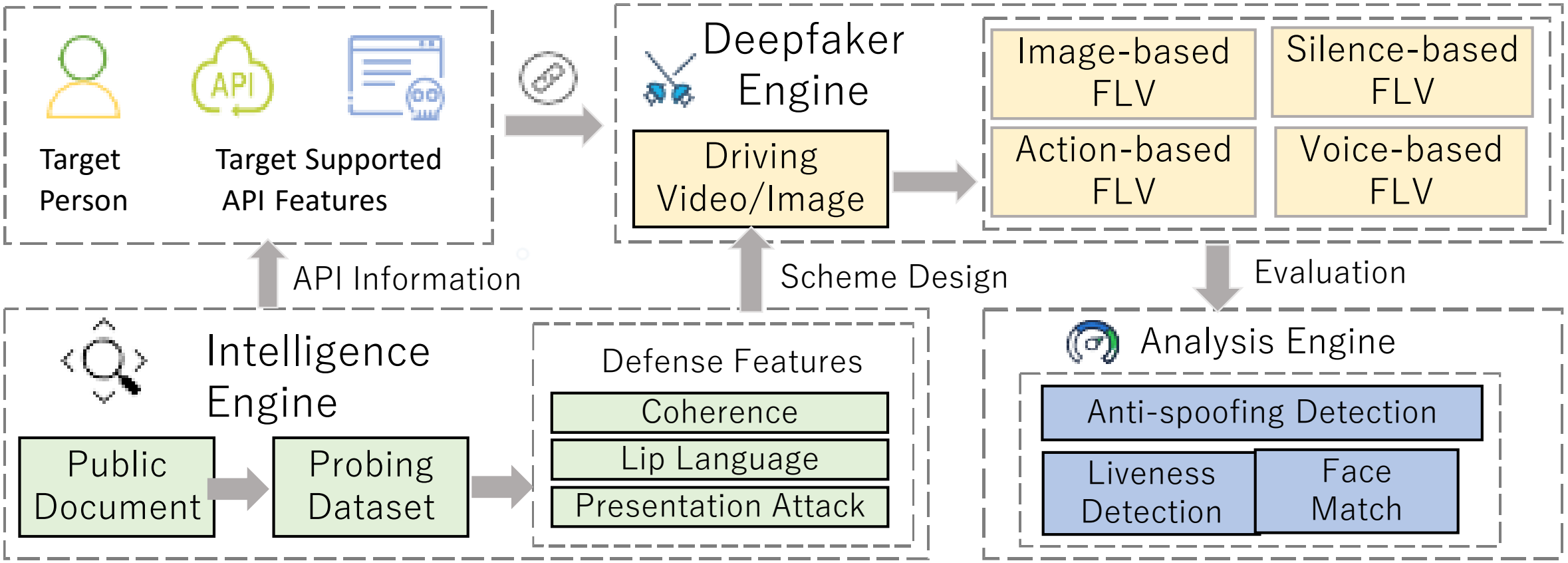
標的画像
(target image)

駆動画像
(driving image/video)

顔の入れ替え
(face swapping)

（資料）Changjiang Li *et al*. (2022) Figure 2

自動検出のプログラムの構成



（資料）Changjiang Li *et al*. (2022) Figure 3

FLVの仕様の一覧表

Platform	Liveness Type: Image	Video: Type: Silence	Voice: Voice Cide Length Range	Voice: Voice Code Type	Voice: Default Code Length	Voice: Lip Language Detection	Action: Action Length Range	Action: Default Action Length	Action: Action Type	Common Detection: Anti-deepfake Detection	Common Detection: Coherence Detection	Common Detection: Replay Attack Detection
BD	●	●	3 - 6	Digits	3 - 6	○	1 - 3	1 - 3	Blink, Turn Right Turn Left, Look Up Chin Down,	●	○	●
TC	●	●	1 - 6	Digits	4	◐	1 - 2	2	Turn Right and Left Blink, Open Mouth	●	○	●
HW	●	○		○			1 - 4	1	Turn Left, Turn Right, Blink, Open Mouth	○	○	●
CW	●	●	4 - 6	Digits	4 - 6	●			○	○	○	●
ST	○	●	4	Digits	4	○			○	○	○	●
iFT	●	●		○					○	○	○	●

Table 1: API intelligence collected from cloud platforms. ● denotes full support; ◐ denotes partial support; ○ denotes no support.

（資料）Changjiang Li *et al*. (2022) Table 1

検証結果とセキュリティ対策

- すべてのタイプのFLVはディープ・フェイクに対して脆弱性あり
- 人種、性別などの属性によって攻撃の成功率に差がある
  - 女性>男性、白人>有色人種 で攻撃の成功率に差
- 敵対的学習（adversarial learning）でなりすまし攻撃の成功率が上昇する
- 駆動画像の品質により攻撃の成功率は変化する

- Deepfakeが作りにくい動作（action）を採用する
- 求める音声と動作の多様性を確保する

本日ご紹介する論文(2) USENIX 2022より

- Changjiang Li *et al.*, "Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era."

  実運用されている顔画像認証のAPIサービスを対象に、ディープ・フェイクを見抜けないことを検証

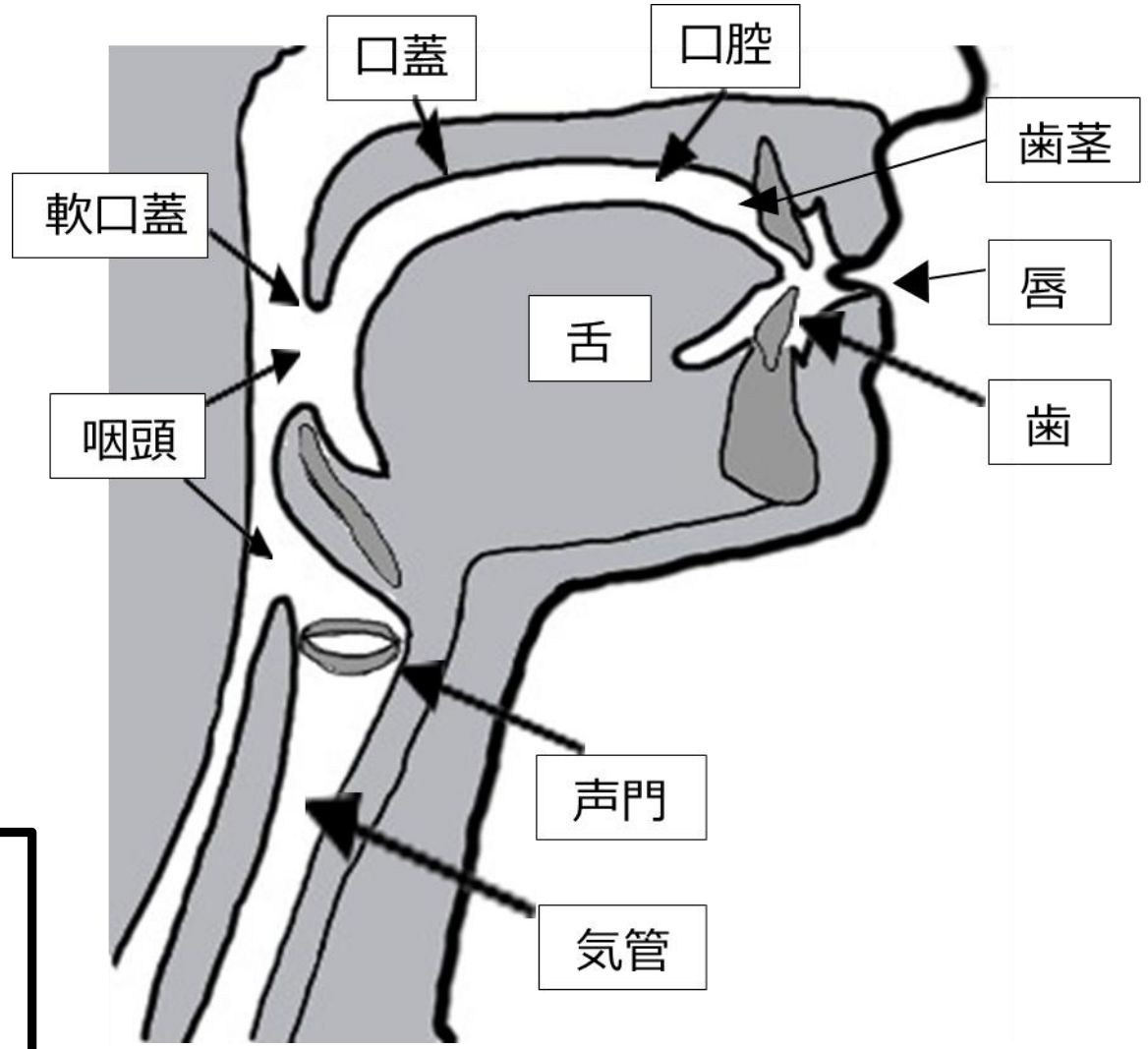
- Logan Blue *et al.*, "Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction."

  人間の発声メカニズムをシミュレートして、ディープ・フェイクによる音声を高精度で見破る手法を提案

人間の発声器官

- 人間の発声器官は構造を持つ
- その中の空洞（声道）を音が拡散して声が出る
- Deepfakeは、上記の物理法則（器官構造や流体力学）を考慮していない

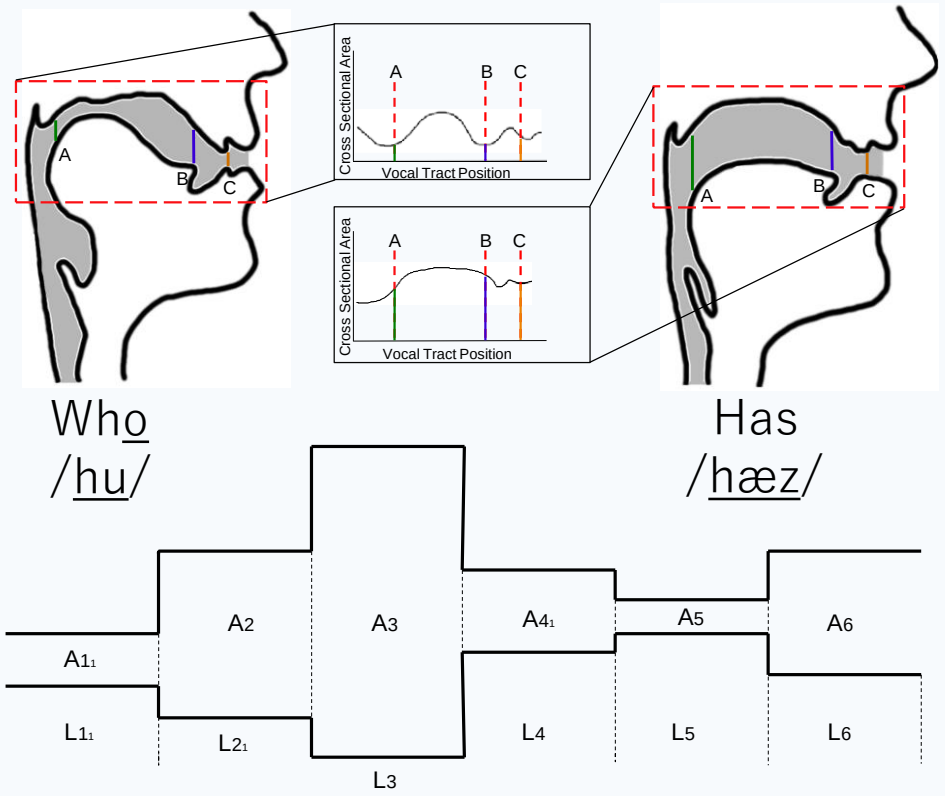
現状、物理法則がもたらす特徴を使えばDeepfake検出できるはず

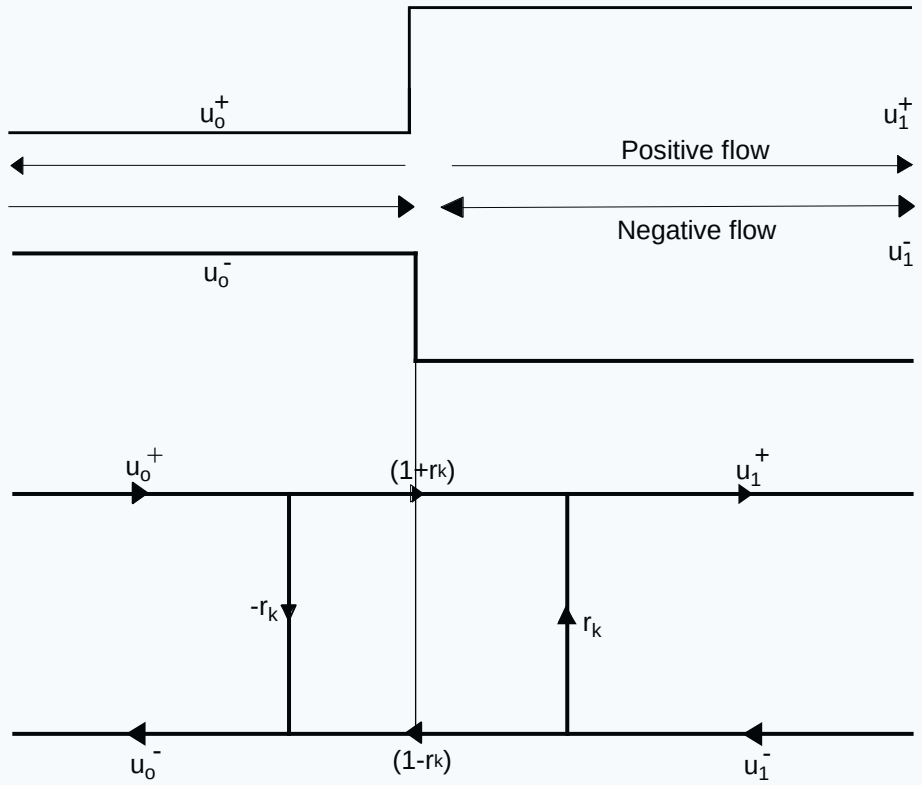


（資料）Logan Blue *et al.* (2022) Figure 1. 当方和訳

発声過程のモデル化

声道の形状と空気の流れを流体力学に基づきモデル化

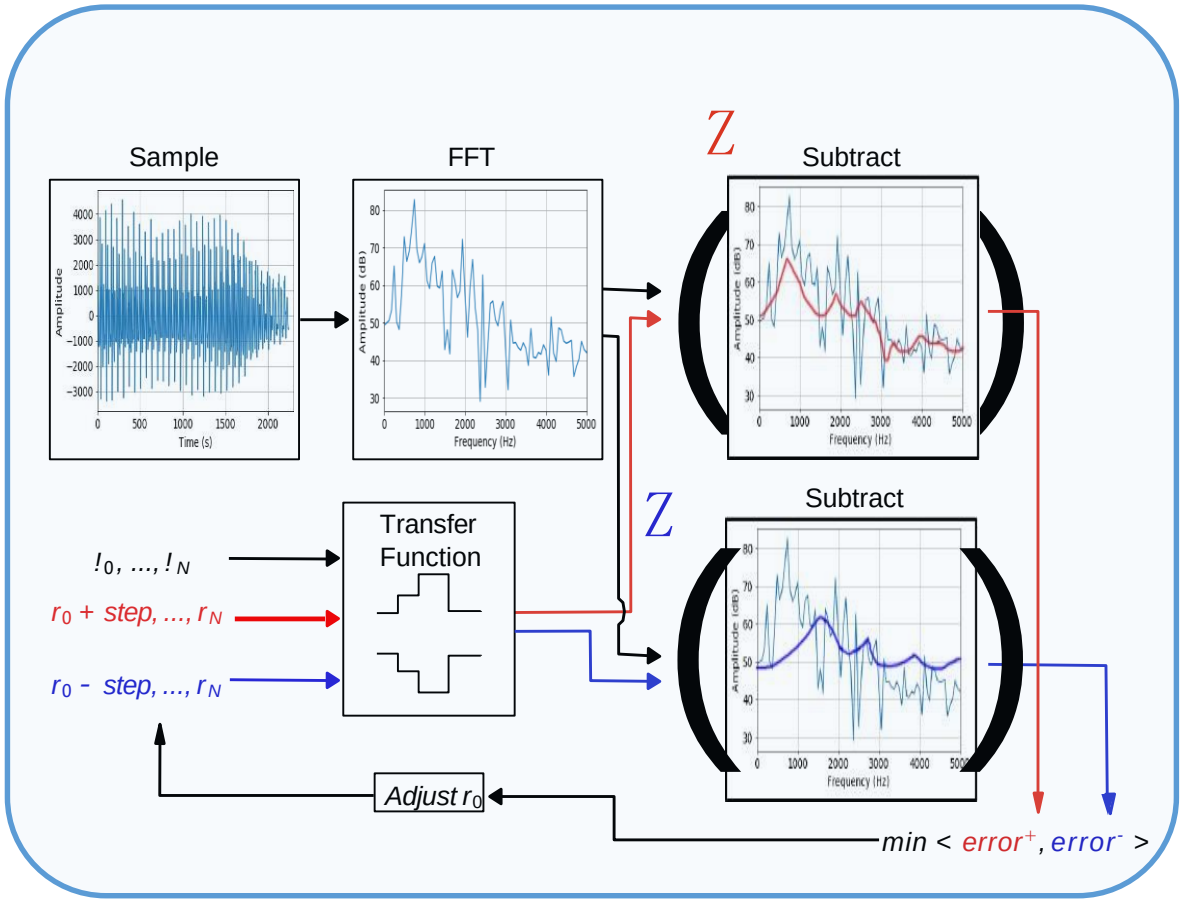


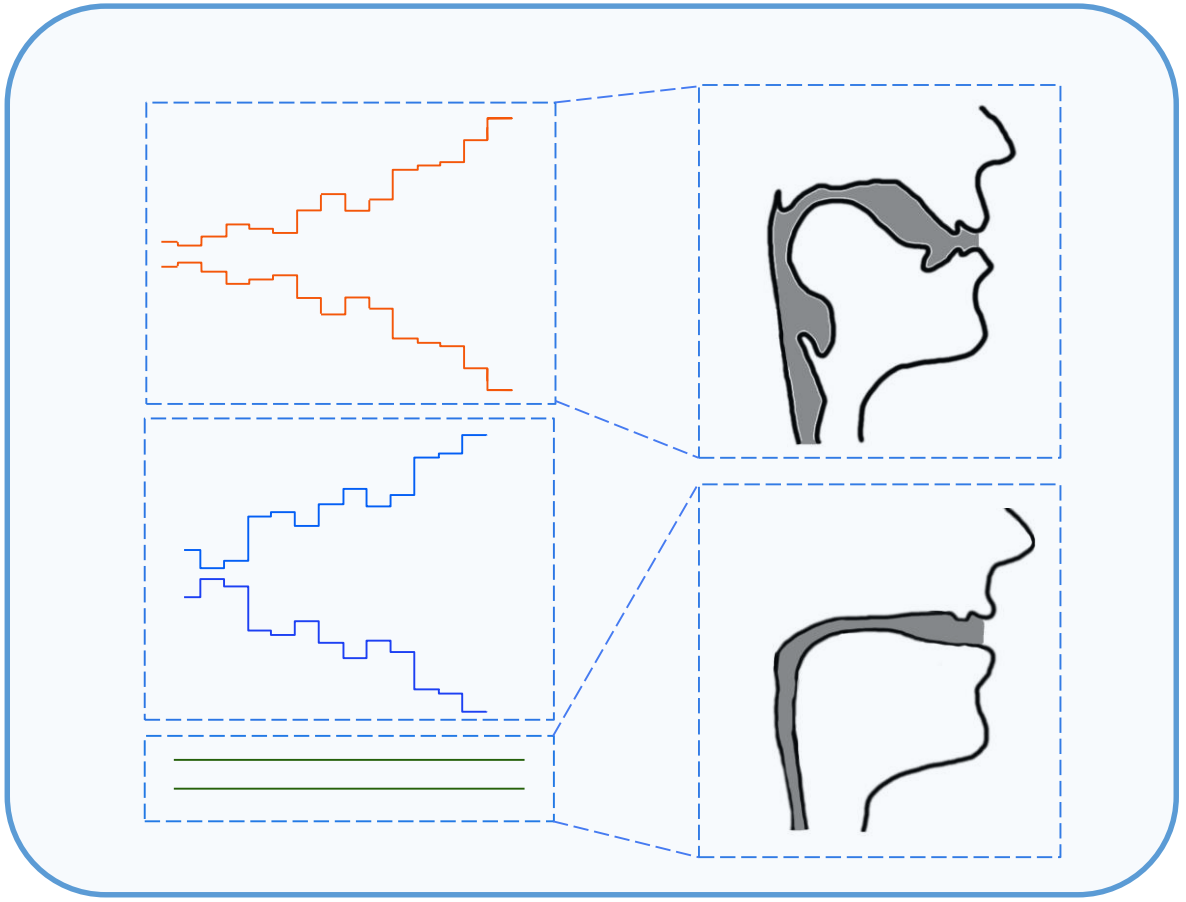


（資料）Logan Blue *et al*. (2022) Figure 3-5

波形から声道の形状を推定

周波数解析で形状を逆算。99%以上の精度でディープ・フェイクを検出





（資料）Logan Blue *et al.* (2022) Figure 6, 10

(考察) ディープ・フェイクとセキュリティ

- 画像/動画/音声の偽造は容易になりつつある。本物とディープ・フェイクを見分けることは、より困難になると見込まれる
- リスク１：ディープ・フェイクにより認証を突破される可能性
  - 検出技術の向上や認証プロセスの複雑化には一定の効果
  - 検出モデルが公開されると敵対的学習で潜り抜ける技術も獲得され、「いたちごっこ」になる恐れ
- リスク２：真正な顔画像等のデータであっても、利用者が本人であることを証明する上で、それ単体では十分な証拠能力を持つとまでは言えなくなる可能性
- 顔画像認証の導入は、なりすまし攻撃を難しくするうえで有用と考えられる。今後は、それに加えて多要素認証を取り入れるなど、リスク低減策の研究動向が注目される。