National Institute of

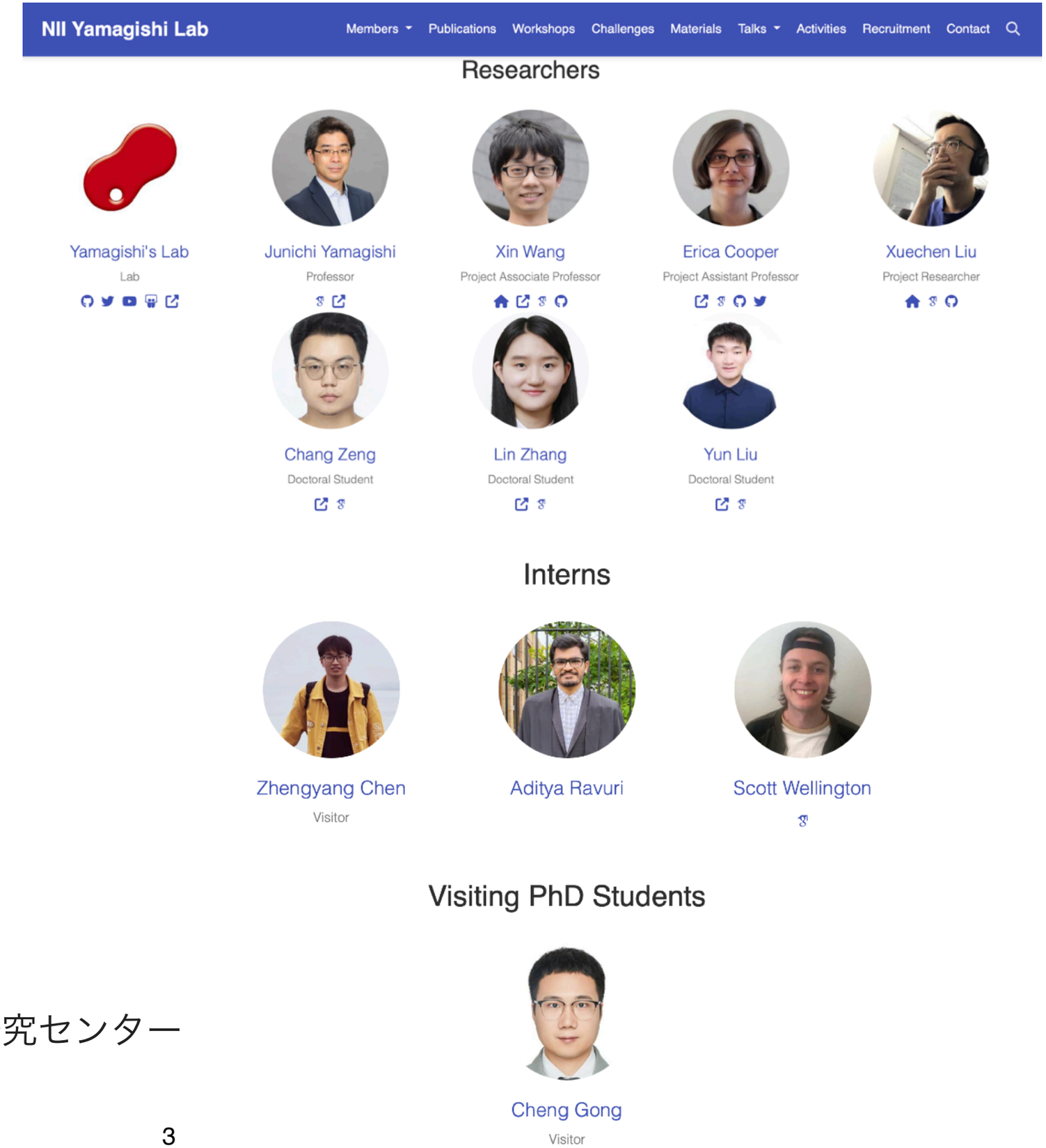
音声のディープフェイク検知はどこまで可能か？

山岸順一
国立情報学研究所
シンセティックメディア国際研究センター

自己紹介

自己紹介とNII山岸研究室について

- **経歴**
  - 2006
    - 東京工業大学
    - 博士後期課程短期修了
  - 2007～2010：
    - 英国エジンバラ大
    - Research fellow
  - 2011～2019：
    - 英国エジンバラ大
    - Career Acceleration Fellow
- 2013～2019：
  - 国立情報学研究所(NII)
  - 准教授
- 2019～現在：
  - 国立情報学研究所 教授
- 2021～現在：
  - 国立情報学研究所
  - シンセティックメディア国際研究センター
  - 副センター長



今日の講演内容の背景

マカフィー株式会社による調査

AI（人工知能）を悪用した音声詐欺が世界で増加中

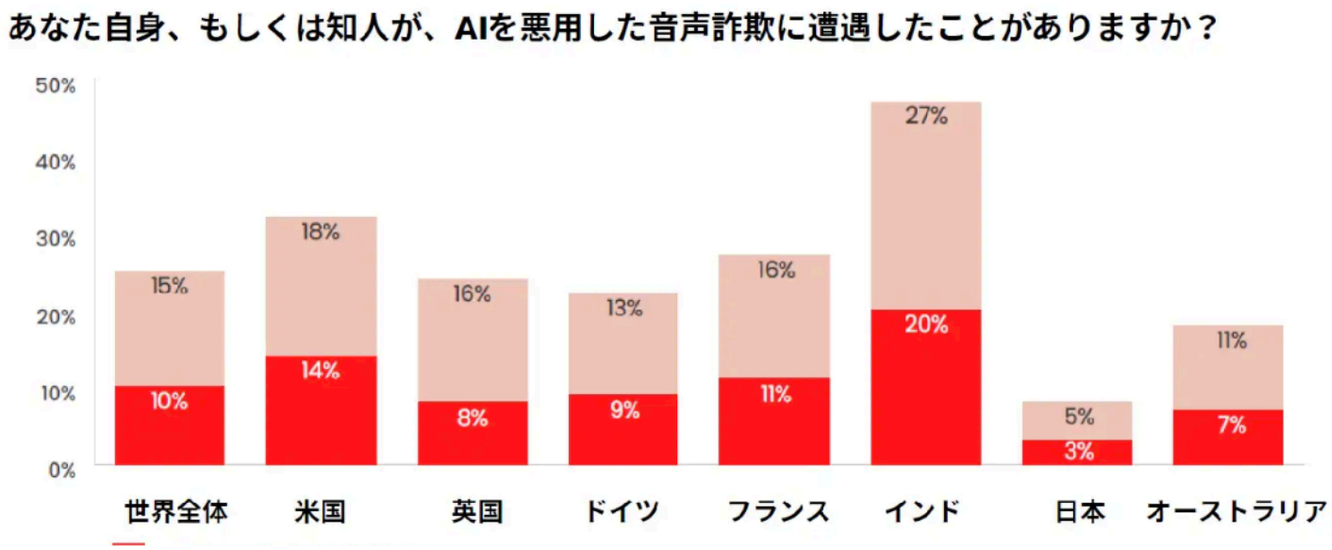
成人の10人に1人が被害に遭遇“音声クローンが3秒で作成できる今”、日本の被害遭遇率は低いものの、今後への注意が必要

マカフィー株式会社

2023年5月17日 16時28分

マカフィー株式会社（本社：東京都渋谷区）は本日、日本を含む世界7ヵ国（日本、米国、英国、ドイツ、フランス、インド、オーストラリア）を対象に実施したAI（人工知能）を悪用したオンライン音声詐欺の現状に関するレポート「The Artificial Imposter」を発表しました。

マカフィーが7ヵ国の18歳以上の成人7,054人を対象に調査を行ったところ、10%が自身がAI音声詐欺に遭遇、15%が知人が遭遇したと回答しました。なお、被害者の77%が実際に金銭被害にあったと回答しています。日本のオンライン音声詐欺の遭遇率は、世界7ヵ国で最も低く、自身が遭遇は3%、知人が遭遇は5%にとどまり、世界平均の約3分の1でした。



引用元

https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf

cloud.google.com

Google Cloud 概要 ソリューシ > ドキュメント サポート Language コンソール

Cloud Text-to-Speech カスタム音声 ガイド リファレンス お問い合わせ 無料で利用開始

フィルタ

Cloud Text-to-Speech カスタム音声
- カスタム音声のドキュメント
  - 概要
  - 基本
  - クイックスタート
  - トレーニング データの要件
  - よくある質問

リソース

Cloud Text-to-Speech Trusted Tester 向けのフォーラム

一般公開 Cloud Text-to-Speech に戻る

Cloud Text-to-Speech のドキュメント

ドキュメント > ガイド

この情報は役に立ちましたか？

Text-to-Speech のドキュメント

このページの内容

カスタム音声

Cloud Text-to-Speech API にカスタム音声が追加されました。この機能により、独自のスタジオ品質の音声録音を使用してカスタム音声モデルをトレーニングし、独自の音声を作成できます。カスタム音声を使用して、Cloud Text-to-Speech API で音声を合成できます。

Custom Voice を実装するには、セールスチームの担当者にお問い合わせください。

カスタム音声のサンプル

次の例を聴くことにより、カスタム音声のサンプルを確認できます。最初の音声例はオリジナルの音声です。その後、オリジナル音声に基づいて 2 つのカスタム音声の例を聴くことができます。

女性 - オリジナル音声	エラー	男性 - オリジナル音声	0:00 0:08
女性 - カスタム音声の例 1	0:00 0:03	男性 - カスタム音声の例 1	0:00 0:03

learn.microsoft.com

Microsoft | Learn ドキュメント トレーニング 認定資格 Q&A コード サンプル 評価 ショー イベント

検索 サインイン

Azure 製品ドキュメント アーキテクチャ Azure について学習 開発 リソース

ポータル 無料アカウント

タイトルでフィルター

- Speech Service のドキュメント
- 概要
  - 音声サービスとは
  - 新着情報
  - 言語と音声のサポート
  - リージョンのサポート
  - 価格
  - クォータと制限
- Speech Studio
- Speech CLI
- Speech SDK
- 音声テキスト変換
- テキストを音声に変換する
  - テキスト読み上げのドキュメント
  - テキスト読み上げの概要
  - テキスト読み上げのクイックスタート
  - 音声を合成する
  - バッチ合成 API
  - SSML を使用して合成を改善する
  - 音声合成の待機時間を短縮する
  - 口形素を使用して顔の位置を取得する
  - カスタム ニューラル音声
    - **カスタム ニューラル音声とは**
    - カスタム ニューラル音声 Lite
    - カスタム ニューラル音声を作成する
    - トレーニング データの種類
    - 音声サンプルを録音する方法
  - Audio Content Creation
  - テキスト読み上げに関する FAQ
- 音声翻訳
- 意図認識
- 話者認識
- キーワード認識
- シナリオの詳細
- 操作方法ガイド

Learn / Azure / Cognitive Services / Speech Service /

カスタム ニューラル音声とは

[アーティクル] • 2023/04/06 • 10 人の共同作成者

フィードバック

この記事の内容

カスタム ニューラル音声 (CNV) は、アプリケーション用に独自にカスタマイズした合成音声を作成できるようにするテキスト読み上げ機能です。 カスタム ニューラル音声を使用すると、人間の発話サンプルをトレーニング データとして提供することで、ブランドやキャラクターの音声を非常に自然な音声で作成できます。

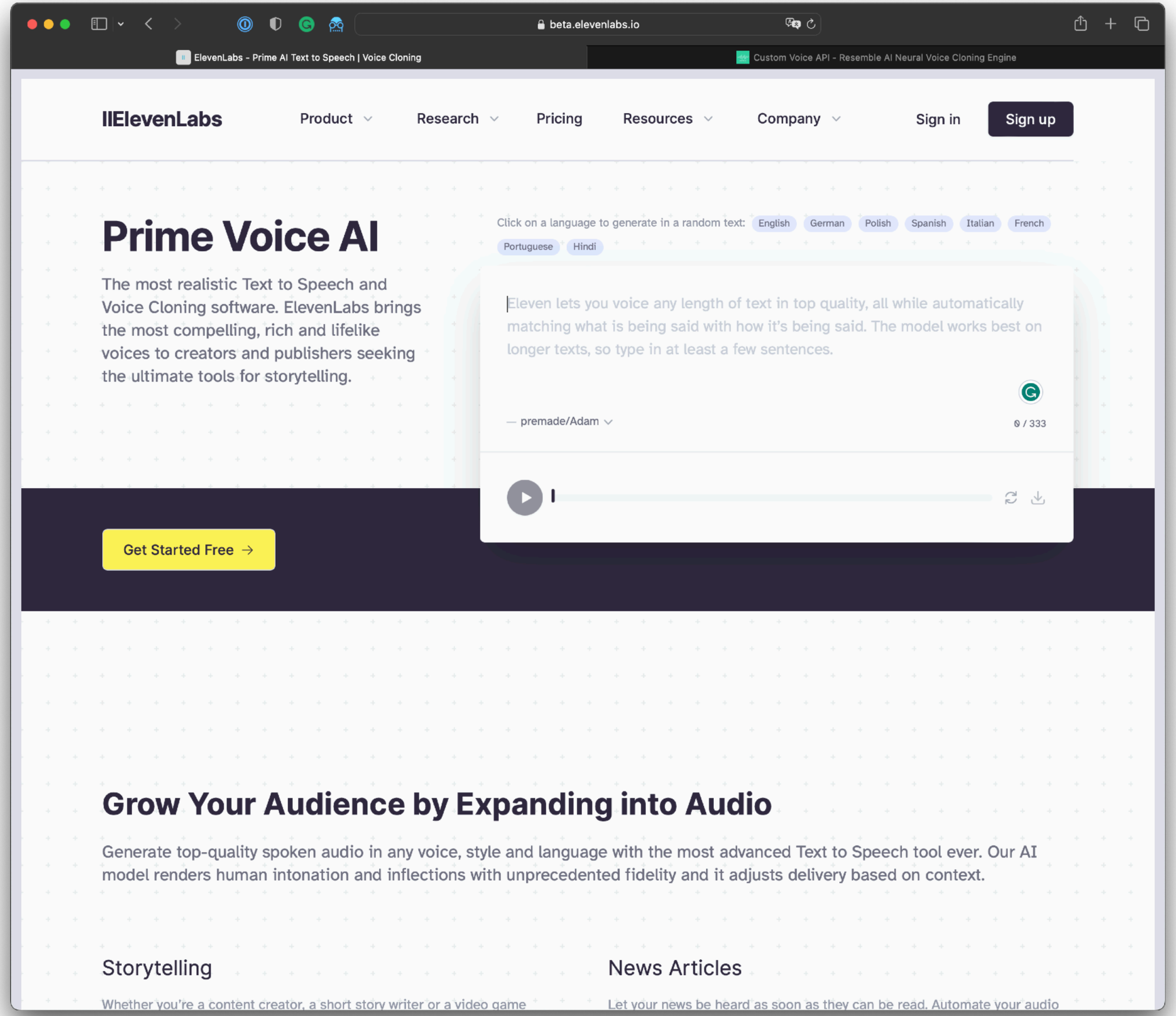
> ⓘ 重要
>
> カスタム ニューラル音声アクセスは、資格と使用条件に基づいて 制限されます。 入力フォーム で アクセスを要求します。
>
> より高品質な音声を作成するためのプロフェッショナルなレコーディングに投資する前に、どなたでもカスタム ニューラル音声 (CNV) Lite にアクセスして、CNV をデモして評価することができます。

テキスト読み上げは、サポートされている各言語の事前構築済みのニューラル音声で、追加設定なしで使用できます。 独自の音声が必要ではない場合は、ほとんどのテキスト読み上げシナリオで、事前構築済みのニューラル音声が非常に効果的に機能します。

カスタム ニューラル音声は、ニューラル テキスト読み上げテクノロジと多言語マルチスピーカーユニバーサル モデルに基づいています。 豊富な話し方の合成音声や、調整可能なクロス言語を作成できます。 カスタム ニューラル音声のリアルで自然な声は、ブランドや擬人化したコンピューターを表し、ユーザーが会話的にアプリケーションと対話することが可能になります。 カスタム ニューラル音声でサポートされる言語を参照してください。

しくみ

カスタム ニューラル音声を作成するには、Speech Studio を使用して、録音された音声とそれに対応するスクリプトをアップロードし、モデルをトレーニングして、音声をカスタム


beta.elevenlabs.io
ElevenLabs - Prime AI Text to Speech | Voice Cloning
Custom Voice API - Resemble AI Neural Voice Cloning Engine
IIElevenLabs
Product
Research
Pricing
Resources
Company
Sign in
Sign up
Prime Voice AI
The most realistic Text to Speech and Voice Cloning software. ElevenLabs brings the most compelling, rich and lifelike voices to creators and publishers seeking the ultimate tools for storytelling.
Click on a language to generate in a random text:
English
German
Polish
Spanish
Italian
French
Portuguese
Hindi
Eleven lets you voice any length of text in top quality, all while automatically matching what is being said with how it's being said. The model works best on longer texts, so type in at least a few sentences.
premade/Adam
0 / 333
Get Started Free →
Grow Your Audience by Expanding into Audio
Generate top-quality spoken audio in any voice, style and language with the most advanced Text to Speech tool ever. Our AI model renders human intonation and inflections with unprecedented fidelity and it adjusts delivery based on context.
Storytelling
Whether you're a content creator, a short story writer or a video game
News Articles
Let your news be heard as soon as they can be read. Automate your audio

resemble.ai

ElevenLabs - Prime AI Text to Speech | Voice Cloning

Custom Voice API - Resemble AI Neural Voice Cloning Engine

RESEMBLE.AI

PRODUCTS USE CASES PRICING SIGN IN REQUEST DEMO

Build Custom AI Voices using a Real-Time API

Rapidly build production-ready integrations with modern tools. Use Resemble's Custom Voice API to fetch existing content, create new clips and even build voices on the fly.

Read API Docs

Access web-generated content programmatically.

Build Voices with your own data.

Generate massive amount of speech content asynchronously.

Generate content in faster-than-realtime in sync with extremely low latency.

iPhoneのアクセシビリティ機能

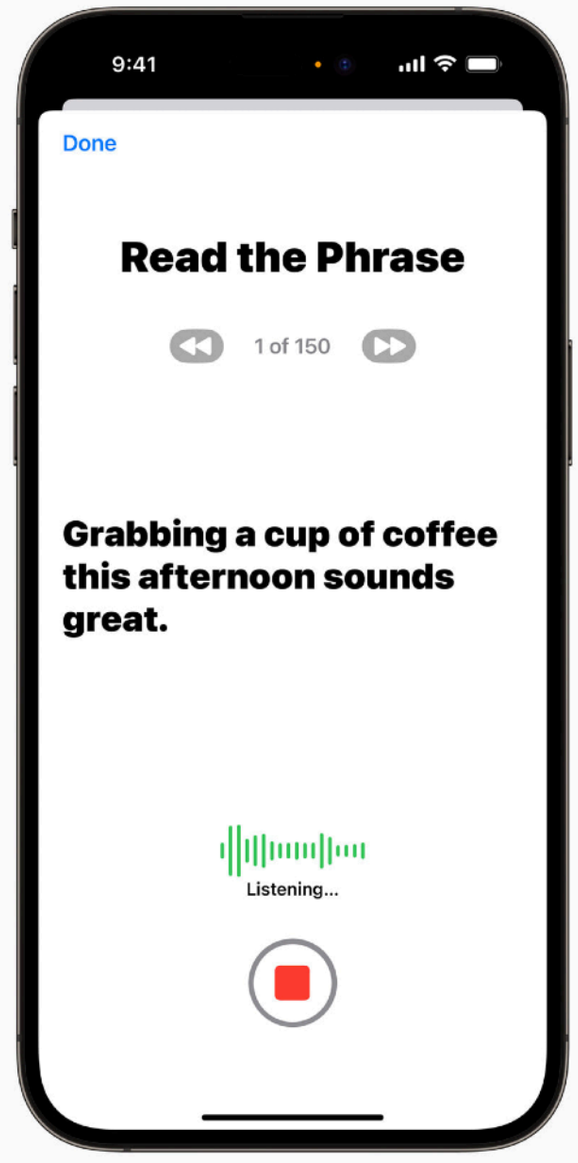
モバイル ALS患者が充実した生活を送れるよう支援

iPhoneで“自分そっくりの合成音声”を作れる ーソナルボイス」、アップルが予告

多根清史 @bigburn、 Yahoo!ニュース 個人
2023/05/17 11:58

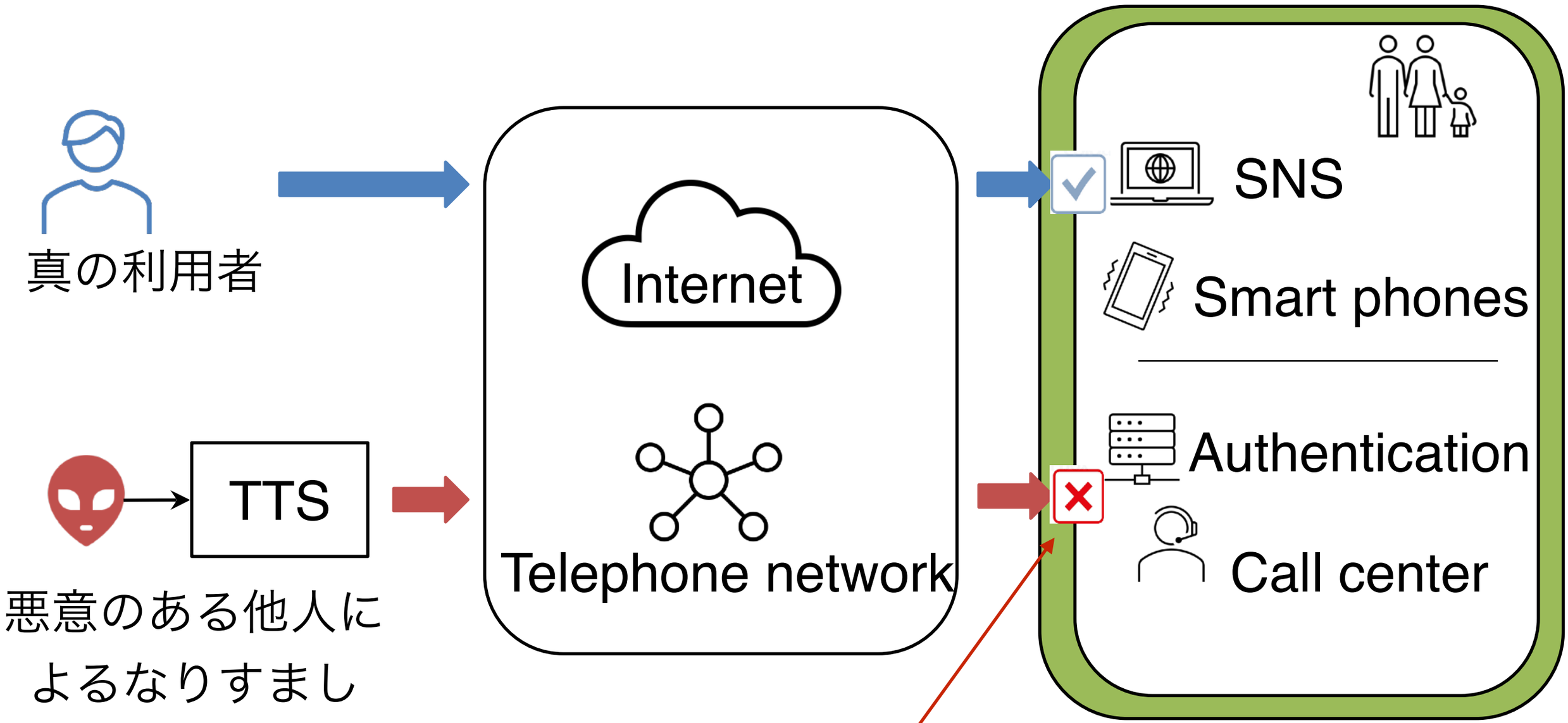
Share

Users can create a Personal Voice by reading along with a randomized set of text prompts to record 15 minutes of audio on iPhone or iPad. This speech accessibility feature uses on-device machine learning to keep users’ information private and secure, and integrates seamlessly with Live Speech so users can speak with their Personal Voice when connecting with loved ones.[1]



音声のディープフェイク検知
ーデータベースー

TTSやVCによるなりすまし・ディープフェイク対策



ディープフェイク検知(本講義での呼び方)

Countermeasure, Anti-spoofing, Presentation attack detection(PAD)とも呼ばれる

ディープフェイク検知の基本方針

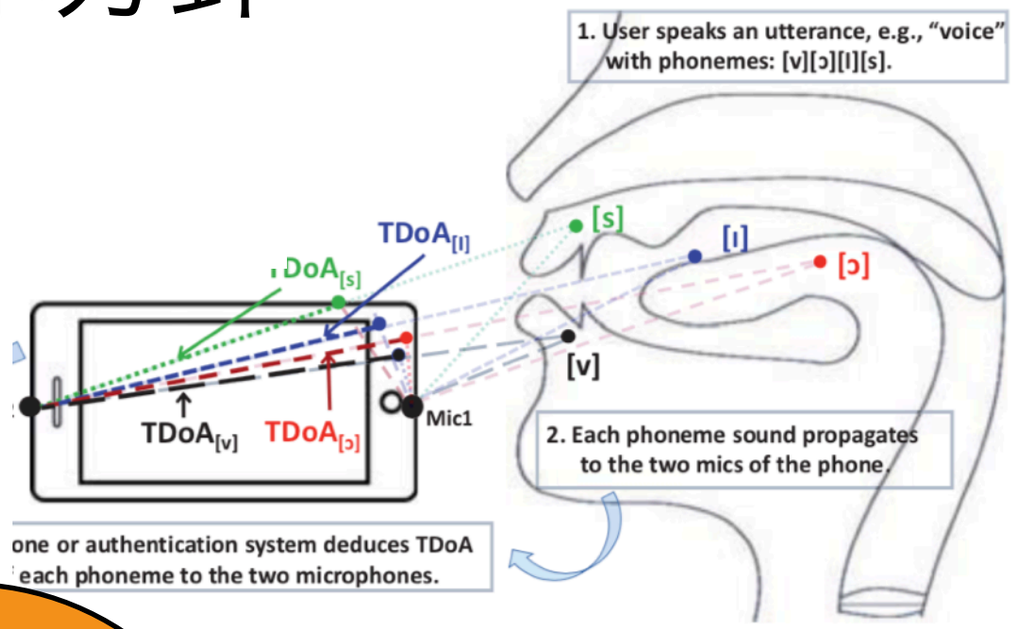
- 何を学習させるか？

ASV spoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

SASV challenge 2022

Artefacts
(speech community)

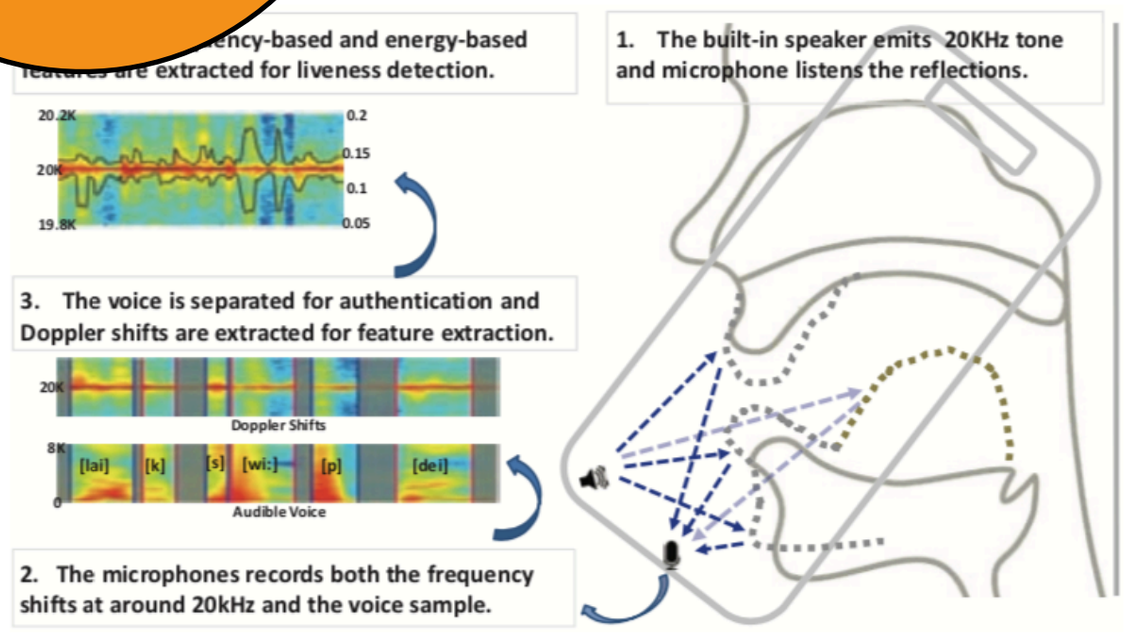
Liveness evidence
(Security community)



Linghan Zhang, Sheng Tan, Jie Yang, Yingying Chen, VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones 23rd ACM Conference on Computer and Communications Security (CCS 2016) Vienna, Austria, October 2016

Linghan Zhang, Sheng Tan, Jie Yang. "Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication". 24th ACM Conference on Computer and Communication Security (CCS 2017).

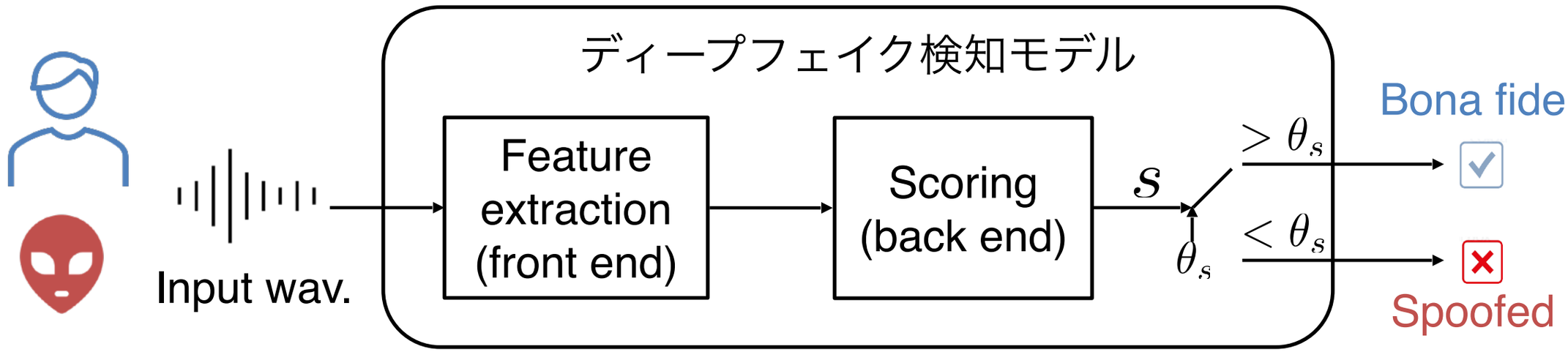
ADD
Audio Deep Synthesis Detection



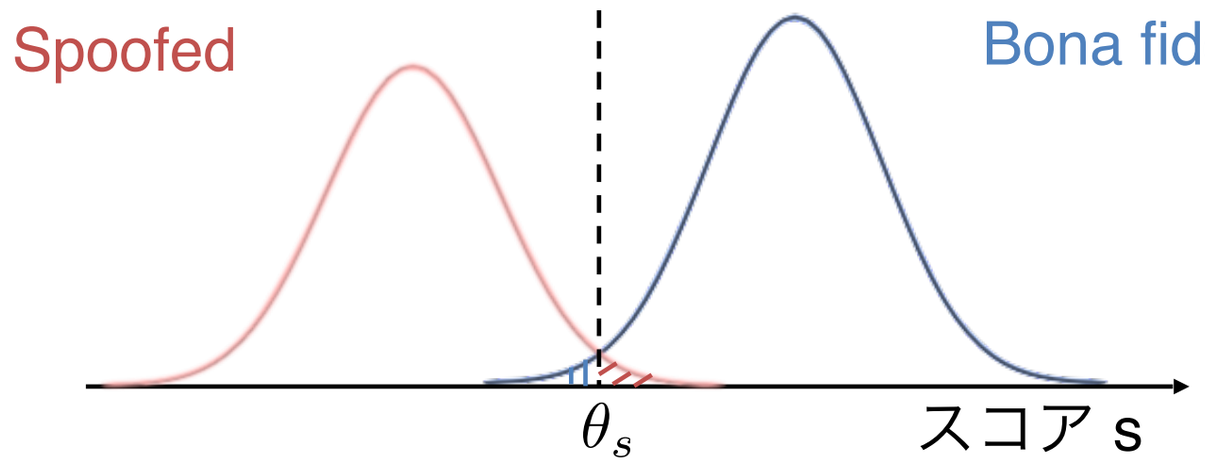
- 本講義ではArtefactsを中心に紹介

ディープフェイク検知は二値分類として定義できるが、そう簡単な話ではない

- 二値分類

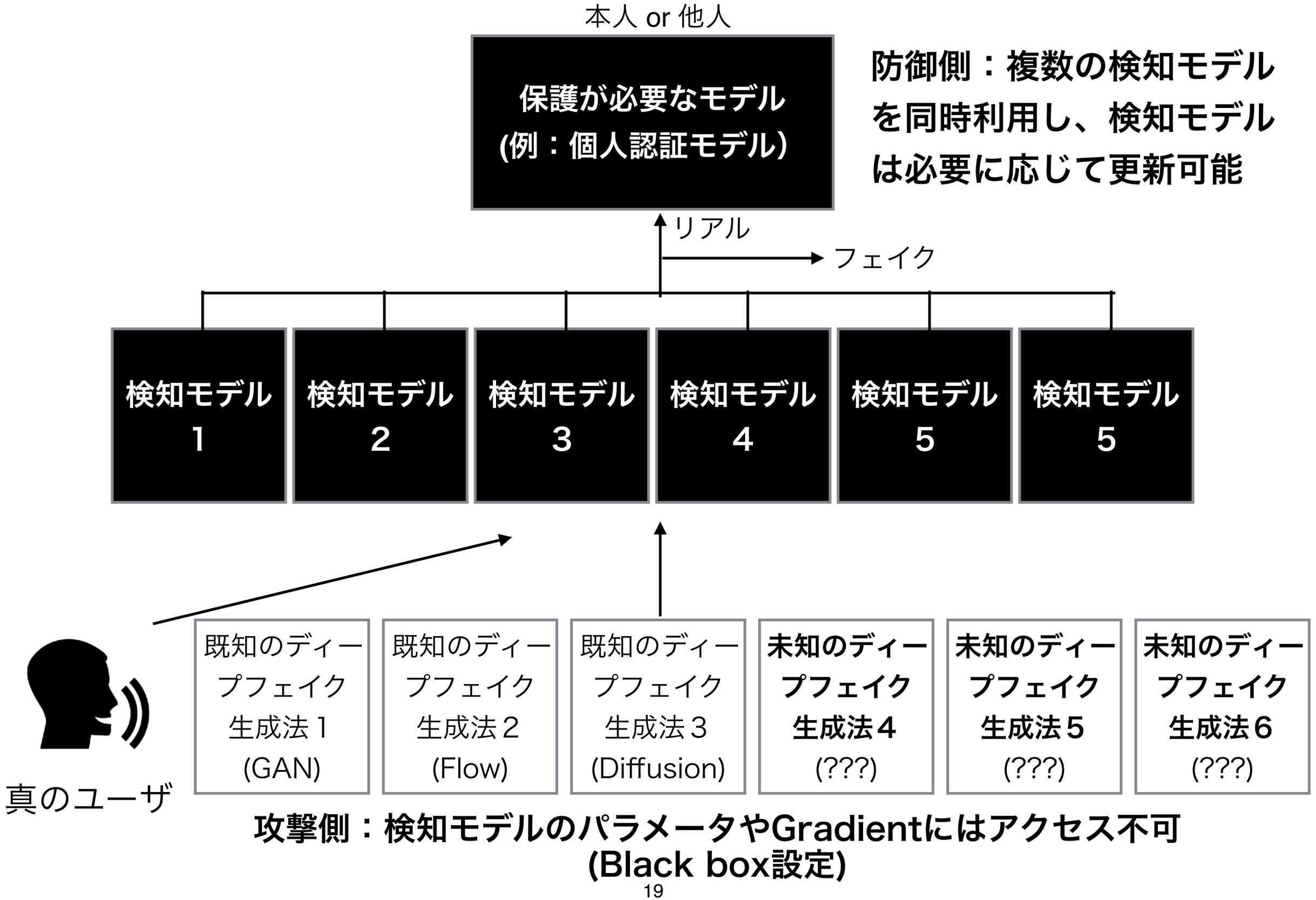


- 検知モデルのスコア



- ディープフェイク生成手法は**多種多様**
- **検知モデルを学習した後にも異なるディープフェイク生成ツールが出現**
  - **事前に全ての生成手法を知る事はできない**

ディープフェイク検知モデルの利用シナリオ



ディープフェイク検知モデル学習用データベース

- ディープフェイク検知モデルは合成音声と自然音声の大量データから学習
- Google, NTT, DFKI等との複数組織との協力により大規模データベースを2019年に構築し公開
- ニューラルボコーダ、Voice Conversion Challengeでベストな手法を含む
- 評価データは主に既知生成手法ではなく、派生・未知の生成手法により構成
- 一般無償公開(ASVspoof 2019 LA dataset) 現在60万回ダウンロード

	Number of trials			Acoustic Model	Waveform generation	Category
	Train	Dev	Eva.			
A01	3800	3716	-	LSTM-RNN	WaveNet-vocoder	TTS
A02	3800	3716	-	LSTM-RNN	WORLD-vocoder	TTS
A03	3800	3716	-	Feedforward NN	WORLD-vocoder	TTS
A04	3800	3716	-	Unit-selection	Waveform concate	TTS
A05	3800	3716	-	Conditional-VAE	WORLD-vocoder	VC
A06	3800	3716	-	GMM-UBM	Spectral filtering	VC
A07	-	-	4914	LSTM-RNN	WORLD & GAN filtering	TTS
A08	-	-	4914	LSTM-RNN	Neural source-filter model	TTS
A09	-	-	4914	LSTM-RNN	Vocaine-vocoder	TTS
A10	-	-	4914	Tacotron	WaveRNN	TTS
A11	-	-	4914	Tacotron	Griffin-Lim	TTS
A12	-	-	4914	-	WaveNet-based TTS	TTS
A13	-	-	4914	Moment matching NN	Waveform filtering	TTS-VC
A14	-	-	4914	LSTM-RNN	STRAIGHT-vocoder	TTS-VC
A15	-	-	4914	LSTM-RNN	WaveNet-vocoder	TTS-VC
A16	-	-	4914	Unit-selection	Waveform concate	TTS
A17	-	-	4914	Conditional-VAE	Waveform filtering	VC
A18	-	-	4914	i-vector & GMM	Glottal vocoder	VC
A19	-	-	4914	GMM-UBM	Spectral filtering	VC

Train & dev

Evaluation

色の意味
既知
派生
未知

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection”, Interspeech 2019, Sept 2019

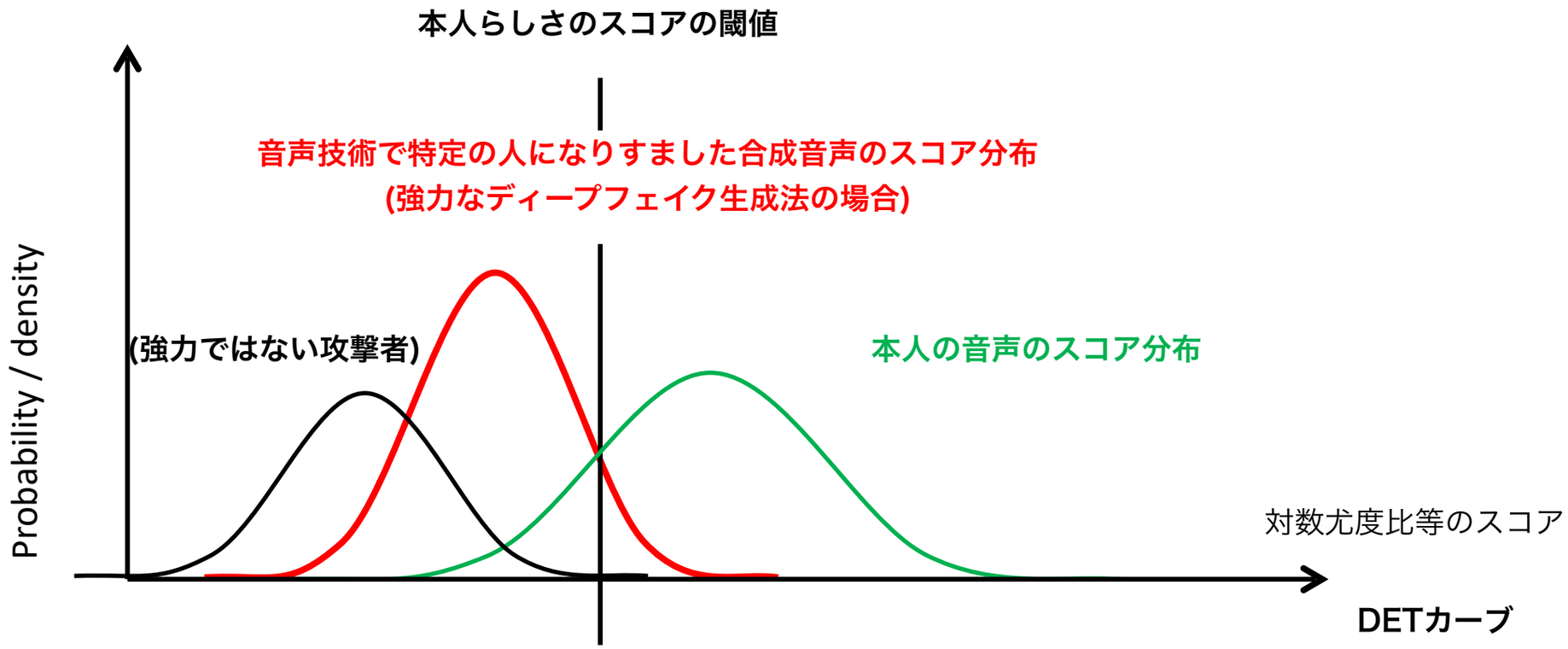
音声サンプル

	Human voice	Synthesized voice A10
Speaker 1	♪	♪
Speaker 2	♪	♪
Speaker 3	♪	♪

Shen, Jonathan, et al. "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

音声のディープフェイク検知
ー指標ー

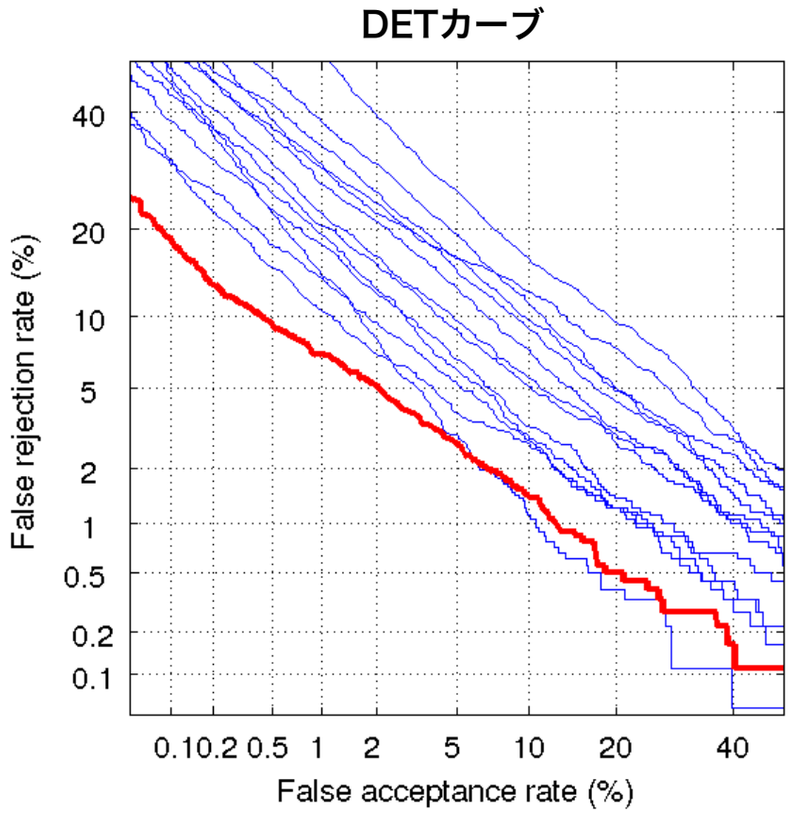
評価指標１：本人受理・誤受理確率



Trial	Decision: Accept	Decision: Reject
本人の音声	Correct accept	**False reject (FRR)**
合成音声	**False alarm (FAR)**	Correct reject

合成音声が本人の音声に近い場合、FARが増加

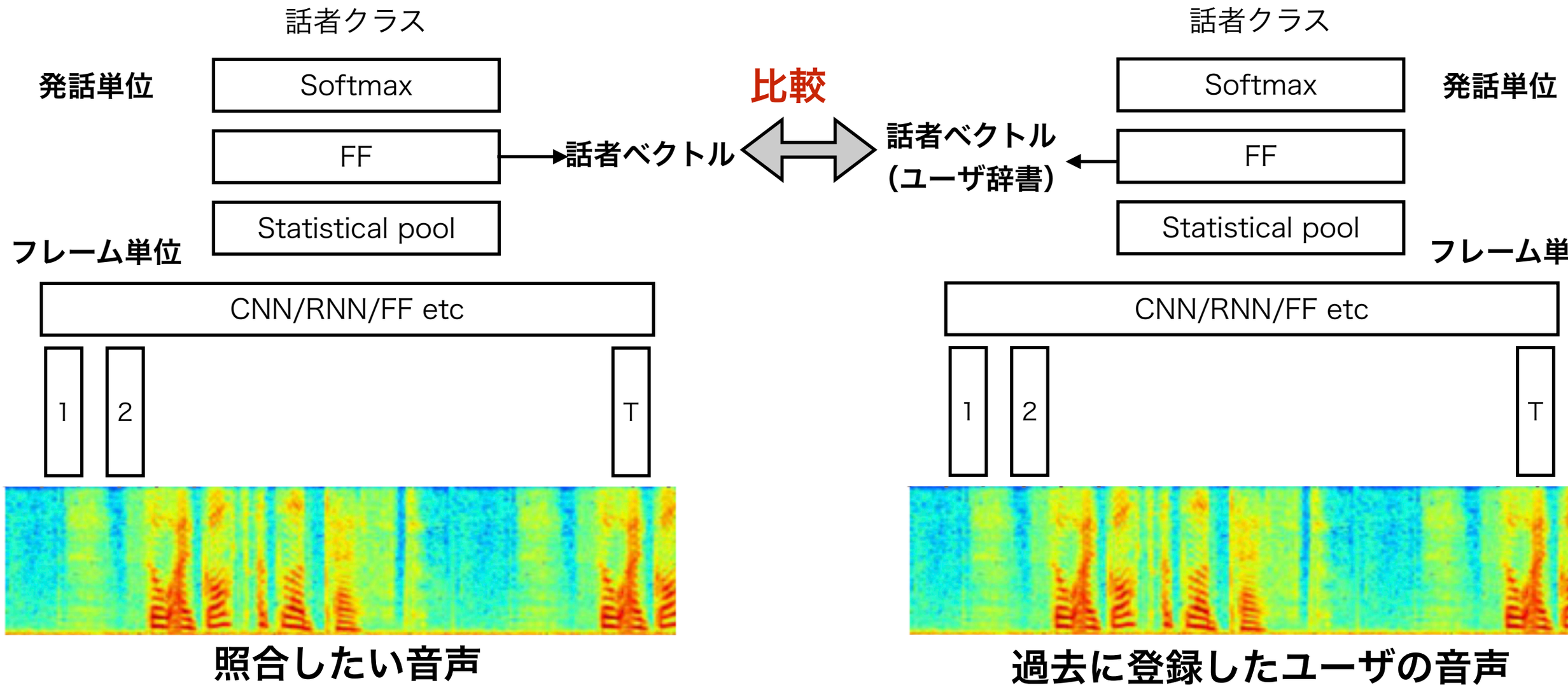
スコアの閾値を調整、
等価誤り率
EER (FAR = FRR)を計算
↓
合成音声が本人の音声に近いほどEERは増加！



Spoofing and countermeasures for speaker verification: a survey
Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Li
Speech Communication 66 130-153 2015年3月

評価指標１：話者ベクトル

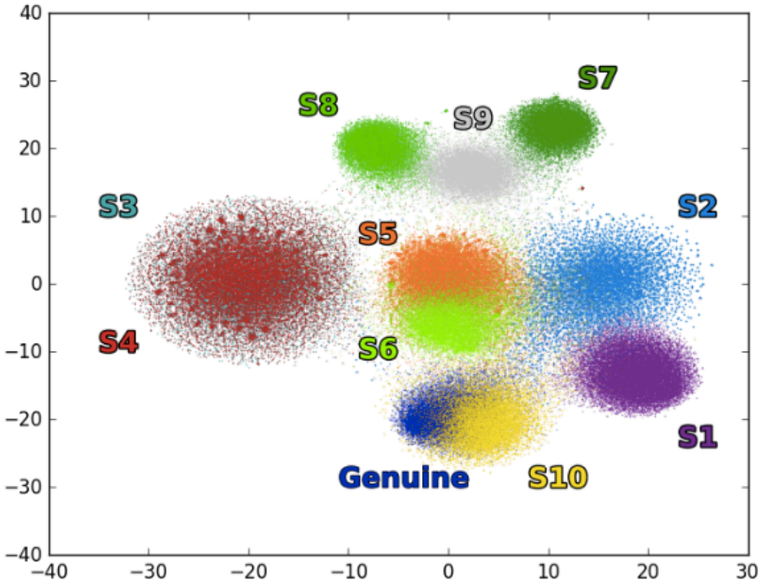
- ニューラルネットワークを利用した話者埋め込みベクトル
- 話者認識の標準技術



評価指標１：本人と合成音声の話者ベクトルを可視化

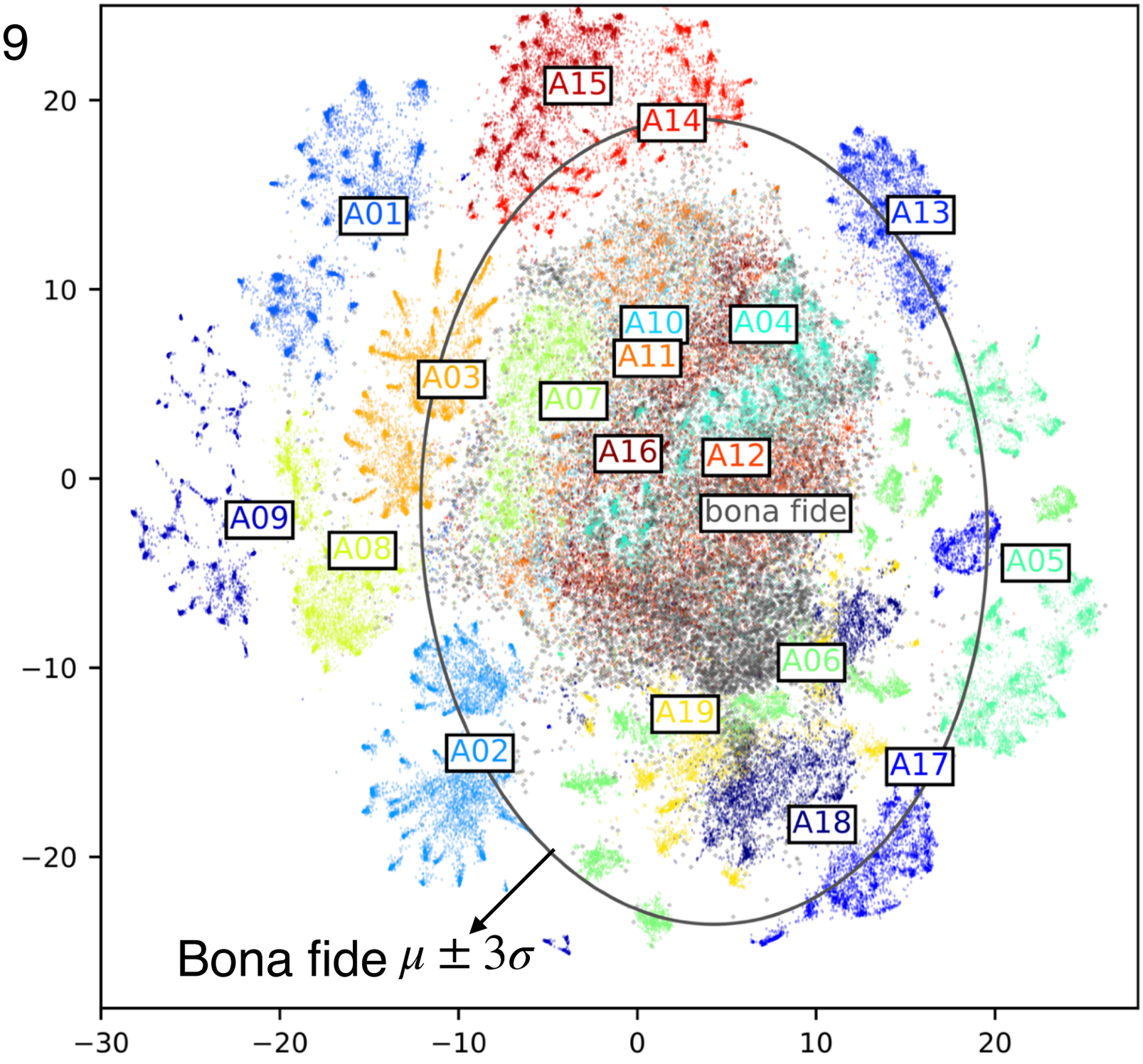
ASVspoof 2019

- 2015年



ASVspoof 2015

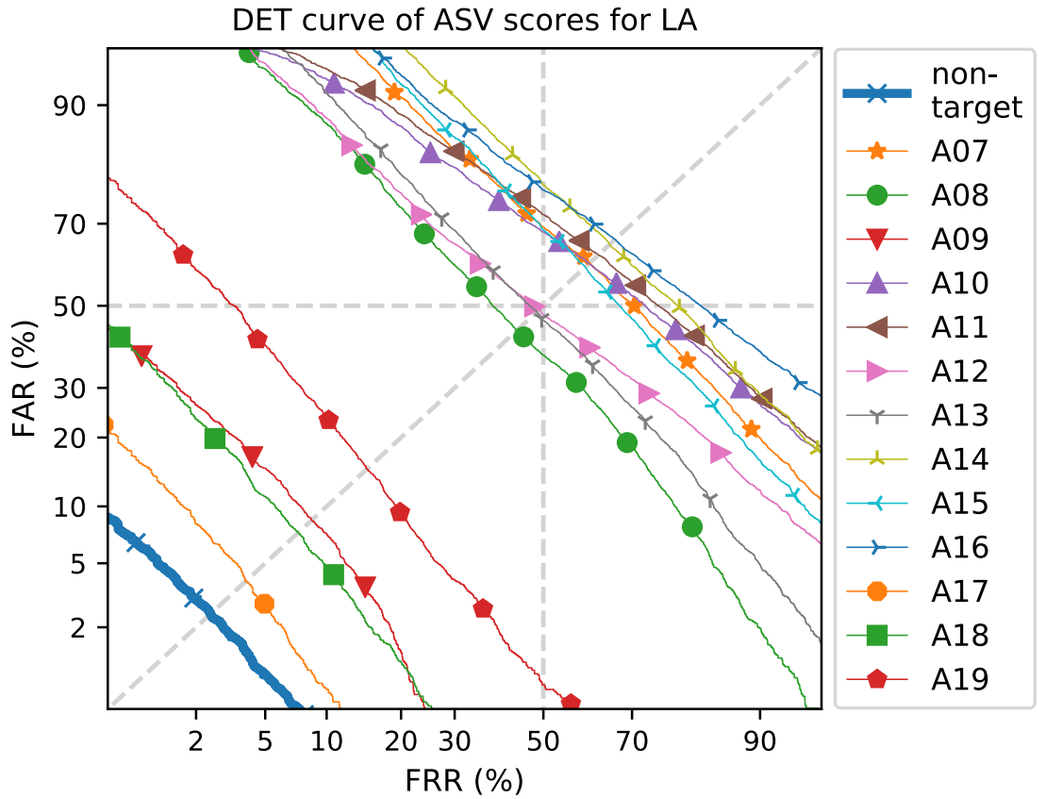
聴覚上も識別困難な
ケースあり



The ASVspoof 2019 database
Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, Zhen-Hua Ling, Computer Speech & Language, 2020

評価指標１：データベース内の合成音声の本人らしさ

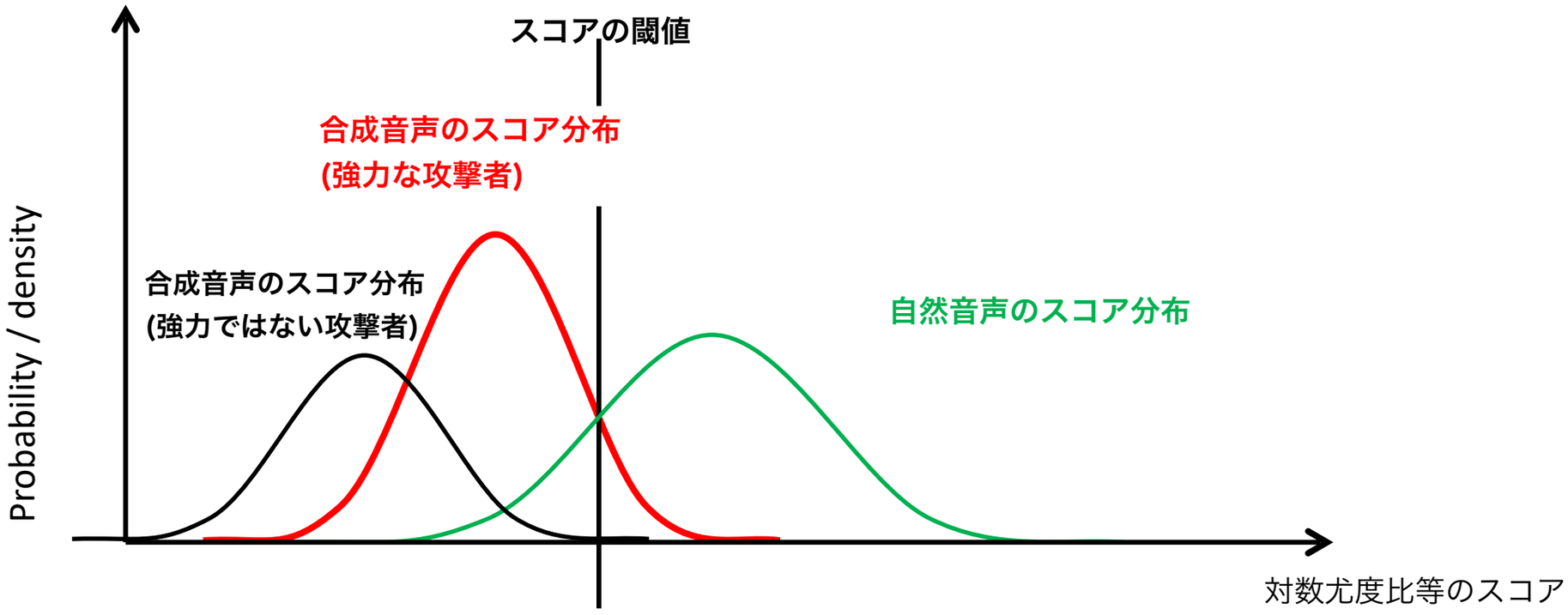
- ASVspoof 2019 LA datasetのサンプルを利用して、声による話者認識システムを実際に攻撃
- 攻撃対象のシステム
  - 数千人の話者を含むVoxCelebデータベースを利用して学習された当時SoTAだった話者照合システム（x-vector / PLDA）



- 合成音声を本人と間違って受理する確率が大幅に増加
- 一部の合成システム(A16)は**本人よりも本人らしさが高いと判定**

評価指標２：ディープフェイク検知モデルの性能

- ディープフェイク検知モデルは本人類似度ではなく、合成音声にのみ存在するアーティファクトを特徴量として利用するので、ディープフェイク検知モデル単独の指標も定義する
- ディープフェイク検知モデルの等価誤り率も同様に定義可能



Trial	Decision: Accept	Decision: Reject
自然音声	Correct accept	**False reject (FRR)**
合成音声	**False alarm (FAR)**	Correct reject

合成音声のアーティファクトを適切に学習し、自然音声との識別に成功するとEERは減少

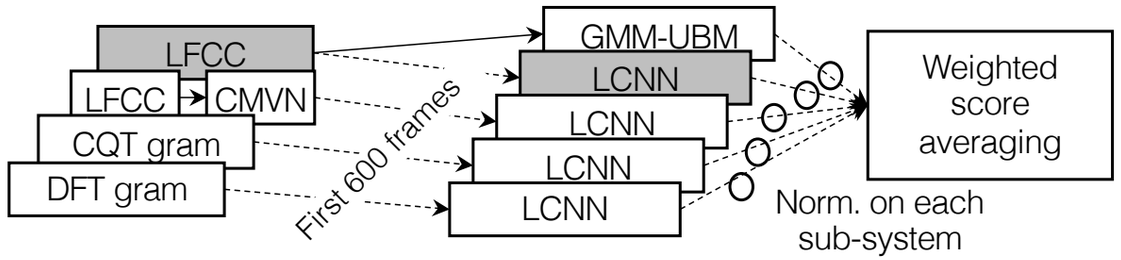
↓

検知モデルの性能が高いとEERが低い

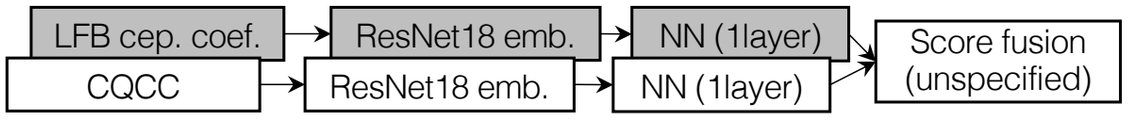
評価指標２：ディープフェイク検知モデルの例

- ASVspoof 2019データベースを利用して構築された50種類以上のディープフェイク検知モデルの一例

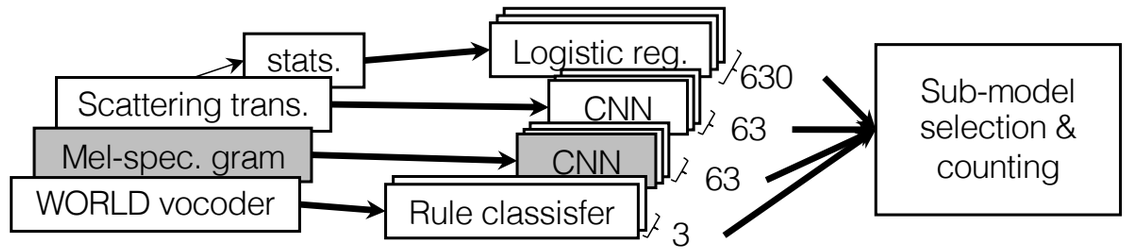
☆★ T45



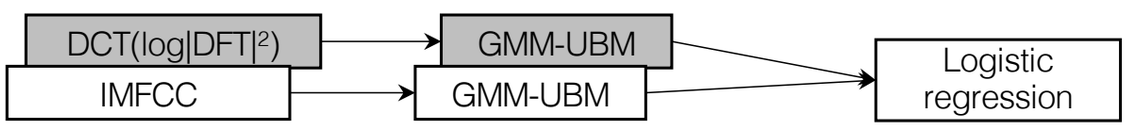
☆★ T24



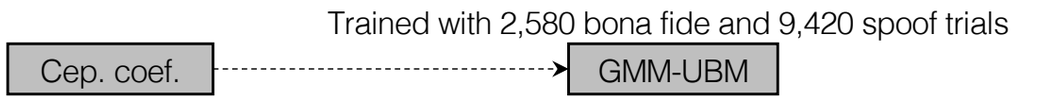
☆★ T39



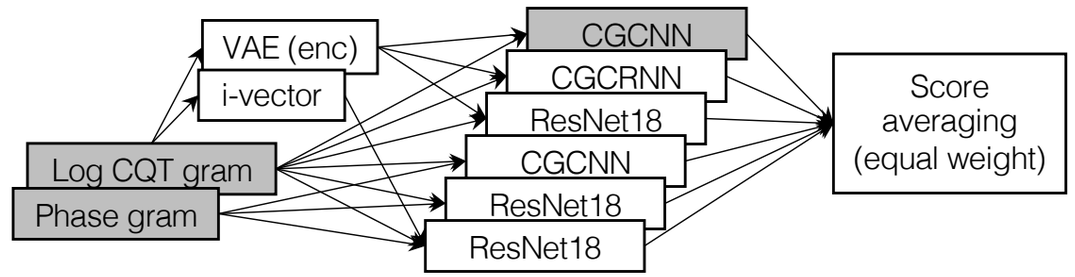
☆ T01



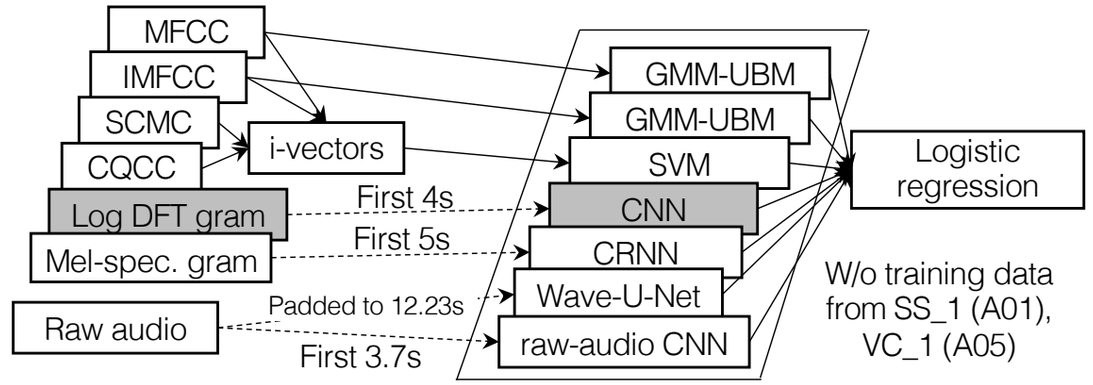
☆ T04



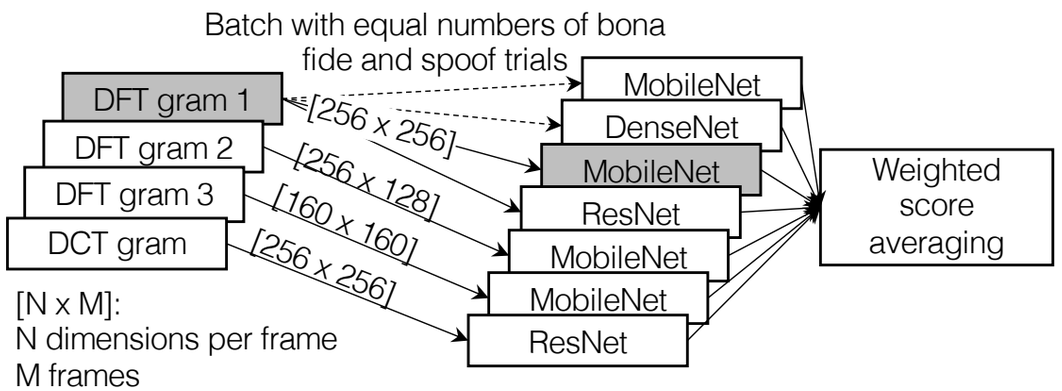
★ T50



★ T60

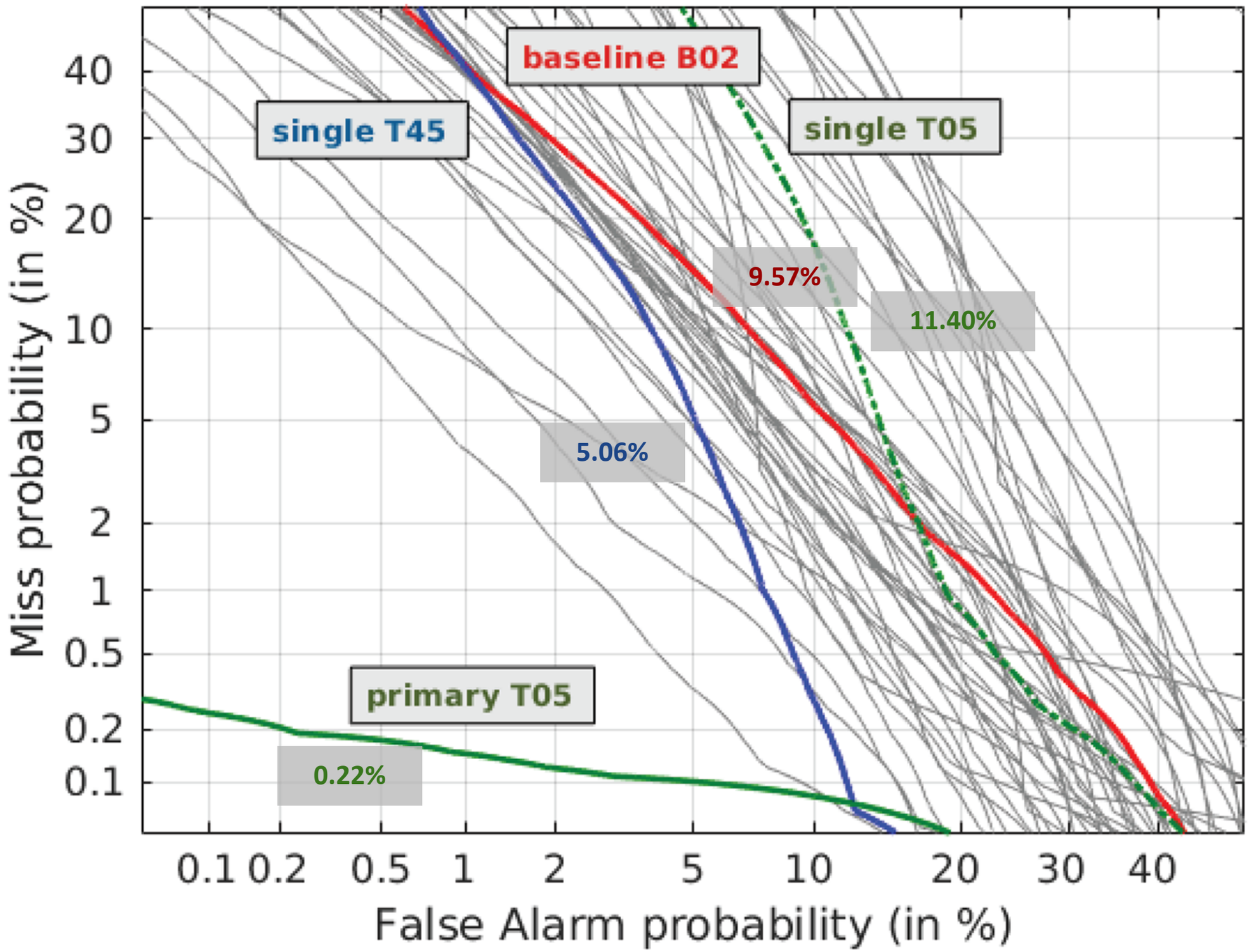


★ T05



Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech", IEEE Transactions on Biometrics, Behavior, and Identity Science

評価指標２：50種類以上のディープフェイク検知モデルはどの程度ディープフェイクをただしく検知できるのか？



評価指標３ ディープフェイク検知・照合の全体エラー：

Tandem-Decision Cost Function (t-DCF)

AI attackers (TTS/VC) → P_a

Human attackers (Non target) → P_b

True users → アーティファクト有無 → ディープフェイク検知

Accept (Real) → 本人らしさ → 音声による話者照合

Accept (Target)

True users → P_d → Reject (Fake)

P_c → Reject (Non target)

$$t-DCF = \sum_{i}^{a,b,c,d} C_i \pi_i P_i$$

C_i Cost for a specific type of errors

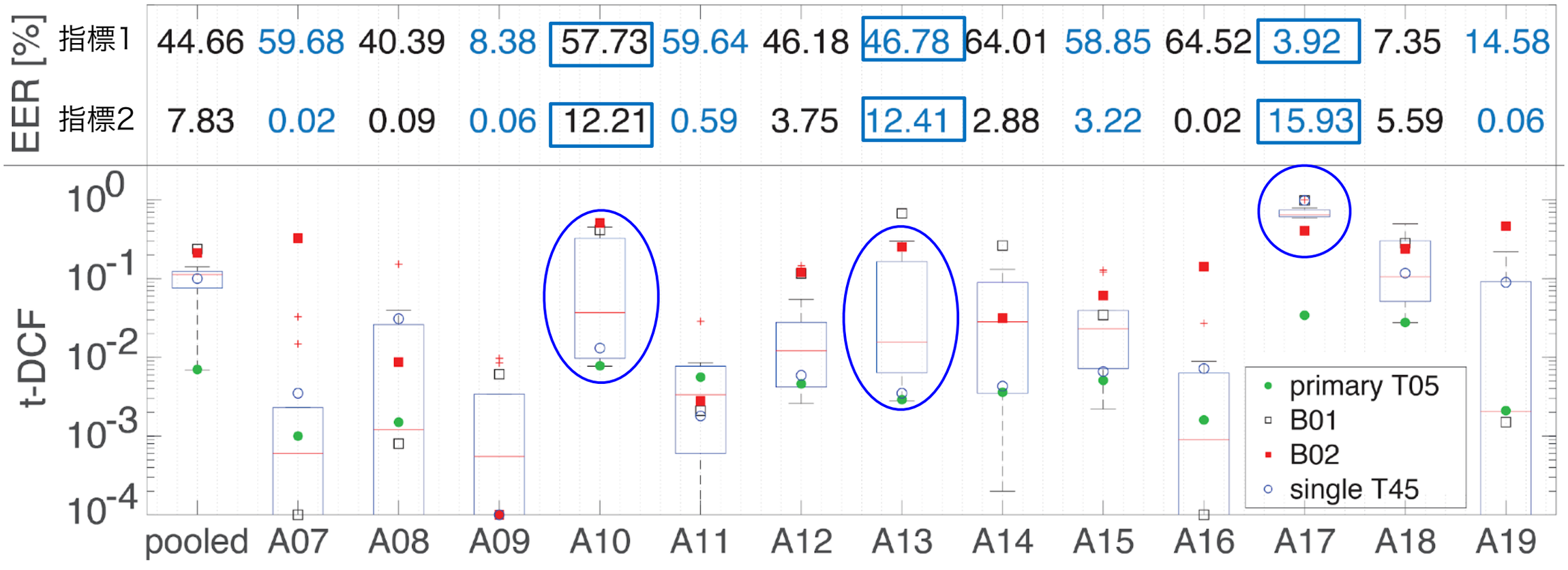
π_i Prior of a specific type of errors

Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, Douglas A. Reynolds "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals" IEEE/ACM Transactions on Audio, Speech, and Language Processing

指標３：全体のエラー分析

- **指標１：本人らしさ, EER [%]**
  - 値が高いと本人らしさが高いと判断
- **指標２：ディープフェイク検知モデル単体の性能, EER [%]**
  - 値が少ないとディープフェイク検知が正しく出てきていると判断
- **t-DCF：統合指標**
  - 値が少ないとディープフェイク検知と後段の個人認証の全体のエラーが少ないと判断

- 50種類以上のディープフェイク検知モデルの中で、t-DCF指標でトップ10だった検知モデルの性能を、ディープフェイク生成手法(A07-A19) 毎にBoxplot表示



ここまでのまとめ

- 予想通り、正しくディープフェイク検知できないモデルが多い
  - 評価セットは既知のディープフェイク生成手法ではなく、主に派生・未知のディープフェイク生成手法により構成されているため

- 全ての指標で概ね正しく検知できたモデルもあった
  - primary T05が素晴らしい検知結果
    - 後ほど詳しく分析

- ディープフェイク生成手法毎の分析から検出が難しかったシステムも一部あった (VC A17)