音声生成AI・ボイスクローンの悪用対策

ディープフェイク検知からアクティブディフェンスまで

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 (H-index=79 2025/06/04)

- 経歴

- 最初の約10年間：音声合成
  - 2002～2006 @ 東京工業大学 博士課程
    - 隠れマルコフモデル、信号処理、**音声合成のパーソナライゼーション**
  - 2006～2010 @ 英国エジンバラ大 ポスドク研究員として多数のEUプロジェクトに参加
    - ボイスクローン自動化研究、**1000人分のボイスクローン、**音声明瞭性変換
  - 2011～2013 @ 英国エジンバラ大 助教相当
    - **VCTKデータベース、ボイスクローンによる攻撃可能性に気がつく**
- 次の約10年間：音声のセキュリティとプライバシー
  - 2013～2019 @国立情報学研究所(NII) 准教授
    - 2015年：ASVspoof2015データベース公開
    - 2018年：**ディープフェイク顔映像検知モデルMesoNet発表**
    - 2018年～2023年：**日仏JST CREST ”VoicePersonae: 声のアイデンティティクローニングと保護”**
    - 2019年：MOSNet（後程VoiceMOSへ発展） & ニューラルソースフィルタモデル 発表
  - 2019～現在 @ 国立情報学研究所 教授・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
    - 2020年：話者匿名化とVoice Privacy Initiative
    - 2021年～現在**：ディープフェイク検出プログラムSynthetiq Vision開発**し、ライセンス開始
    - 2023年：ディープフェイク検知に関しIEEE Biometric Councilの5-year Highest Impact Award受賞
    - 2024年4月～現在：**JST AIP加速課題 “フェイクメディア検出技術の社会実装加速と普及” 代表**
    - 2024年10月～現在：**NIIの研究分担者として偽情報対策のNEDO Kプロに参画**　自動ファクトチェックなど
- 趣味的研究
  - 落語音声合成、MIDI-to-ピアノ音、MIDI-to-ギター音

- 研究グループ：yamagishilab.jp

- プロジェクト准教授1名、プロジェクト助教1名 ポスドク研究員5名、海外大学インターン生3名、博士学生1名

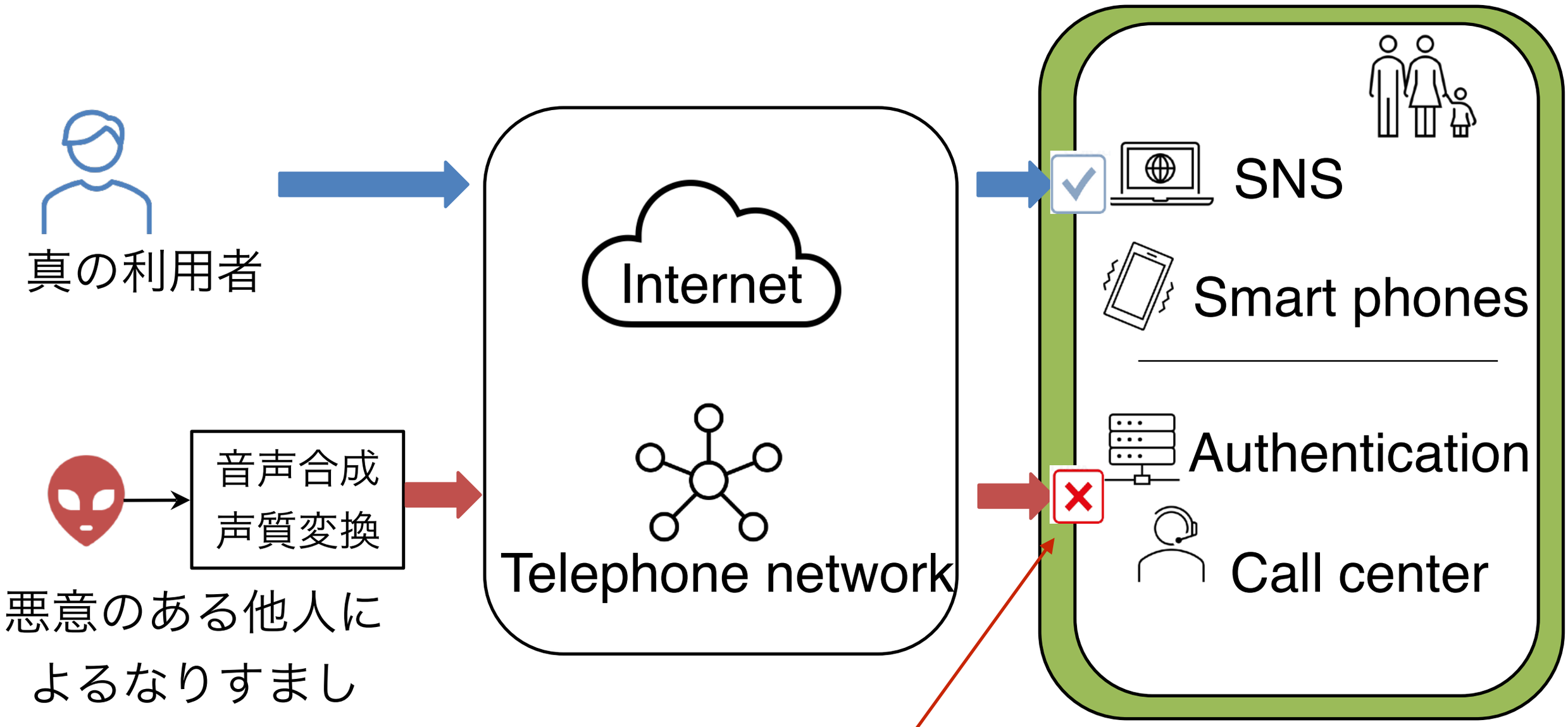
本講演の構成

- **パート１：ディープフェイク検知**
  - 1-1 データベースと指標
  - 1-2 検知モデルの何が重要？
  - 1-3 実環境での評価
  - 1-4 未知の生成手法の検知
  - 1-5 安定運用継続のための学習データ自動選択
- **パート２：プロアクティブディフェンス**
  - 2-1 音声の透かし技術
  - 2-2 話者匿名化

パート1

ディープフェイク音声の検知

ボイスクローンによるなりすまし・ディープフェイク対策



ディープフェイク検知(本資料での呼び方)

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

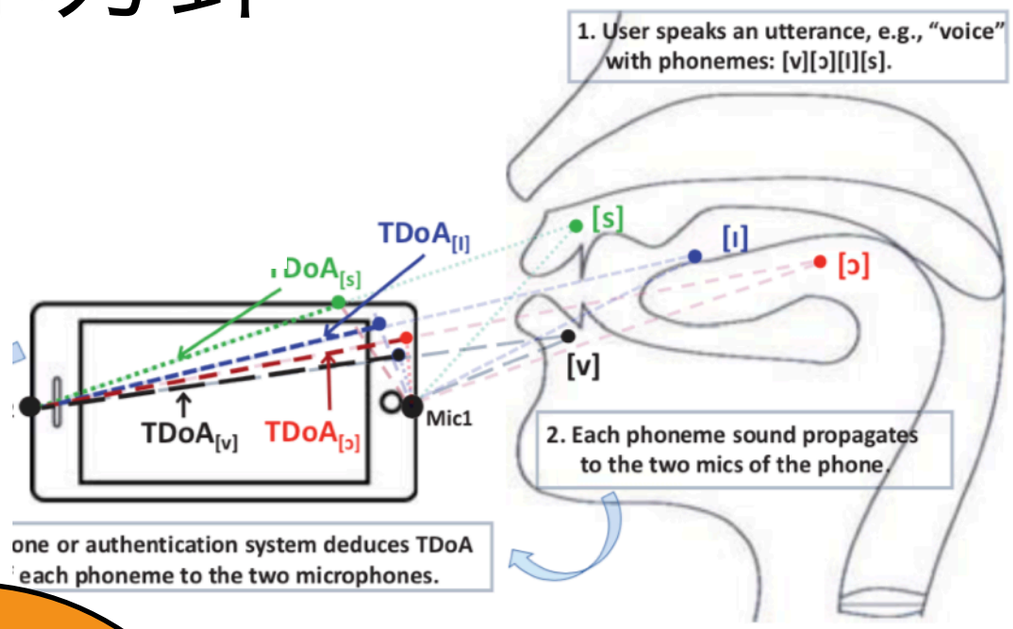
SASV challenge 2022

Artefacts

Liveness evidence

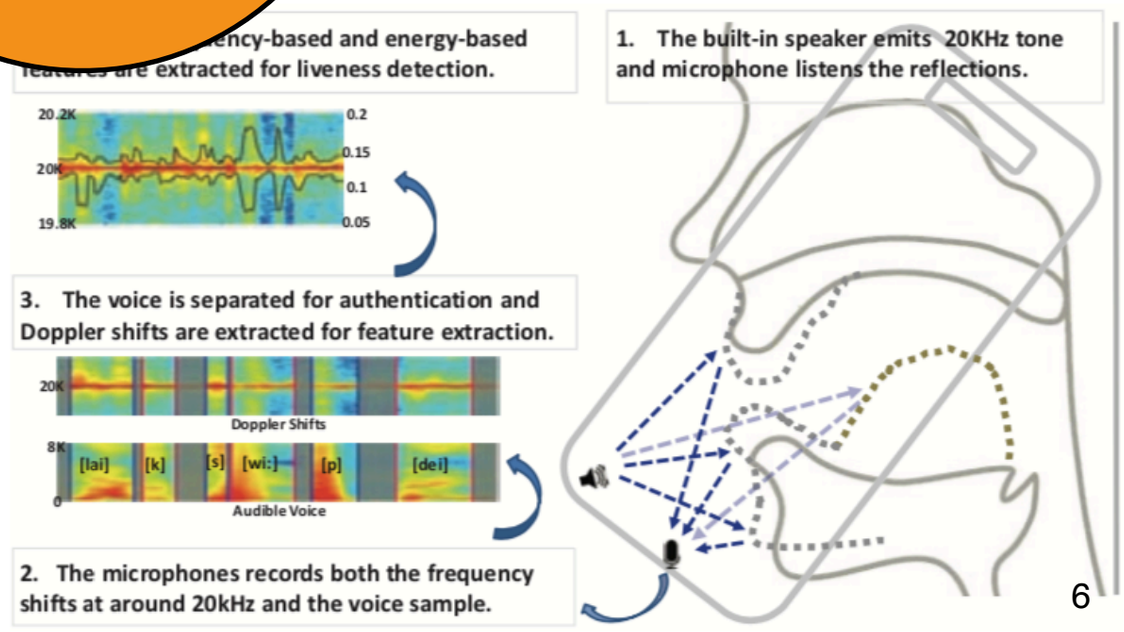
ADD
Audio Deep Synthesis Detection

- 本資料ではArtefactsを中心に紹介



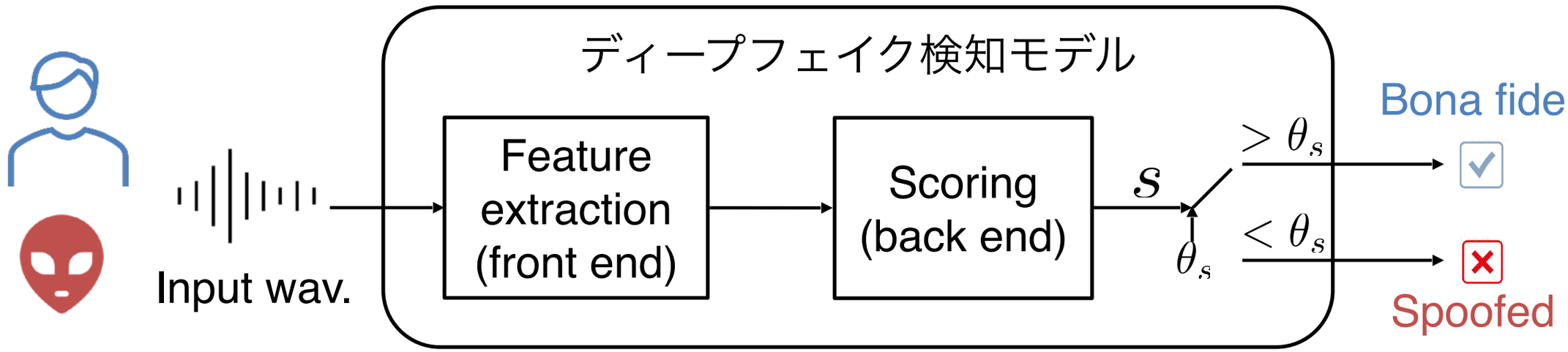
Linghan Zhang, Sheng Tan, Jie Yang, Yingying Chen, VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones 23rd ACM Conference on Computer and Communications Security (CCS 2016) Vienna, Austria, October 2016

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).

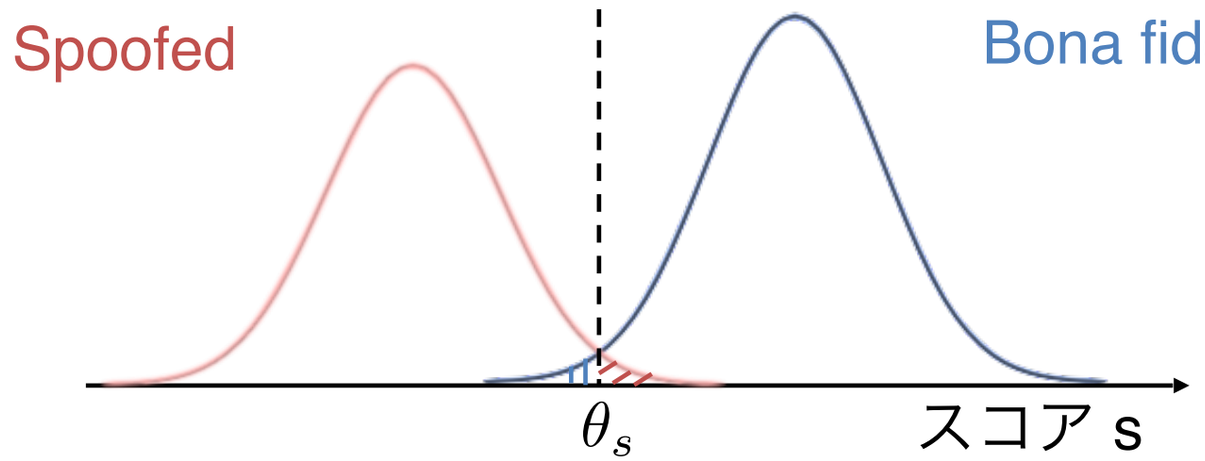


ディープフェイク検知は二値分類の様だが、、

- 二値分類

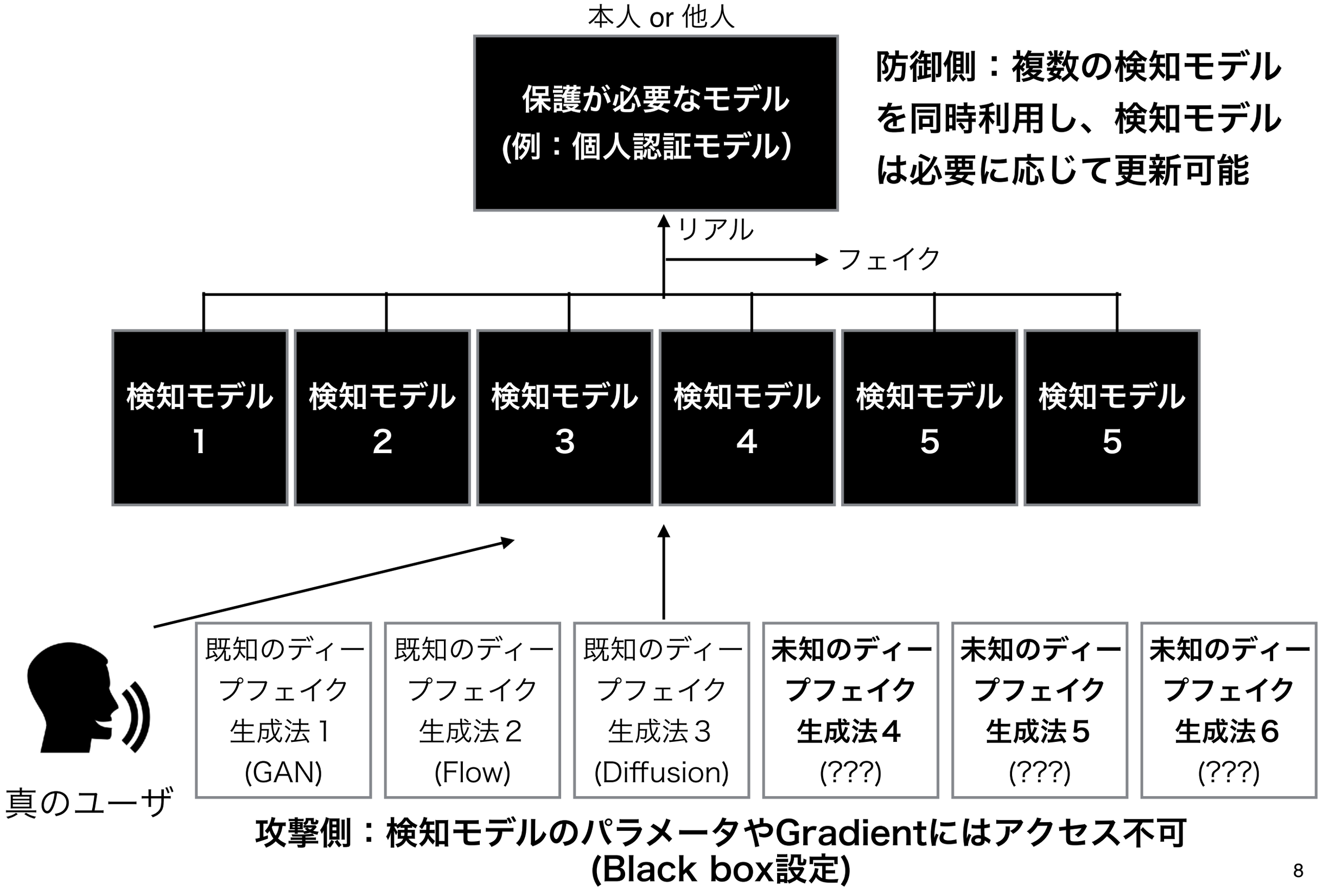


- 検知モデルのスコア



- 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→強力なドメインシフト**

ディープフェイク検知モデルの利用シナリオ



パート 1-1
音声のディープフェイク検知
ーデータベースと指標ー

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

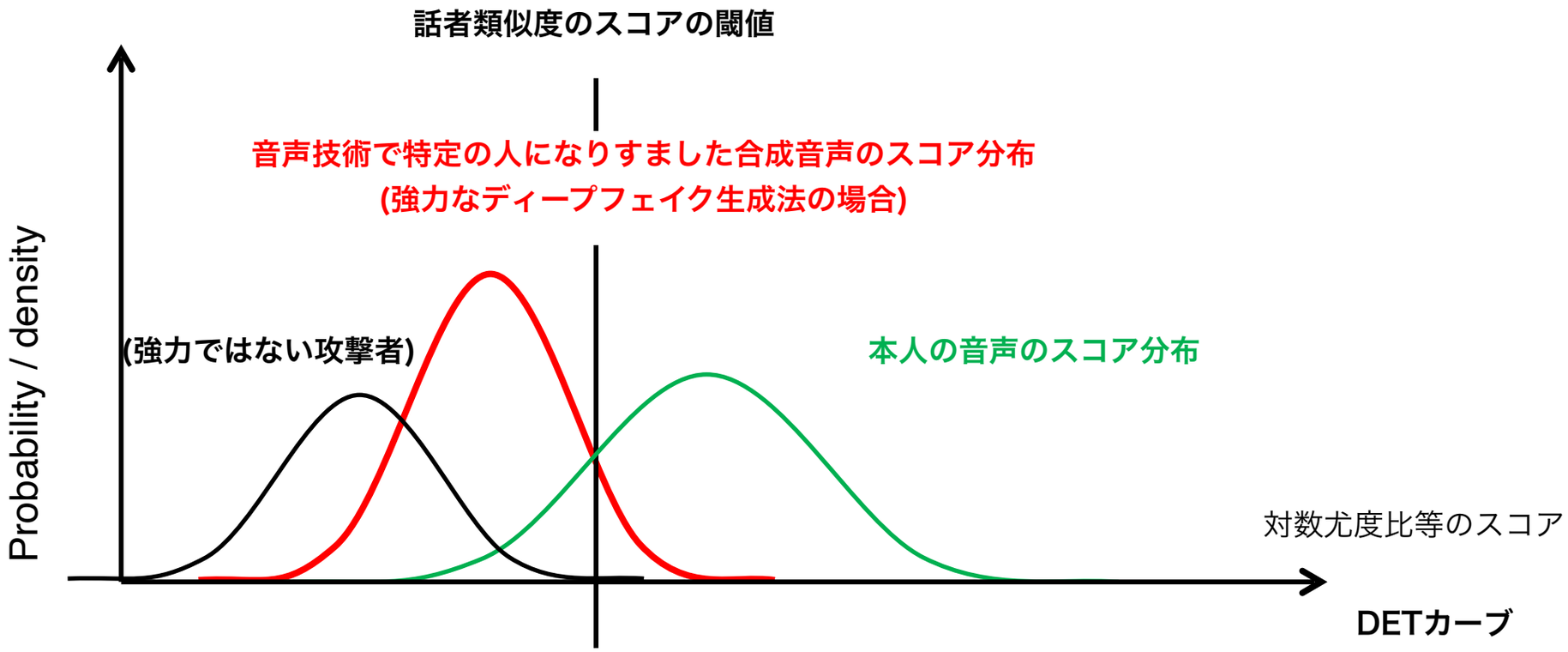
Train & dev

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

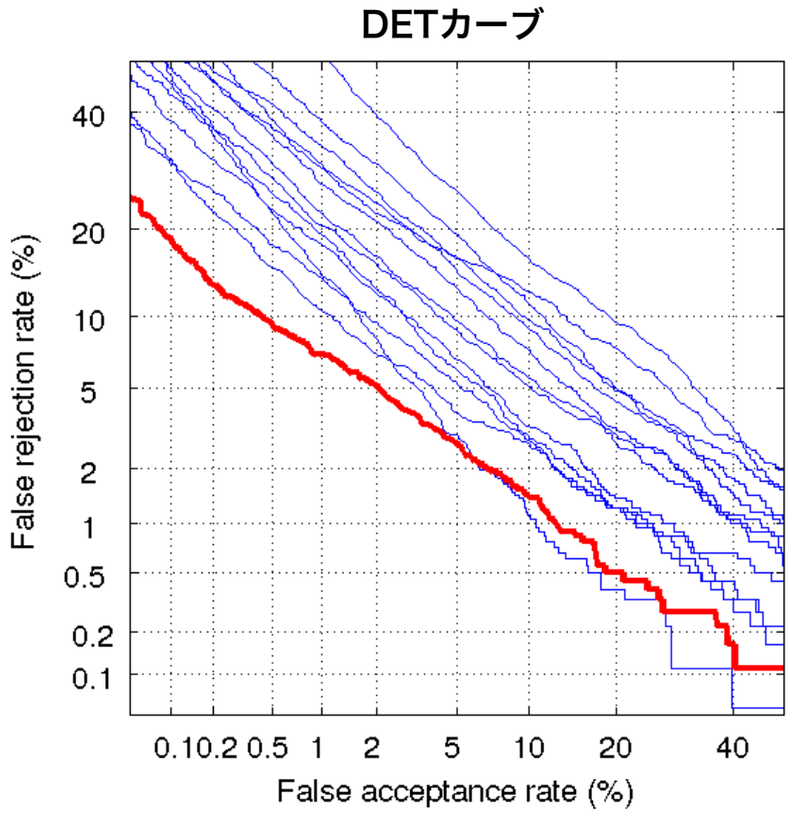
評価指標１：話者類似度の受理・誤受理確率



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

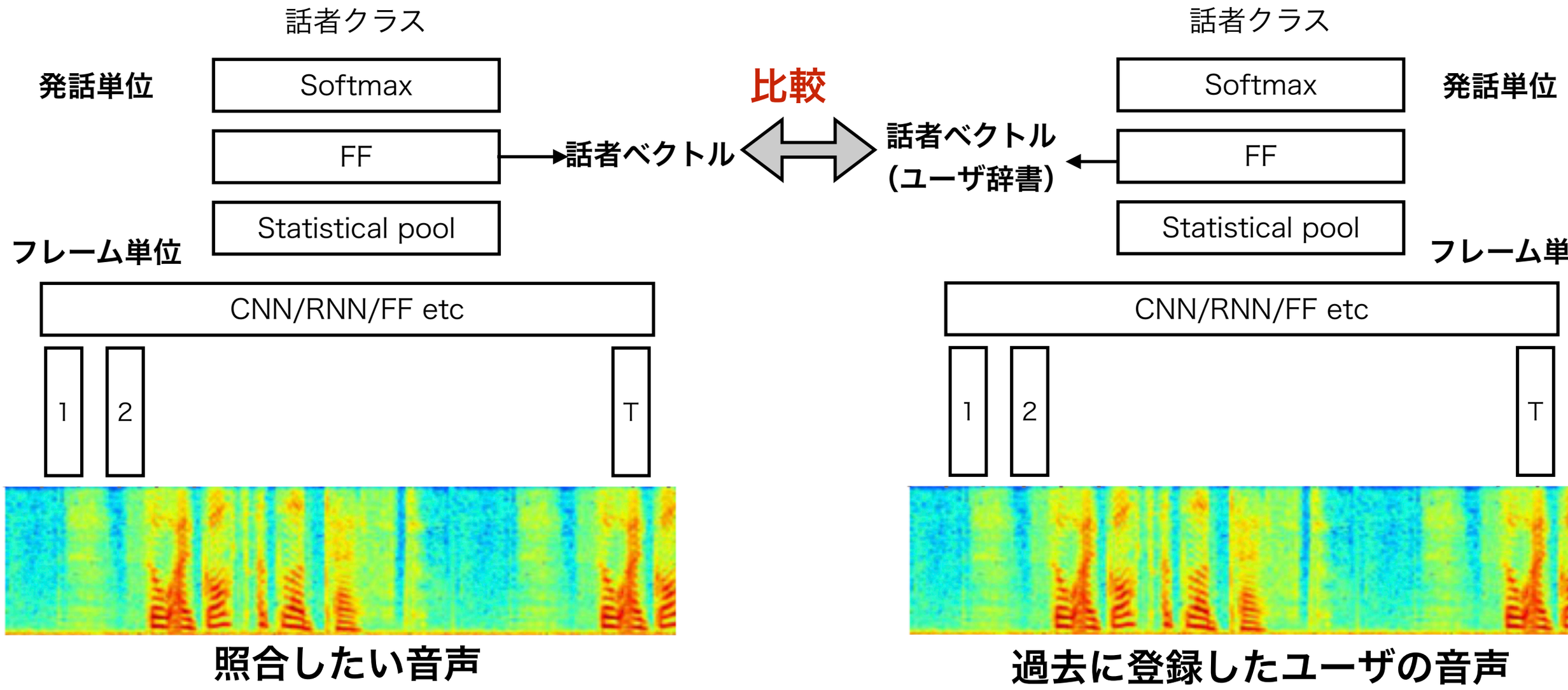
スコアの閾値を調整、
等価誤り率
EER (FAR = FRR)を計算
↓
合成音声が本人の音声に近いほどEERは増加！



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2月

評価指標１：話者ベクトルを利用して本人らしさ計測

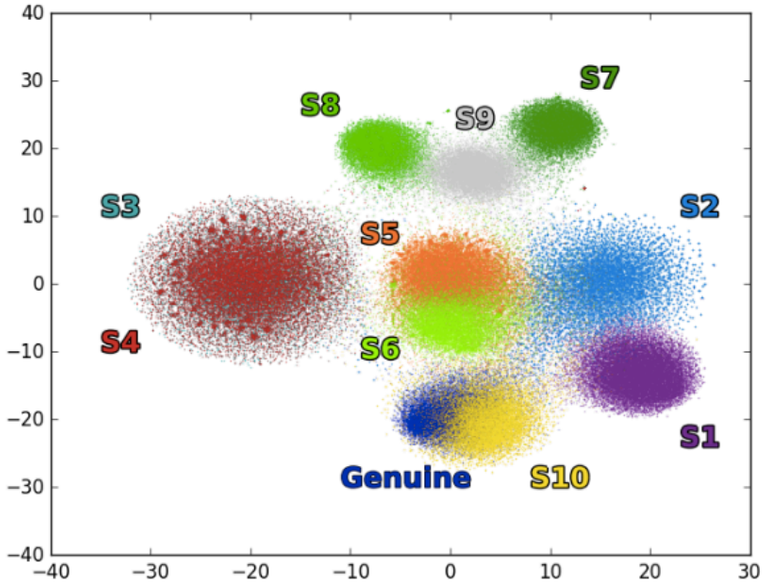
- ニューラルネットワークを利用した話者埋め込みベクトル
- 話者認識の標準技術



評価指標１：本人と合成音声の話者ベクトルを可視化

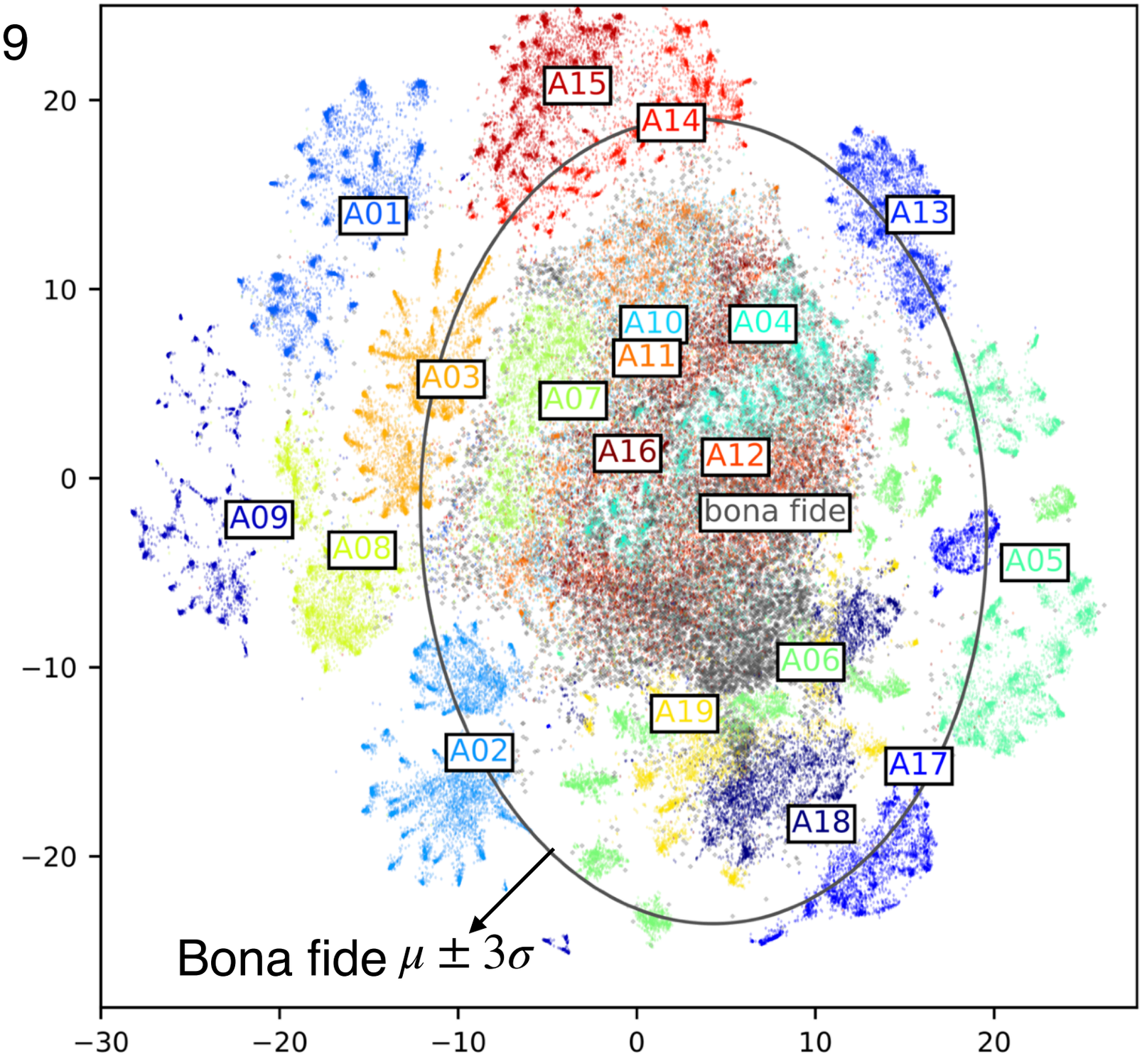
ASVspoof 2019

- 2015年



ASVspoof 2015

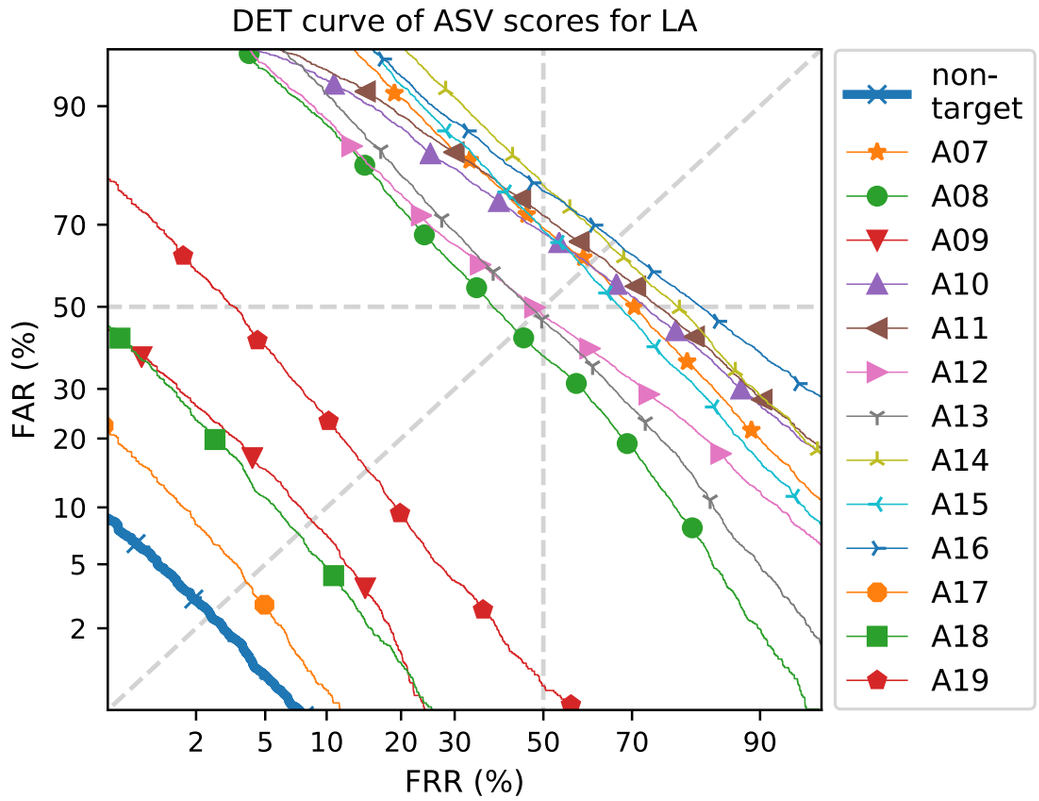
聴覚上も識別困難な
ケースあり



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：合成音声の本人らしさを数値化

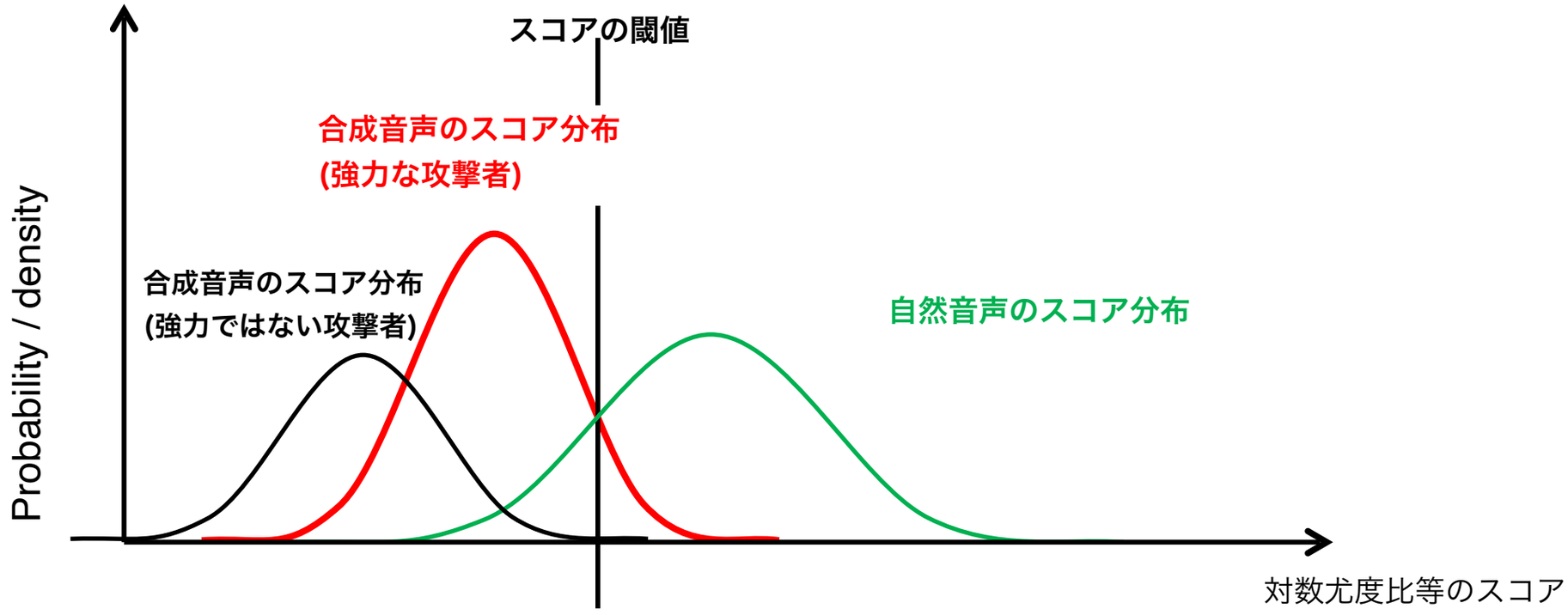
- ASVspoof 2019 LA datasetのサンプルを利用して、声による話者認識システムを実際に攻撃
- 攻撃対象のシステム
  - 数千人の話者を含むVoxCelebデータベースを利用して学習された当時SoTAだった話者照合システム（x-vector / PLDA）



- 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：合成音声のアーティファクト量を計測

- 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量を計測することでディープフェイク検知モデル単独の指標も定義可能
- 等価誤り率も同様に定義可能



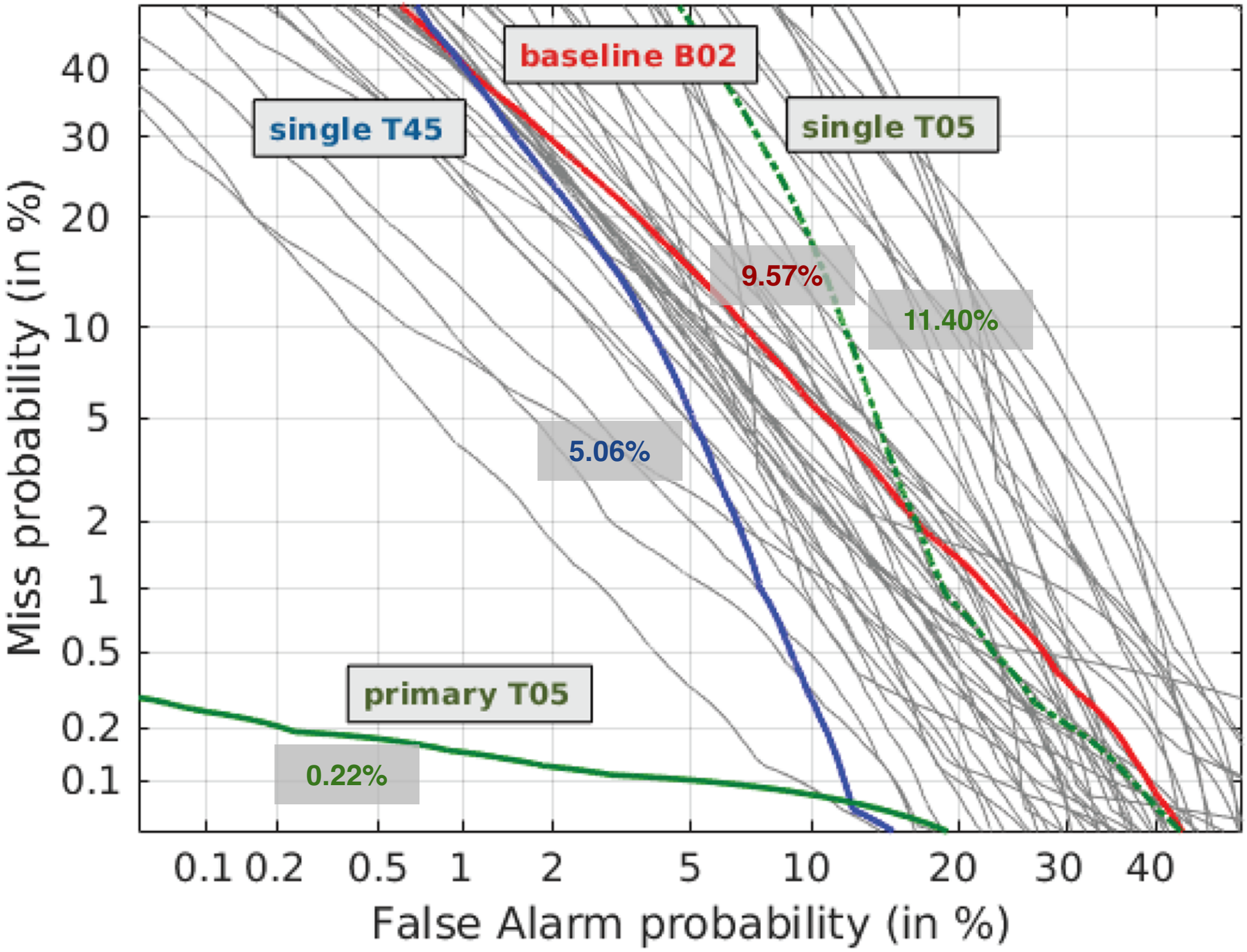
Trial	Decision	
	Accept	Reject
自然音声	Correct accept	**False reject (FRR)**
合成音声	**False alarm (FAR)**	Correct reject

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

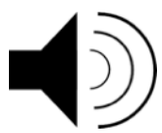
↓

検知モデルの性能が高いとEERが低い

評価指標２：アーティファクト量を計測し、DETプロットで比較



評価指標３ ディープフェイク検知・本人照合の全体エラー



AI attackers
(TTS/VC)

P_a

Human attackers
(Non target)

P_b

True
users

ディープフェイク検知
（アーティファクト）

Accept
(Real)

音声による話者照合
（本人類似度）

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

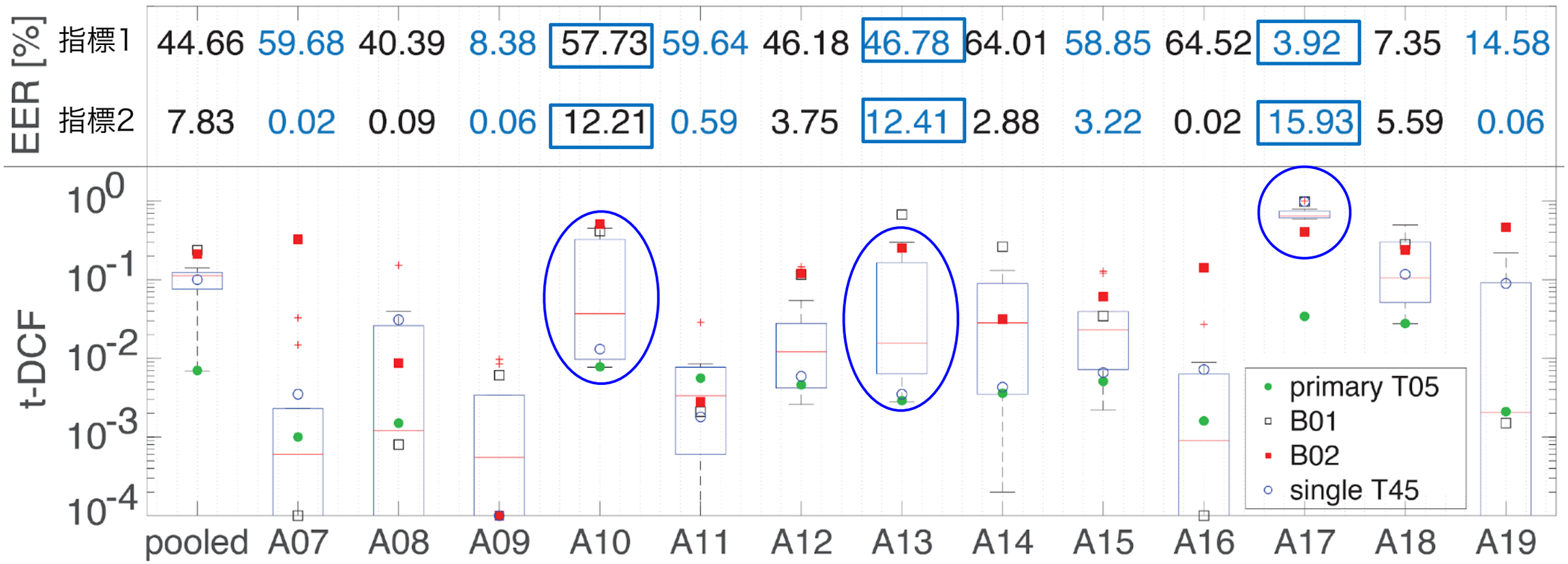
π_i Prior of a specific type of errors

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：ディープフェイク検知・本人照合の全体エラー

- **指標１：話者類似度, EER [%]**
  - 値が高いと話者類似度が高いと判断
- **指標２：アーティファクト量, 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

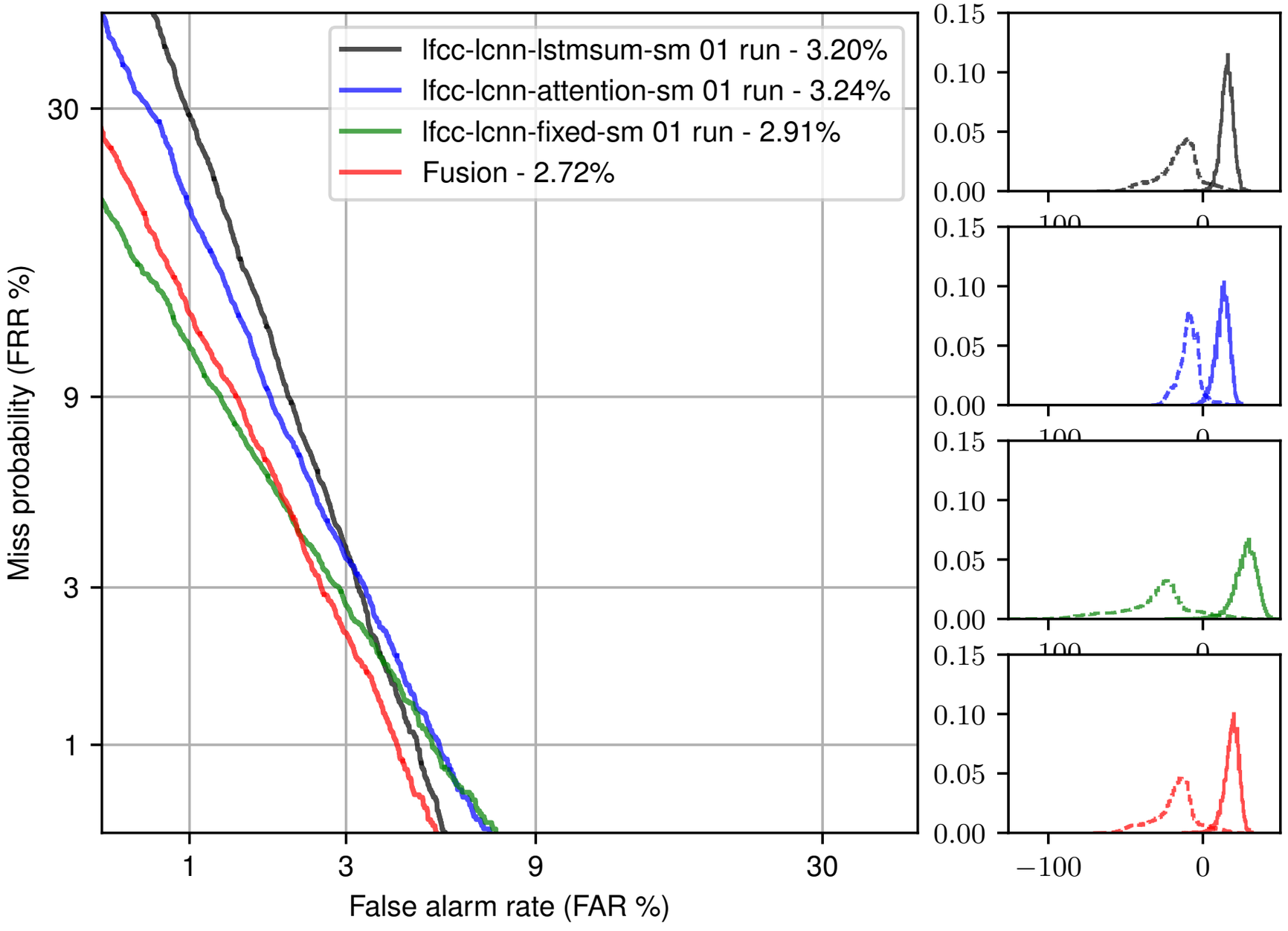
- 50種類以上のディープフェイク検知モデルの中で、t-DCF指標でトップ10だった検知モデルの性能を、ディープフェイク生成手法(A07-A19) 毎にBoxplot表示→どの生成手法が問題が判明！



パート 1-2
音声のディープフェイク検知
ー検知モデルの何が重要？ー

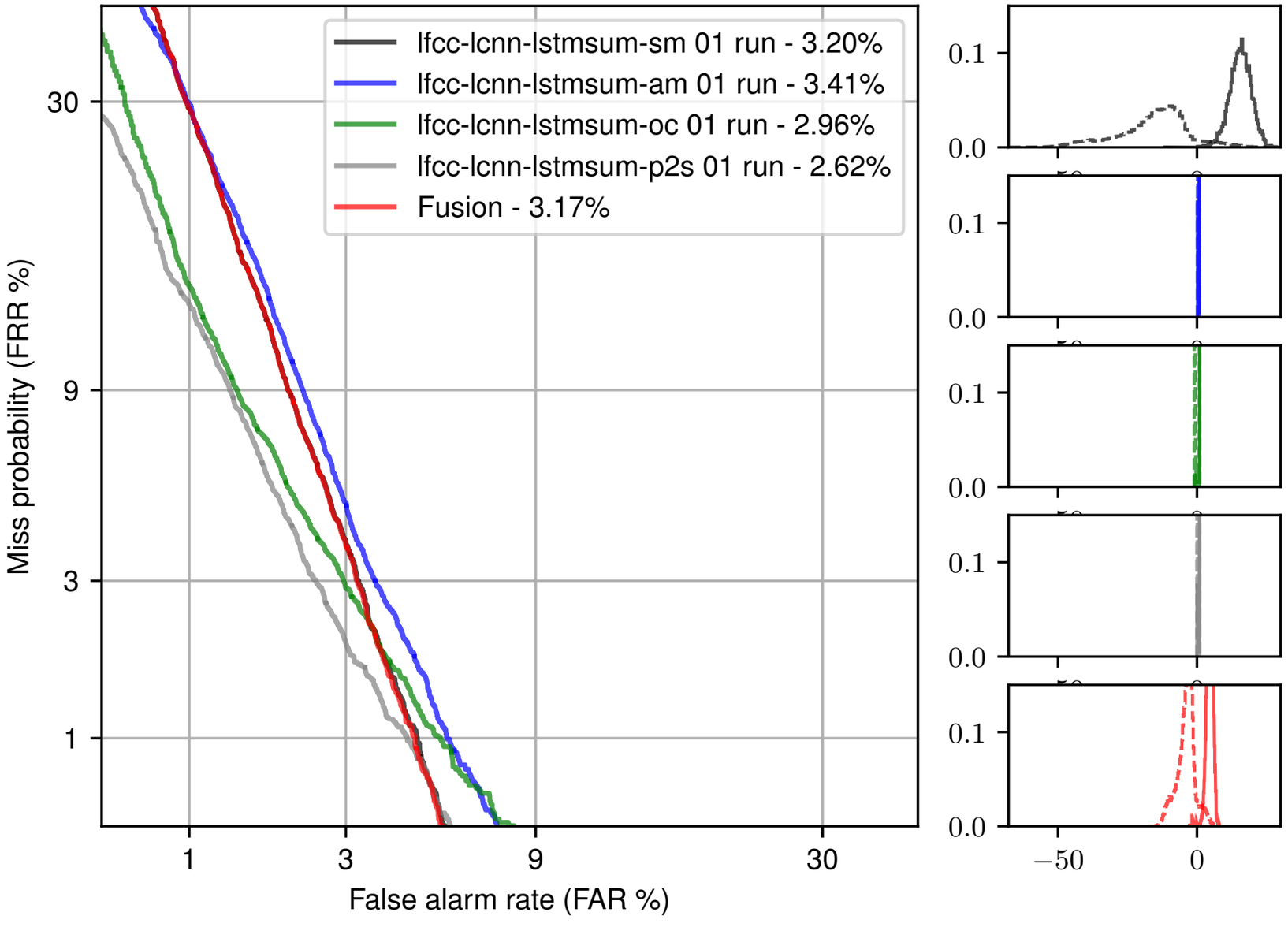
異なるモデル構造

- モデルの構造の一部をLSTM、セルフアテンション、パディングに変更
- 特徴量・学習基準は全て同じ
- これらの異なるモデル構造による検知性能の差は大きくない



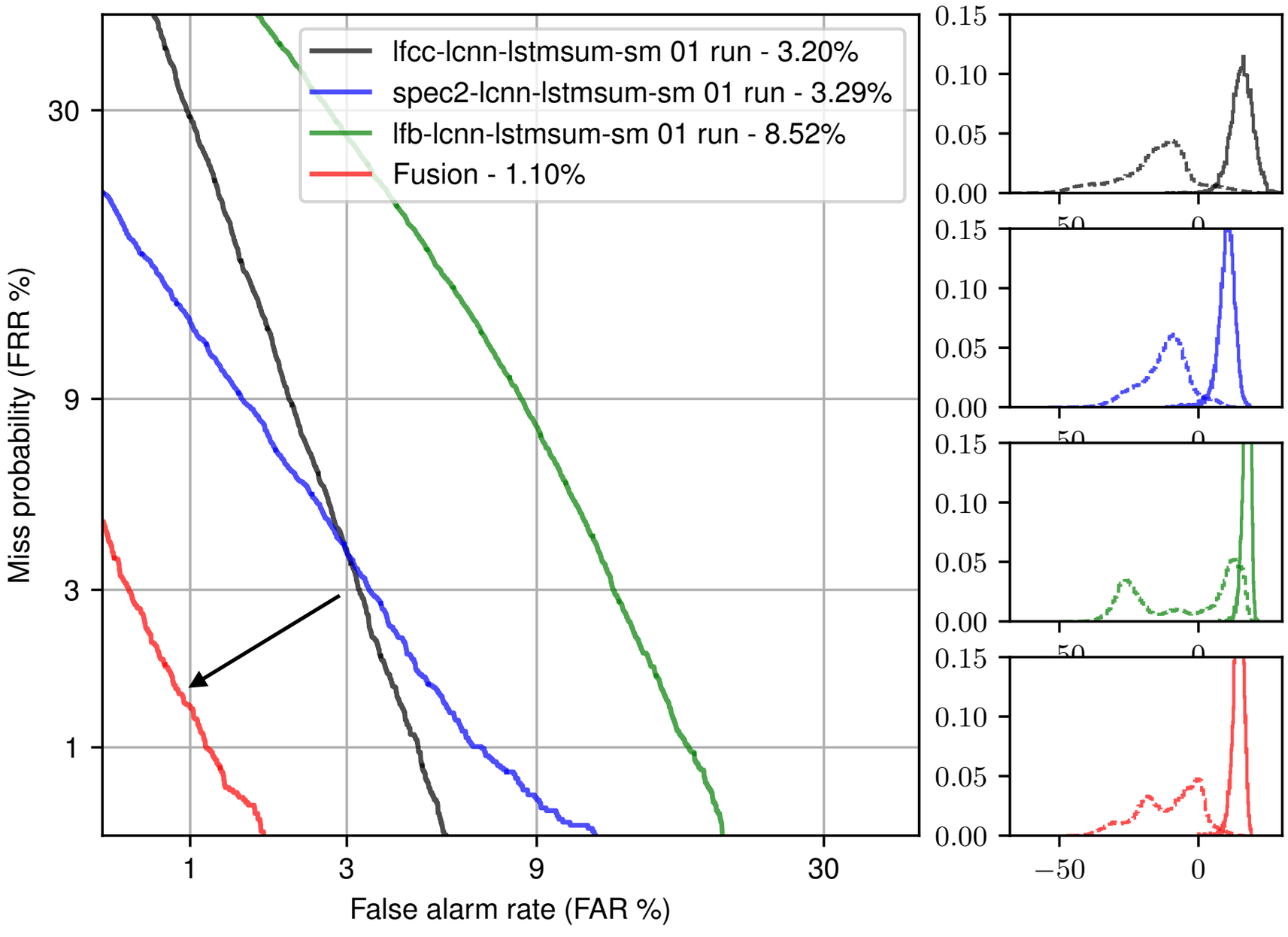
異なる学習基準

- モデルを学習する基準を変更 (softmax, A-softmax, one-class softmax, P2S)
- 特徴量・モデルの構造は一緒
- これらの異なる学習基準による検知性能の差は大きくない



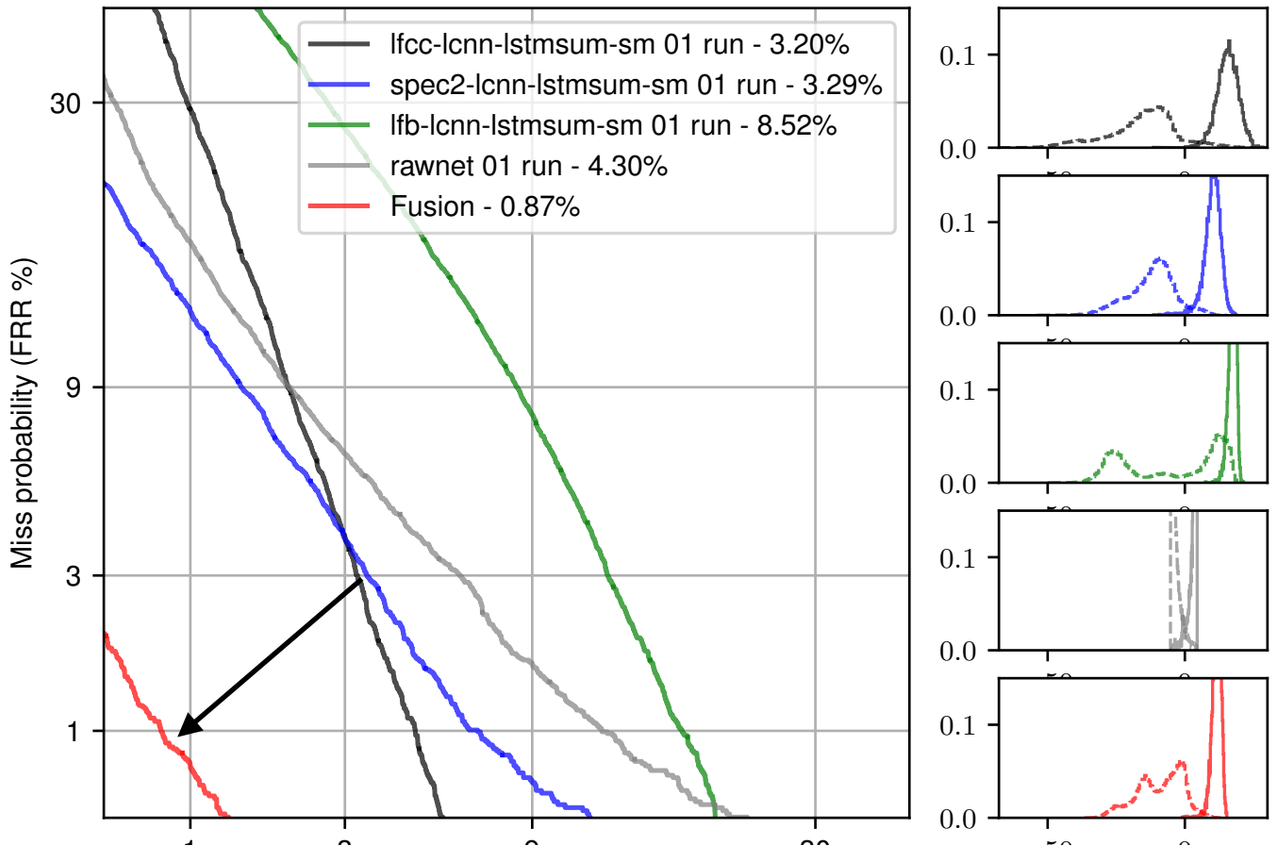
異なる特徴量によるモデル

- 識別モデルの学習基準・構造は一緒
- 周波数特徴量のみを変更 (LFCC, Spectrogram, LFB)
- これらの検知モデルの性能の差は大きい。融合するとさらに改善



特徴量を学習するニューラルネットワークも追加

- 特徴量自身を学習するニューラルネットワークRawNetを追加
- 周波数特徴量のみを変更した複数モデルとRawNetとをフュージョン
- 検知性能がさらに改善



Ref.	Model	EER (%)	min t-DCF		Aug.
			legacy	v2.0	
[84]	GMM-RawNet2 (L+S1)	1.12	0.0330	-	
[62]	LFCC-STFT Capsule	1.07	0.0328	-	
Fig.9b	LCNNs & RawNet	0.87	0.0237	0.0849	
2019 LA top primary	7 sub-models (T05)	0.22	0.0069	0.0692	

Xin Wang, Junichi Yamagishi, “A Practical Guide to Logical Access Voice Presentation Attack Detection”
Frontiers in Fake Media Generation and Detection, Springer, May 2022

本分析の知見

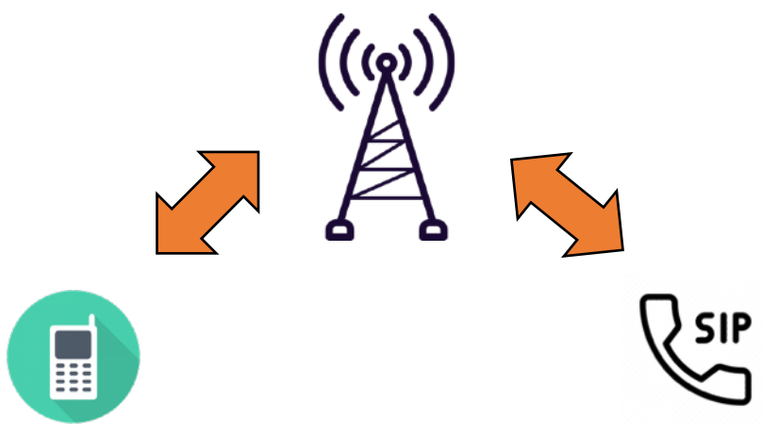
- 多様なディープフェイク生成手法により合成された音声を高精度に識別するためには、様々な特徴量が必要
- そのアンサンブルも有効である

- 評価セットには多種多様な生成手法が含まれており、当然ながら合成音声のアーティファクトも多様である
  - 1つの特徴量空間で識別境界学習するのは難しい
  - 色々な特徴量空間で、ディープフェイク検知に有効な特徴を見つけ、識別を行う事が重要である
  - 特徴量を見つけられれば識別モデルの構造はそこまで重要でない

パート 1-3
音声のディープフェイク検知
ー実環境での評価ー

ディープフェイク検知モデル追加評価セット

- 2021年に**電話越しディープフェイク音声検知を行うシナリオ**(ASVspoof 2021 LA)、動画配信サイトの**圧縮された音声に対してディープフェイク音声検知を行うシナリオ**(ASVspoof 2021 DF)の**評価セットを新たに公開** [16]
  - 電話越しでディープフェイク音声検知を行うシナリオ用データセットは、VoIP通信による転送をアジアとEU間で実際に行いデータ収集
  - 一部のテストデータはさらにIP通信網と公衆交換電話網(PSTN)との変換が伴う通信と設定し、Nuance社の協力のもとディープフェイク音声の転送を行い、データ収集
- 上記追加評価セットには、**100種類以上の未知のディープフェイク生成手法を追加し、その影響を分析**[17]
  - Voice Conversion Challenge 2018の変換音声
  - Voice Conversion Challenge 2020の変換音声
- **狙い：2020年の変換音声は、2019年に構築したディープフェイク検知モデルの学習データベース(ASVspoof 2019)内の音声生成手法より新しい生成手法なので検知が難しいはず**

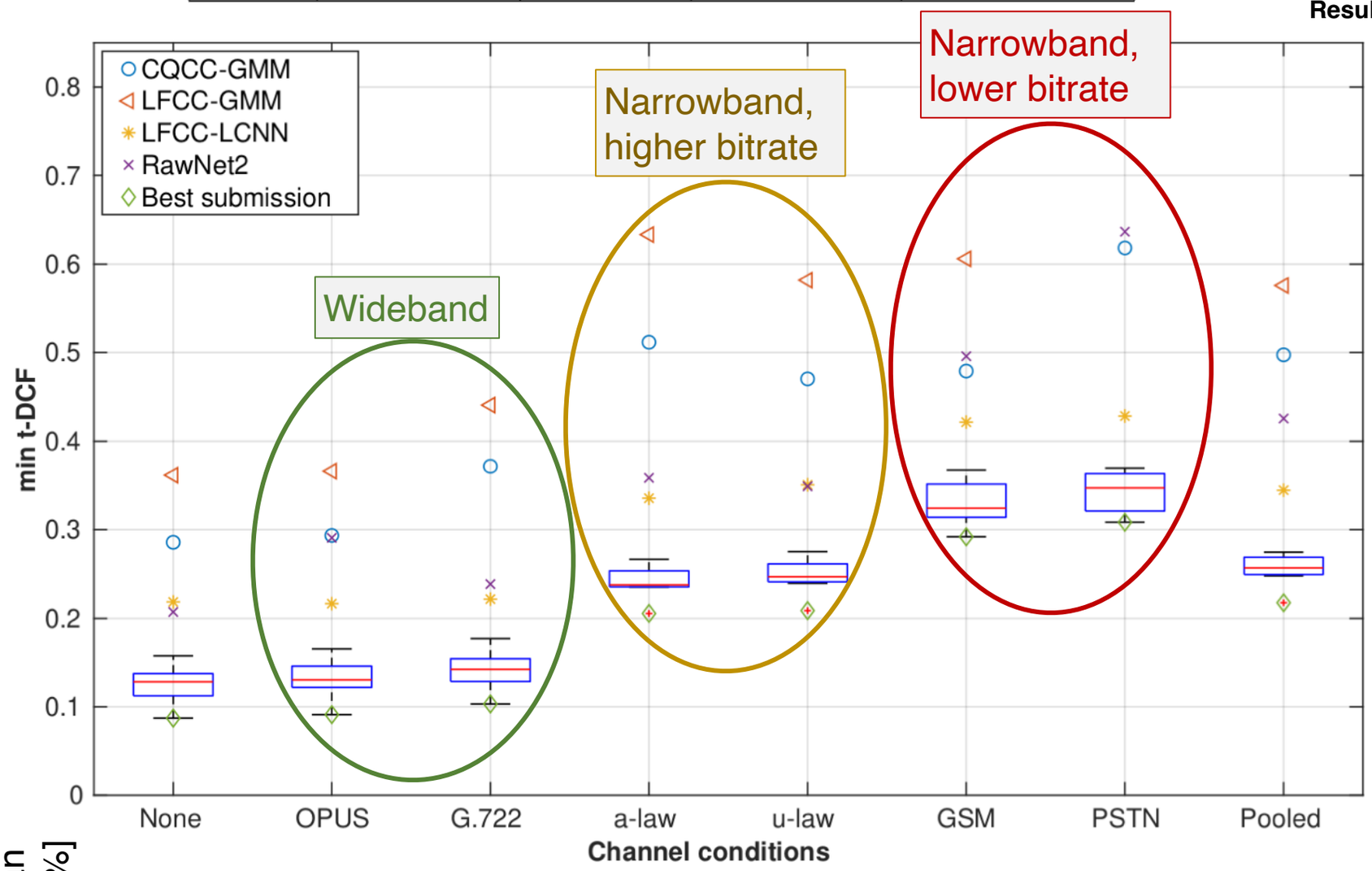


[16] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, Héctor Delgado, “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” The ASVspoof 2021 workshop, Sept. 2021
[17] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, Kong Aik Lee, “Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, June 2023

電話越しでのディープフェイク検知の性能

Cond.	Codec	Audio bandwitdh	Transmission	Bitrate
LA-C1	-	16 kHz	-	250 kbps
LA-C2	a-law	8 kHz	VoIP	64 kbps
LA-C3	unk. + μ-law	8 kHz	PSTN + VoIP	- / 64 kbps
LA-C4	G.722	16 kHz	VoIP	64 kbps
LA-C5	μ-law	8 kHz	VoIP	64 kbps
LA-C6	GSM	8 kHz	VoIP	13 kbps
LA-C7	OPUS	16 kHz	VoIP	VBR ∼16 kbps

Results of top-10 submissions



CM median EER [%]								
	1.87	2.31	2.14	2.33	2.32	3.48	5.59	3.17

圧縮音声に対するディープフェイク検知の性能

Compression conditions:

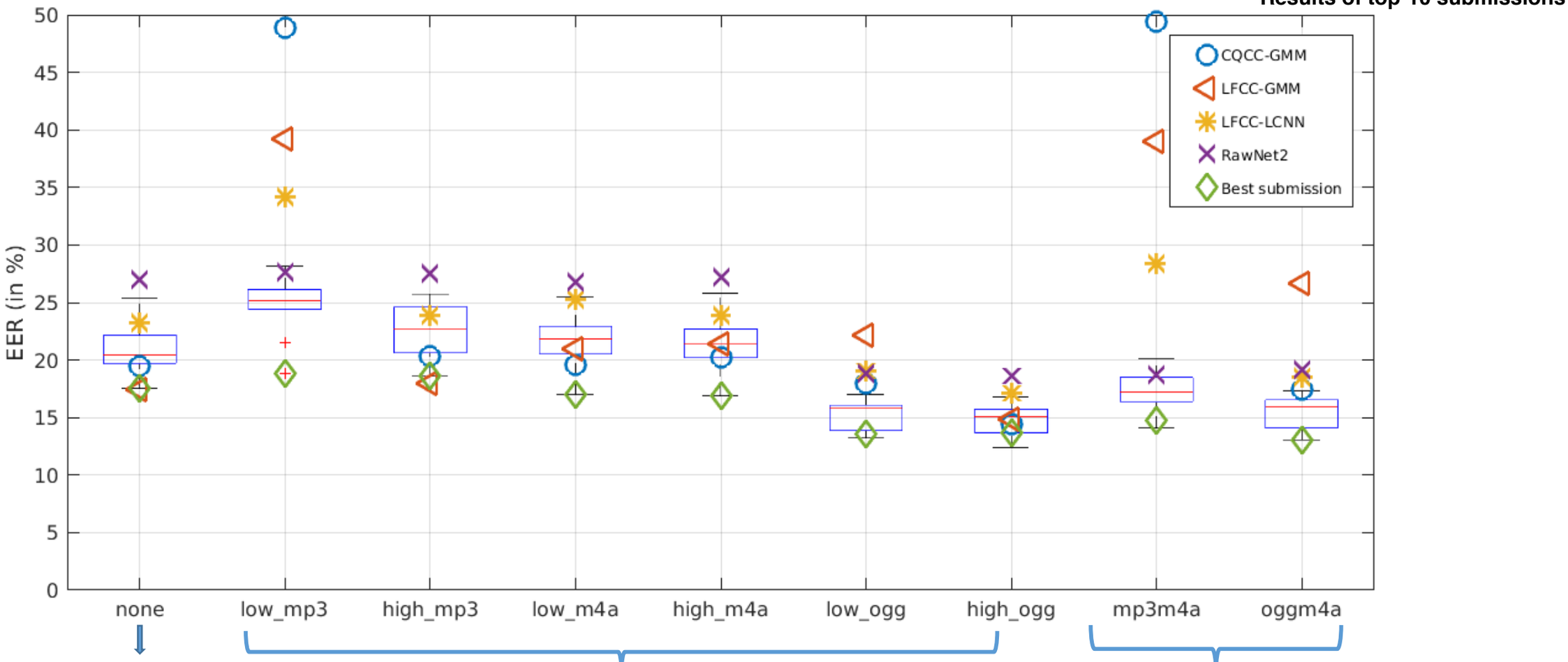
Cond.	Compression (Quality)	VBR settings (kbps)
C1	None	-
C2	mp3 (low)	~80-120
C3	mp3 (high)	~220-260
C4	m4a (low)	~20-32
C5	m4a (high)	~96-112
C6	ogg (low)	~80-96
C7	ogg (high)	~256-320
C8	mp3 (low) => m4a (high)	~80-120=>~96-112
C9	ogg (low) => m4a (high)	~80-96=>~96-112

No compression

Single compression

Double compression

Results of top-10 submissions



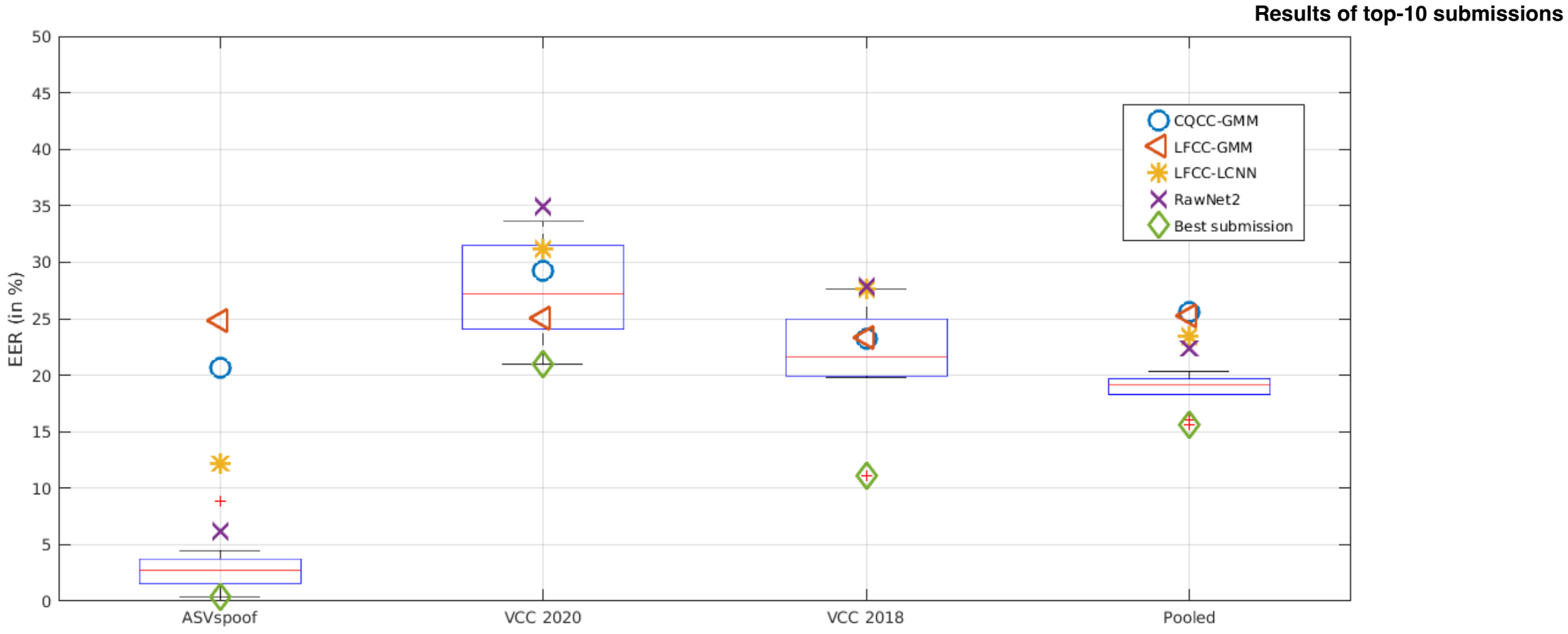
No compression

Single compression

Double compression

新たに追加されたディープフェイク生成手法の検知

- 新たに追加した100種類以上の未知のディープフェイク生成手法の検知
- は予想通り難しい。特に**2020年の方のサンプルの検知が難しかった**

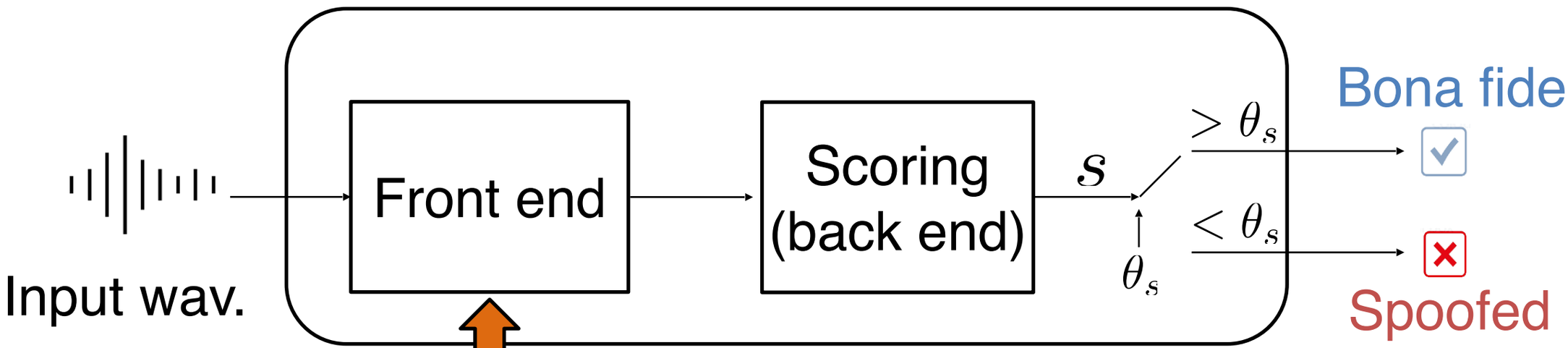


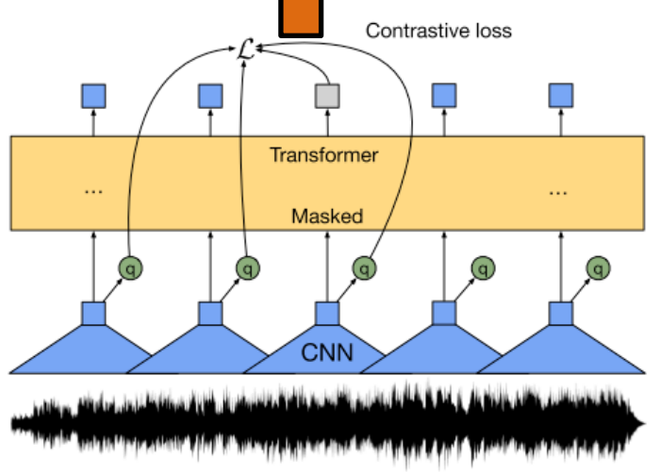
- ディープフェイク検知モデルの特徴量やモデル構造が2019に公開した
  データベース内の合成・変換手法の検出に特化し過ぎている可能性が高い
- より汎化性能の高い特徴量を利用する必要あり

パート 1-4
音声のディープフェイク検知
ー未知の生成手法をどう頑健に検知するか？ー

頑健で汎化性能の高いSSL特徴量

- 周波数系特徴量ではなく、wave2vec 2.0やHuBERT等の大量の自然音声波形により事前学習されたSSLモデル（いわゆる基盤モデル）を特徴量抽出モデルとして導入する





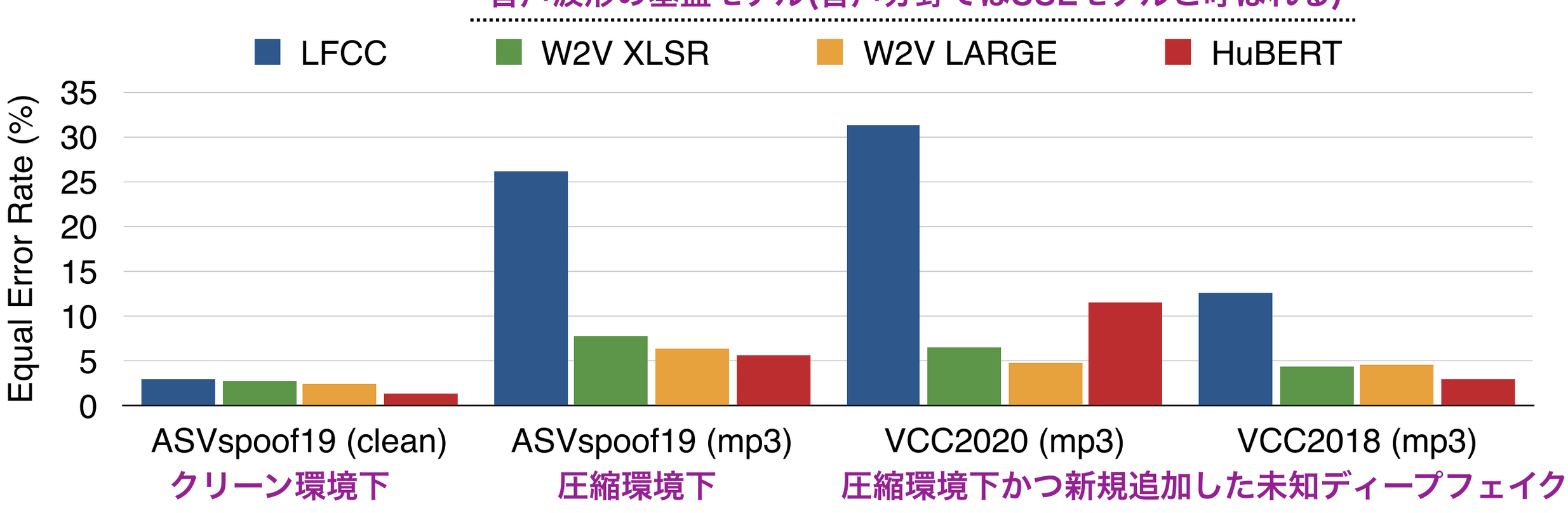
Pre-trained wav2vec 2.0 model (XLSR-53, pre-trained on multi-lingual data, released by Meta)

Model	Architecture	Hours	Languages	Datasets	Model
XLSR-53	Large	56k	53	MLS, CommonVoice, BABEL	download

未知手法によるディープフェイク検出性能向上

- **SSL特徴量を導入すると、新たに追加した100種類以上の未知のディープフェイク手法も劣悪条件下で検出可能になった**
  - 真の理由は不明だが、大量の自然音声により学習されたSSL特徴により自然音声側の境界が明確に定まる？
- RawBoostと命名したデータ拡張手法[21]と上記特徴量の組み合わせにより更に改善可能

音声波形の基盤モデル(音声分野ではSSLモデルと呼ばれる)



[20] Xin Wang, Junichi Yamagishi, “Investigating self-supervised front ends for speech spoofing countermeasures” Speaker and Language Recognition Workshop (Odyssey 2022), June 2022
[21] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, Nicholas Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” Speaker and Language Recognition Workshop (Odyssey 2022), June 2022
[22] Xin Wang, Junichi Yamagishi, “Investigating Active-learning-based Training Data Selection for Speech Spoofing Countermeasure,” IEEE Spoken Language Technology Workshop (SLT 2022), June 2023

最近の取り組み：更なる巨大基盤モデルの利用

音声の代表的な基盤モデル

ID	Model type	Training data	#.para	Out dim.
W2V-Small	Wav2vec 2.0	Librispeech	95 M	768
W2V-Large	Wav2vec 2.0	CommonVoice, Switch-board	317 M	1024
MMS-300M	Wav2vec 2.0		317 M	1024
MMS-1B	Wav2vec 2.0		965 M	1280
XLS-R-1B	Wav2vec 2.0		965 M	1280
XLS-R-2B	Wav2vec 2.0		2.2 B	1920

- 95Mから2.2Bパラメータまでの基盤モデル利用
- 18,000時間以上の合成音声と56,000時間以上の人間の自然音声からなる大規模データセットを新たに用意
- 日本語を含む100言語以上を含む
- 巨大基盤モデルを大規模データでディープフェイク検知タスクで追加事前学習
  - https://github.com/nii-yamagishilab/AntiDeepfake
  - https://huggingface.co/nii-yamagishilab

AIP加速課題用の学習セット

Database	Language	Num. of Hours	Attack
AISHELL3 [2]	Mandarin	real: 85.62 h fake: 0 h	-
ASVspoof2019-LA [3]	English	real: 11.85 h fake: 97.80 h	TTS, VC
ASVspoof2021-LA [4]	English	real: 16.40 h fake: 116.10 h	TTS, VC
ASVspoof2021-DF [4]	English	real: 20.73 h fake: 487.00 h	TTS, VC
ASVspoof5 [5]	English	real: 413.49 h fake: 1808.48 h	TTS, VC
CFAD [6]	Mandarin	real: 171.25 h fake: 224.55 h	Vocoded
CNCeleb2 [7]	Mandarin	real: 1084.34 h fake: 0 h	-
Codecfake [8]	English, Mandarin	real: 129.66 h fake: 808.32 h	Neural Codec
CodecFake [9]	English	real: 0 h fake: 660.92 h	Neural Codec
CVoiceFake [10]	5 Languages	real: 315.14 h fake: 1561.16 h	Vocoded
DECRO [11]	English, Mandarin	real: 35.18 h fake: 102.44 h	TTS, VC
DFADD [12]	English	real: 41.62 h fake: 66.01 h	TTS
Diffuse or Confuse [13]	English	real: 0 h fake: 231.66 h	TTS
DiffSSD [14]	English	real: 0 h fake: 139.73 h	TTS
DSD [15]	English, Japanese, Korean	real: 100.98 h fake: 60.23 h	TTS, VC
FLEURS [16]	102 Languages	real: 1388.97 h fake: 0 h	-
FLEURS-R [17]	102 Languages	real: 0 h fake: 1238.83 h	Restored
HABLA [18]	Latin American Spanish	real: 35.56 h fake: 87.83 h	TTS, TTS-VC
LibriTTS [19]	English	real: 585.83 h fake: 0 h	-
LibriTTS-R [20]	English	real: 0 h fake: 583.15 h	Restored
LibriTTS-Vocoded	English	real: 0 h fake: 2345.14 h	Vocoded
LJSpeech [21]	English	real: 23.92 h fake: 0 h	-
MLADD [22]	38 Languages	real: 0 h fake: 377.96 h	TTS
MLS [23]	8 Languages	real: 50558.11 h fake: 0 h	-
SpoofCeleb [24]	Multilingual	real: 173.00 h fake: 1916.2 h	TTS
VoiceMOS [25]	English	real: 0 h fake: 448.44 h	TTS
VoxCeleb2 [26]	Multilingual	real: 1179.62 h fake: 0 h	-
VoxCeleb2-Vocoded	Multilingual	real: 0 h fake: 4721.46 h	Vocoded
WaveFake [27]	English, Japanese	real: 0 h fake: 198.65 h	TTS
Total	Over 100 Languages	real: 56.37 kh fake: 18.28 kh	-

巨大基盤モデルによるゼロショット評価結果

- 学習に利用していない音声生成手法を含む複数のデータセット上でゼロショットによる評価（ドメイン適用やチューニングは一切なし）

Model ID	# of Params.	DA	Evaluation Set				
			ADD2023	DEEP-VOICE	FakeOrReal		In-the-Wild
			Track-1.2-R2-Test	Segmented Full Set	original-Test	norm-Test	Full Set
W2V-Small	95 M	×	19.35	16.2	1.10	6.56	4.70
		✓	13.15	9.87	22.11	17.89	4.27
W2V-Large	317 M	×	12.58	4.93	0.80	1.44	2.23
		✓	13.23	4.51	0.63	**0.97**	1.91
MMS-300M	317 M	×	11.31	2.92	**0.51**	2.71	2.02
		✓	7.93	2.42	1.60	5.88	2.94
MMS-1B	965 M	×	9.57	2.35	0.97	1.06	1.87
		✓	8.96	2.46	1.39	1.73	1.84
XLS-R-1B	965 M	×	6.62	**2.18**	3.16	10.87	1.35
		✓	5.45	2.46	5.23	9.81	1.33
XLS-R-2B	2.2 B	×	6.90	2.71	1.39	1.78	1.33
		✓	**4.64**	2.32	2.66	1.73	**1.24**
HuBERT-XL	964 M	×	35.11	15.01	3.84	15.36	18.11
		✓	18.9	5.69	2.49	3.17	5.23
Best reported by others	-	-	13.05	-	4.88	3.93	1.99

- 大規模モデルは種々の未知生成手法を**安定的に**検知可能

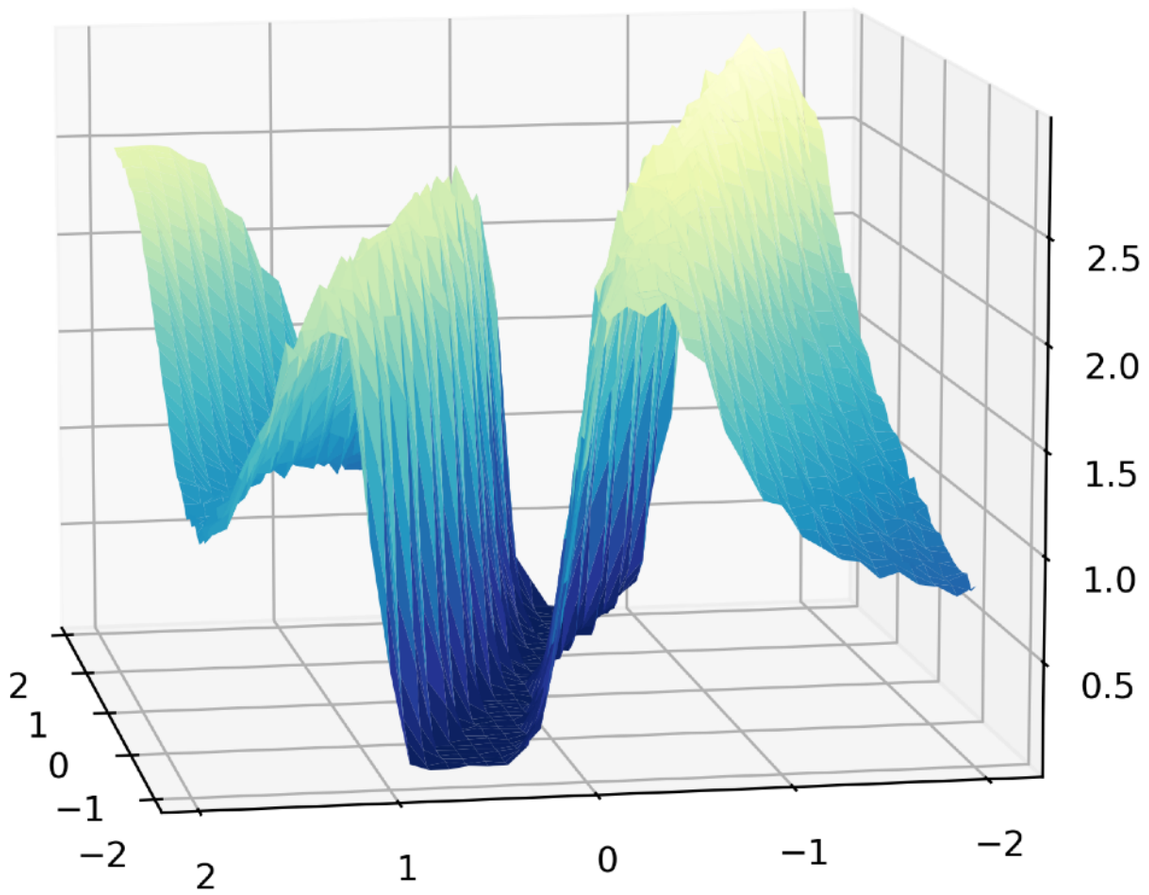
力技ではなくもっとスマートに汎化性能を向上できないか？

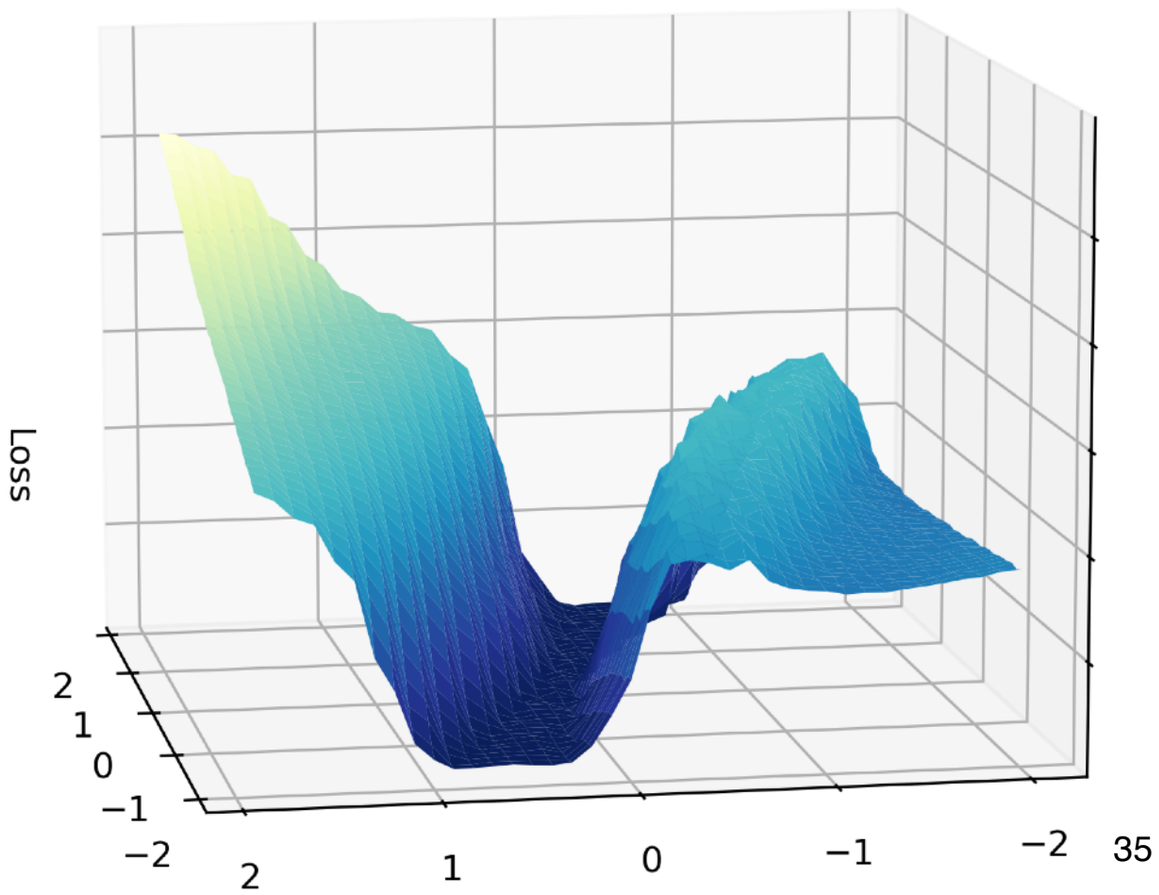
- 未知の生成手法によるテストセットにはドメインシフトが起き、学習データから最適化で得られたモデルパラメータが最適と限らない
- Sharpness-aware minimization (SAM) を導入

$$L_S(w) + \lambda \|w\|_2^2 + [\max_{\|\epsilon\|_2 \leq \rho} L_S(w+\epsilon) - L_S(w)]$$

現在のパラメータw近傍におけるSharpness

- データシフトの影響を受けにくい様「平たい最適点」を探す





SAMによる実験結果

Table 1: *Performance comparison of different models trained with either Adam or SAM across multiple test sets. The table reports the mean ± standard deviation EER (%) over three runs.* ***Bold*** *values indicate the lower value for each model.*

学習に利用されたデータベースとは大きく異なるデータセット（生成手法・言語・収録環境）の場合

Model	Optimizer	19LA	21LA	21DF	ITW	FOR	WF	ADD	SC
AASIST	Adam	1.96 ± 0.61	8.10 ± 1.33	21.92 ± 2.61	34.31 ± 6.73	45.85 ± **0.68**	38.68 ± 3.02	40.48 ± 3.98	45.27 ± **3.01**
	SAM	**1.71 ± 0.55**	**4.62 ± 0.81**	**19.58 ± 1.18**	**33.34 ± 3.07**	**33.51** ± 12.37	**33.85 ± 2.15**	**33.84 ± 2.80**	**43.98** ± 4.71
W2V-Base+Linear	Adam	1.78 ± 0.51	5.82 ± 1.01	13.10 ± 3.00	16.85 ± 2.74	12.01 ± **2.07**	**30.34 ± 6.71**	30.47 ± 9.03	34.98 ± **1.66**
	SAM	**1.21 ± 0.34**	**3.39 ± 0.89**	**11.93 ± 0.55**	**13.66 ± 1.82**	**9.57** ± 2.78	36.95 ± 7.40	**24.29 ± 3.57**	**34.87** ± 2.97
W2V-Base+AASIST	Adam	2.81 ± 0.79	4.78 ± 0.57	10.37 ± 1.19	18.29 ± **0.47**	11.11 ± 1.73	**27.79** ± 3.29	36.22 ± 5.28	45.89 ± 3.56
	SAM	**1.32 ± 0.38**	**3.12 ± 0.39**	**10.00 ± 0.76**	**16.21** ± 2.02	**7.92 ± 1.37**	34.29 ± **2.88**	**28.48 ± 3.73**	**34.83 ± 0.46**
W2V-Large+Linear	Adam	1.37 ± 0.40	3.11 ± 0.64	6.79 ± 0.38	14.96 ± 1.41	**13.24** ± 2.44	23.97 ± 7.30	30.97 ± 9.91	48.35 ± **1.40**
	SAM	**0.88 ± 0.10**	**2.58 ± 0.37**	**6.37 ± 0.30**	**12.22 ± 1.41**	14.65 ± **1.79**	**16.69 ± 1.59**	**37.31 ± 6.68**	**42.89** ± 5.73
W2V-Large+AASIST	Adam	1.22 ± 0.60	4.53 ± 0.99	7.17 ± **0.17**	16.36 ± **1.80**	11.58 ± 1.81	27.89 ± 2.76	33.60 ± 3.40	**39.62 ± 4.50**
	SAM	**1.04 ± 0.47**	**3.60 ± 0.16**	**6.82** ± 0.32	**15.46** ± 2.09	**10.02 ± 0.82**	**25.12 ± 1.47**	**31.41 ± 3.28**	39.67 ± 5.14
W2V-XLSR+Linear	Adam	0.34 ± 0.06	**1.32 ± 0.35**	4.27 ± **0.43**	6.00 ± **0.51**	4.55 ± 1.19	9.87 ± 3.24	22.85 ± 2.78	25.82 ± **1.89**
	SAM	**0.20 ± 0.05**	1.87 ± 0.39	**3.38** ± 0.47	**5.99** ± 0.80	**3.69 ± 0.90**	**7.66 ± 1.24**	**21.71 ± 2.08**	**25.65** ± 2.74
W2V-XLSR+AASIST	Adam	0.34 ± 0.13	1.85 ± 0.25	3.61 ± **0.32**	6.89 ± 1.19	**4.56 ± 0.72**	16.92 ± 7.36	**19.67** ± 1.67	**27.50 ± 1.98**
	SAM	**0.25 ± 0.12**	**1.71 ± 0.27**	**3.44** ± 0.54	**6.34 ± 0.62**	5.18 ± 1.48	**14.36 ± 4.74**	21.36 ± **0.59**	29.93 ± 3.30

- 学習時と条件が異なる際のディープフェイク検出の困難さをSharpnessがある程度説明可能（相関係数=0.4~0.9）

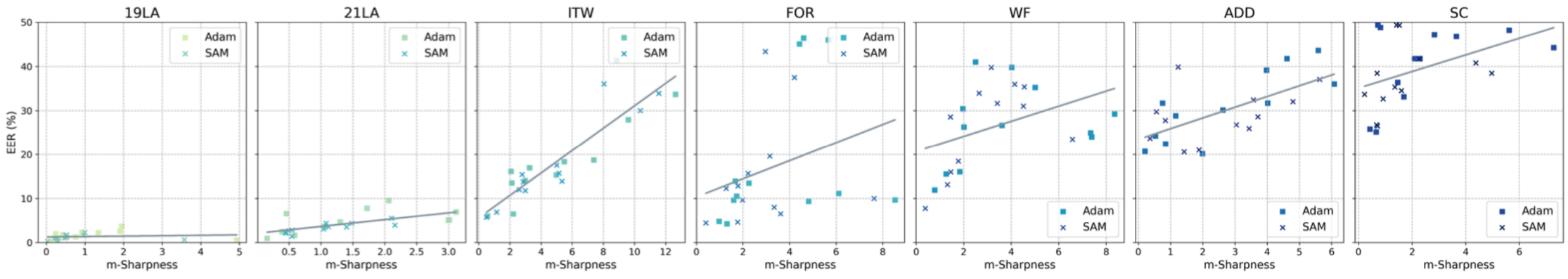


Figure 3: *Scatter plots of sharpness (x-axis) and EER (%) (y-axis) across seven datasets. Each subplot corresponds to one dataset, with square markers for Adam systems and cross markers for SAM systems. A linear regression trend line is included for each group to represent the relationship between the two variables.*

本分析の知見

- 未知・派生生成手法により合成された音声を高精度に識別するためには、やはり特徴量が大事

- 多種多様な生成手法による合成音声を予め全て収集する事は不可能
- しかしその一方、大量の自然音声を事前に収集する事は可能

- 大量の自然音声で学習されたSSLモデルにより特徴量はディープフェイク検知に非常に有用

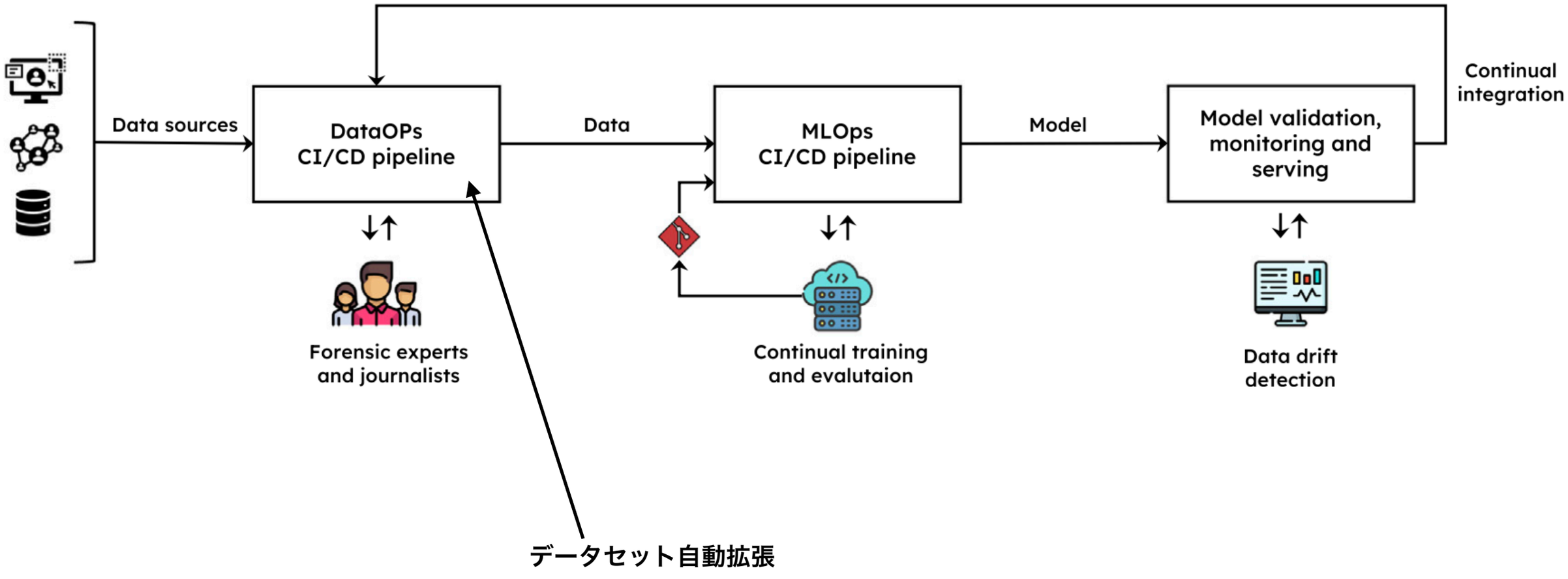
- 巨大SSLモデルおよびその追加学習は非常に有効

- SAMで更に少し改善できる

パート 1-5
音声のディープフェイク検知
ー安定運用継続のための学習データ自動選択ー

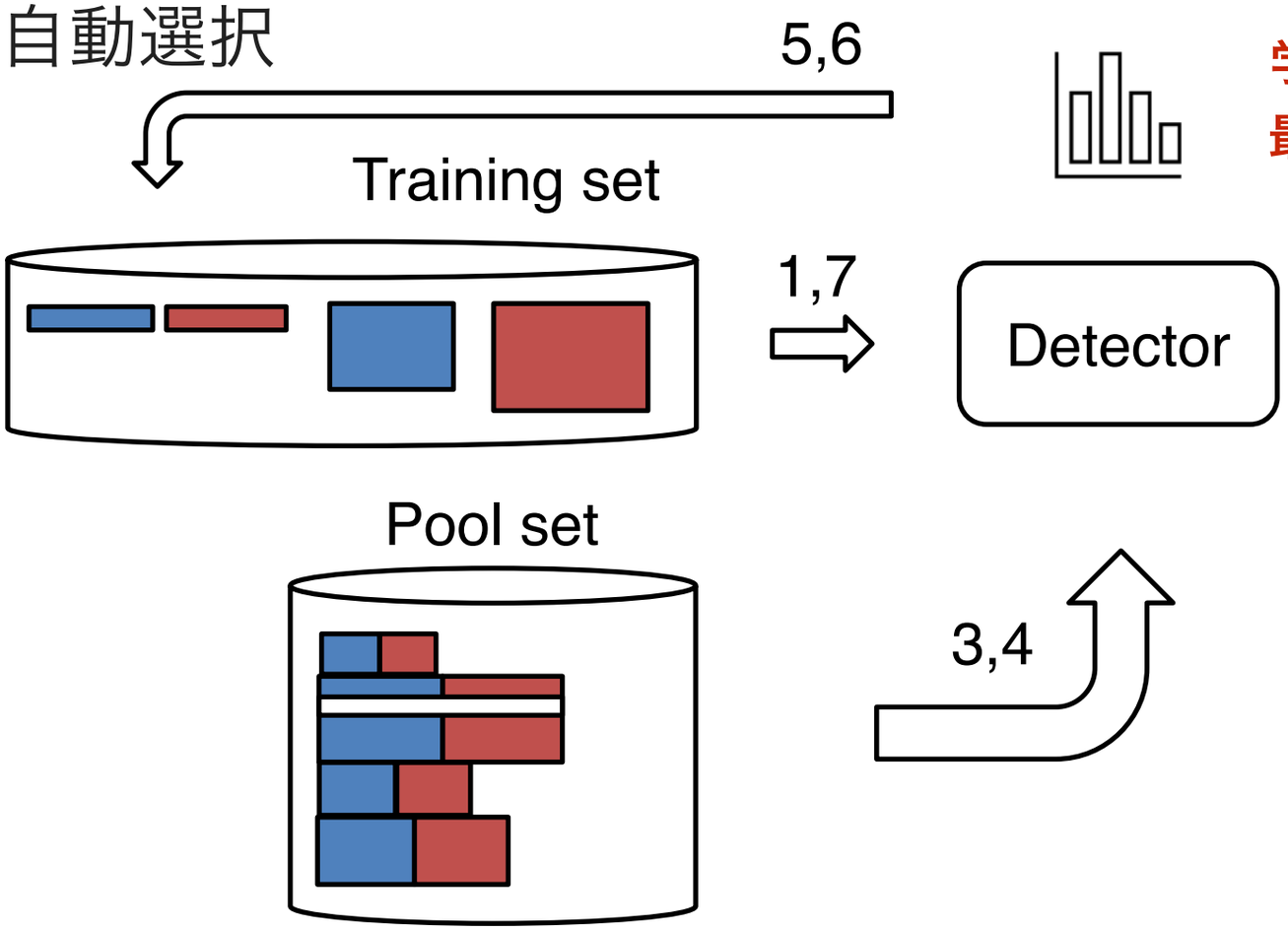
ディープフェイク検知の実運用にはモデルの定期更新が必須

- いかにディープフェイク検知モデルを頑健にしても、**ある段階でデータセット再構築とモデルの再学習によるパラメータ更新が絶対必要**になる
- **どの様に追加データを選び、どの様にモデルを継続学習するか？**
  - 「DataOps」や「MLOps」とも呼ばれる



学習用データベース自動拡張用のアクティブデータ選択

- モデルの定期更新の際には、**以前のディープフェイク手法に加えて、新たに出現したディープフェイク手法を検知できる様にモデル重みを対応させる事が最重要課題（過去の手法を忘れてはならない）**
- 検知モデルの**信頼度**に基づきデータを学習データに追加すべきデータを自動選択



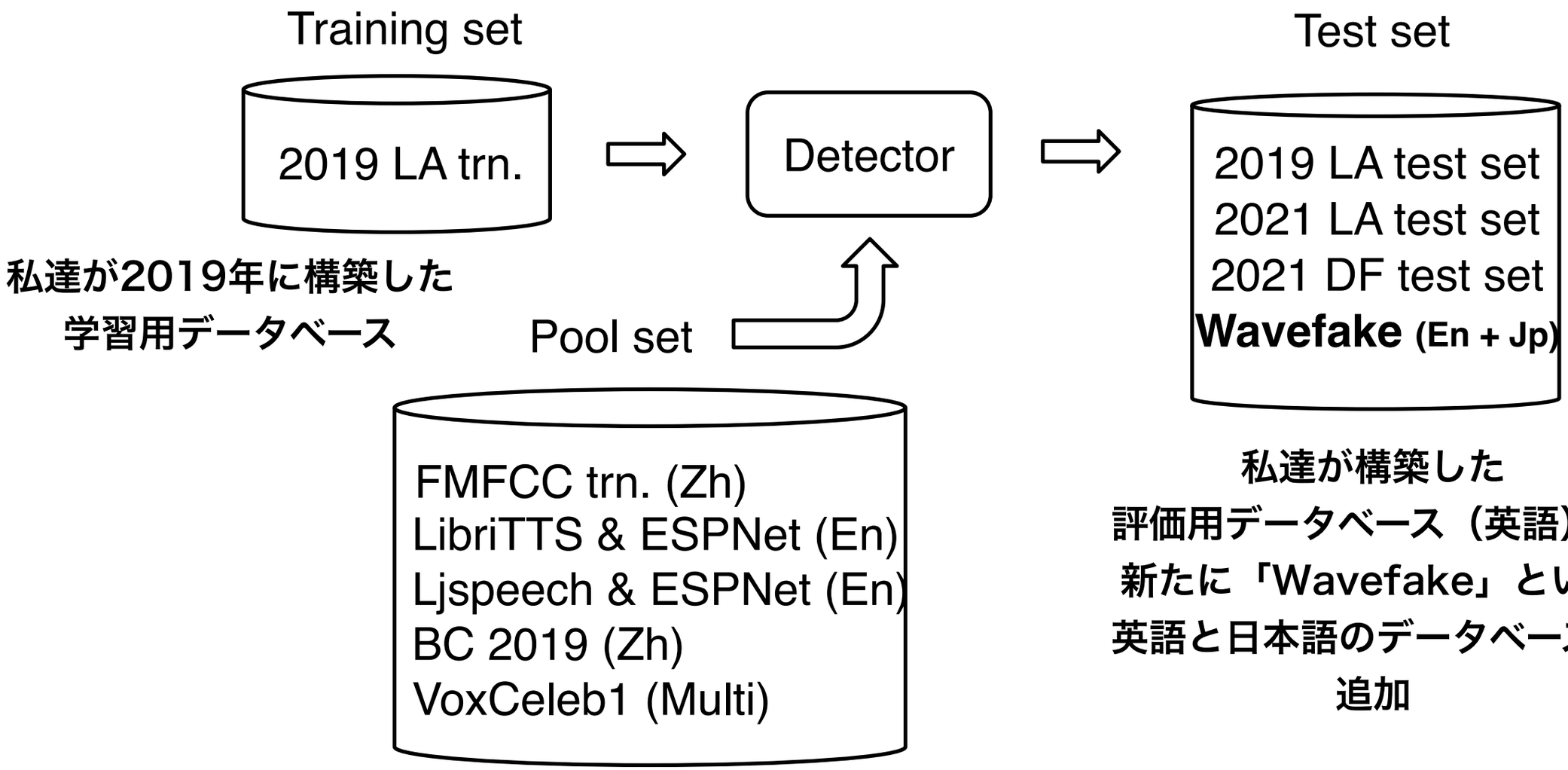
学習データに追加すべき
最重要データを自動選択

追加データ候補：
未知の生成モデルによるデータもあるが冗長

Pseudo code

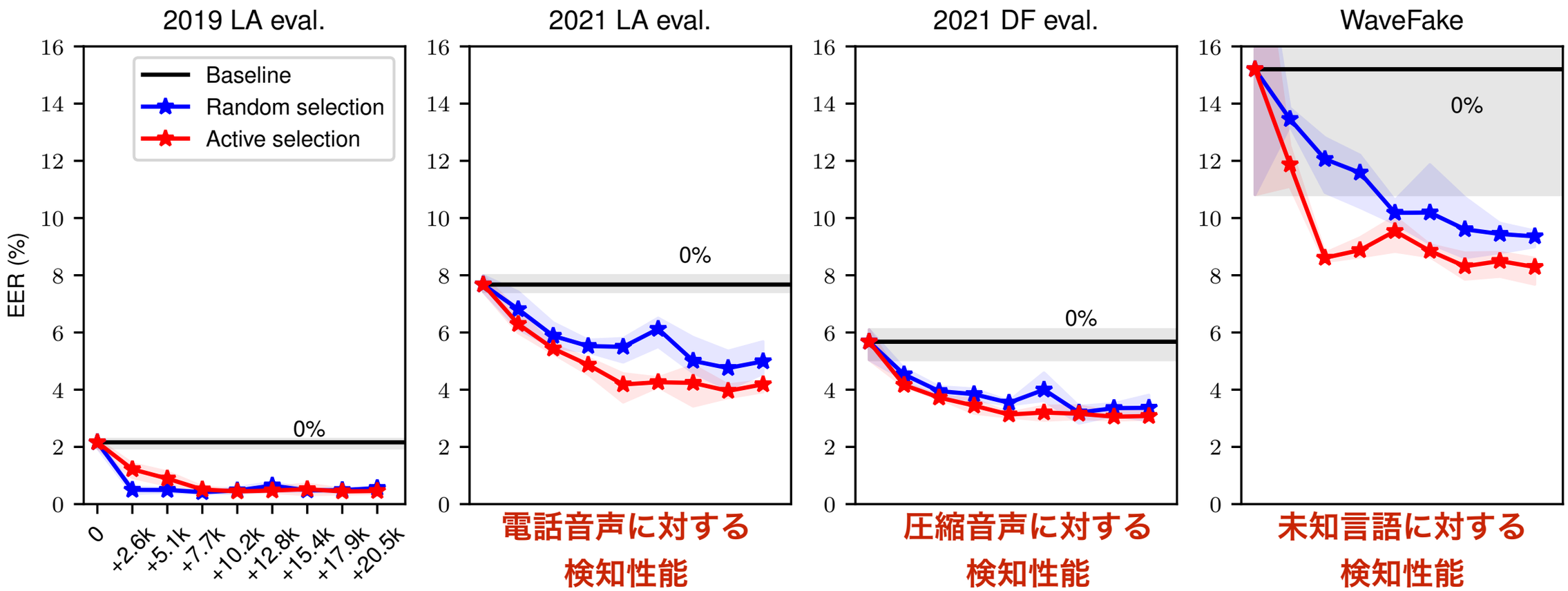
```
1 Train Detector using seed trn. set
2 While loop not ends
3   For trial in pool:
4       usefulness = Detector(trial)
5   Select most_useful_trials
6   Add to training set
7   Fine tune Detector
```

実験条件


Training set
2019 LA trn.
私達が2019年に構築した
学習用データベース
Pool set
FMFCC trn. (Zh)
LibriTTS & ESPNet (En)
Ljspeech & ESPNet (En)
BC 2019 (Zh)
VoxCeleb1 (Multi)
追加収集されたデータと想定し
徐々に追加
Detector
Test set
2019 LA test set
2021 LA test set
2021 DF test set
Wavefake (En + Jp)
私達が構築した
評価用データベース（英語）に
新たに「Wavefake」という
英語と日本語のデータベースを
追加

学習データセットの自動拡張結果

- SSL特徴を利用したLCNNモデルをデータベース拡張する毎に更新

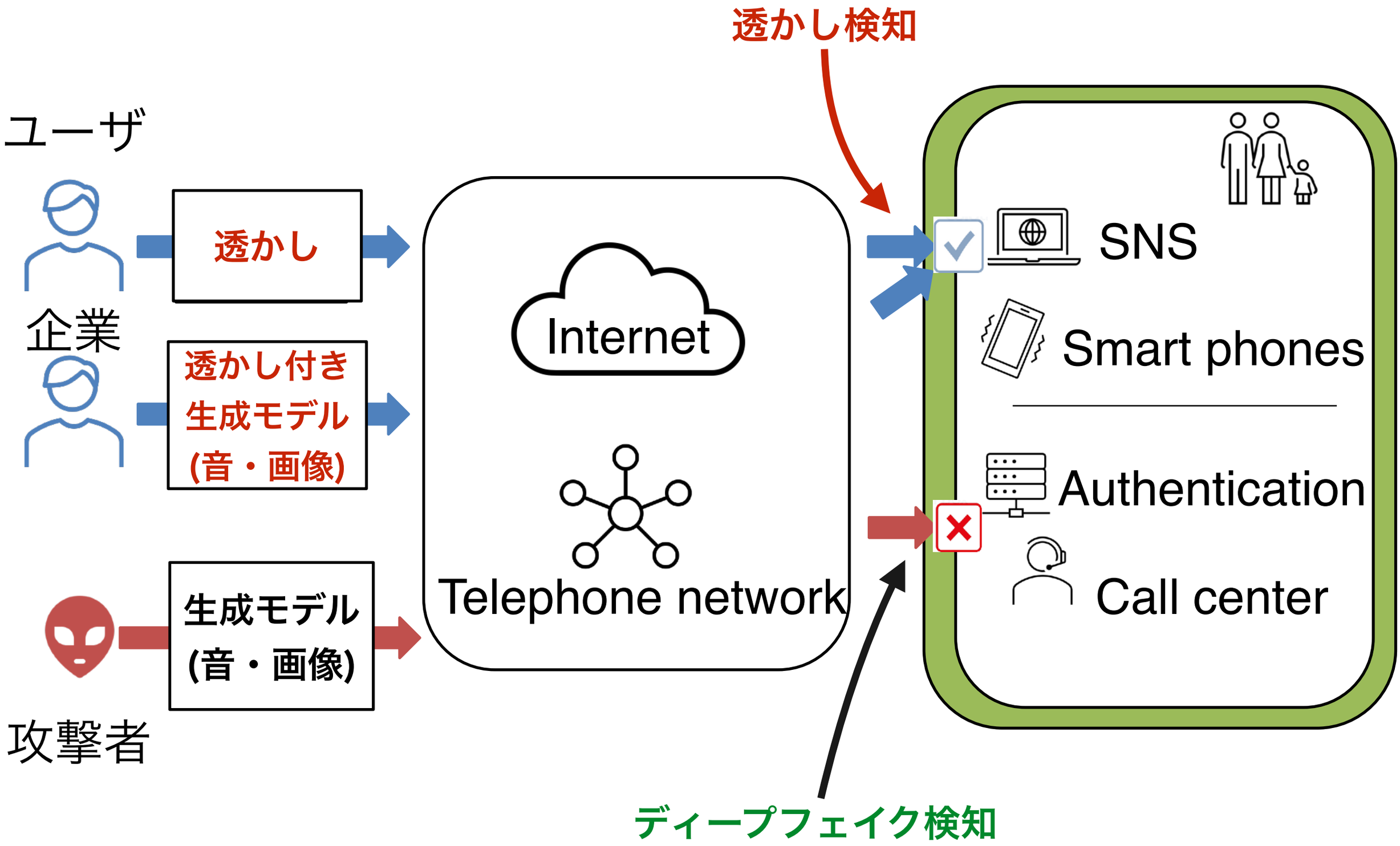


- 自動拡張された学習データベースを利用したディープフェイク検知モデルの性能はどのテストセットにおいても徐々に改善
- ランダム選択より良い
- 異なる言語に対してはまだ改善の余地あり

パート2-1

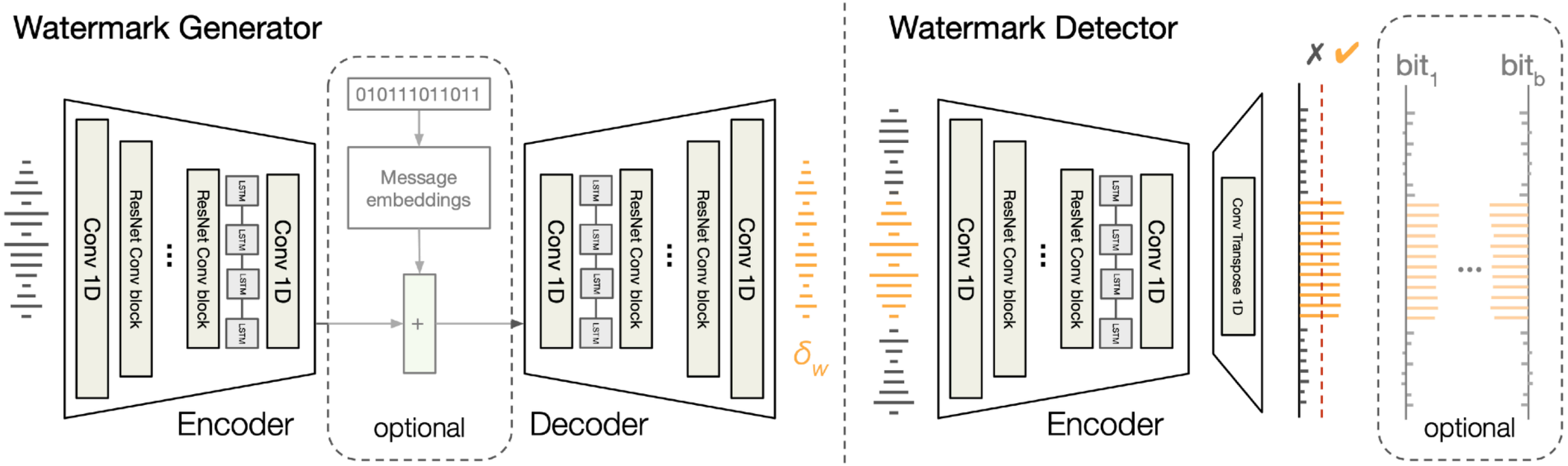
プロアクティブディフェンス：透かし

ディープフェイク検知(Passive)と透かし(Proactive)



ニューラルネットワークによる透かし入り音声波形の合成

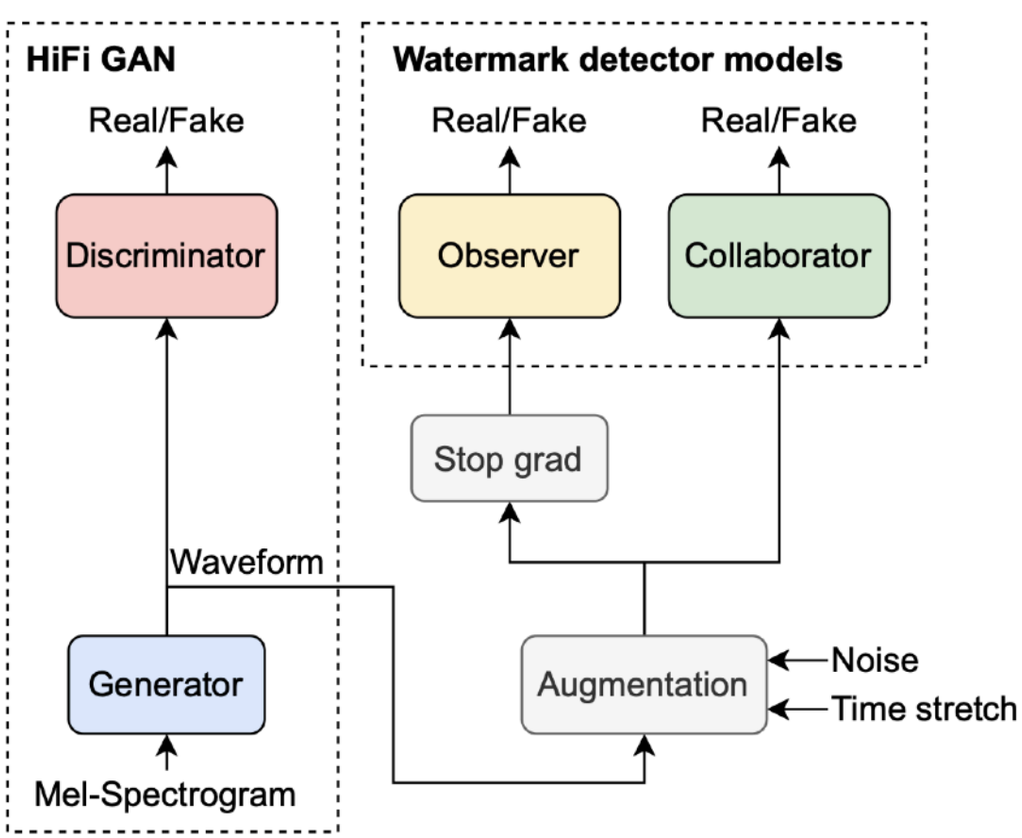
- ニューラルネットワークによる音声波形のオートエンコーダモデル
- 潜在変数空間にマルチビットの透かしを埋め込む
- 透かし情報を聴覚的には聞こえない様にするために、合成された音声は自然音声に近づける様に学習する
- その一方、合成された音声波形から透かしを抽出できる様にエンコーダモデルも同時に学習



Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, Hady Elsahar
“Proactive Detection of Voice Cloning with Localized Watermarking” ICML 2024

透かし情報が入った音声を合成するニューラルボコーダ

- 音声生成AIを提供する企業が利用することを想定
- 音を合成した後に透かしを埋め込むのでは遅延が発生
- ニューラルボコーダで音声波形を合成する際に、透かし情報が既に埋め込まれた状態で音声波形を予測できる様にモデル学習基準を工夫
  - Discriminatorを騙して自然音声と変わらない特長を有する音声を生成する基準と、聴覚上は聞こえない微少ノイズにより合成音声である と判定する基準とを同時に考慮・学習
  - これらの基準は矛盾しないことに注意



L. Juvela and X. Wang, "Collaborative Watermarking for Adversarial Speech Synthesis," *ICASSP 2024*

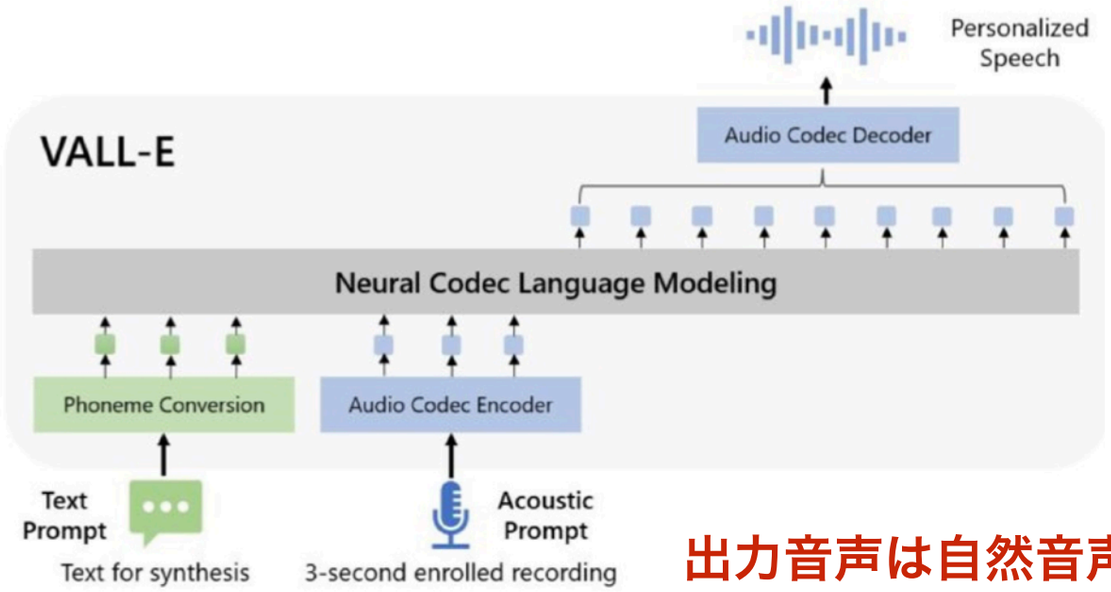
ディープフェイク検知と透かしの比較：２値分類の場合

ディープフェイク検知 **透かし検知**

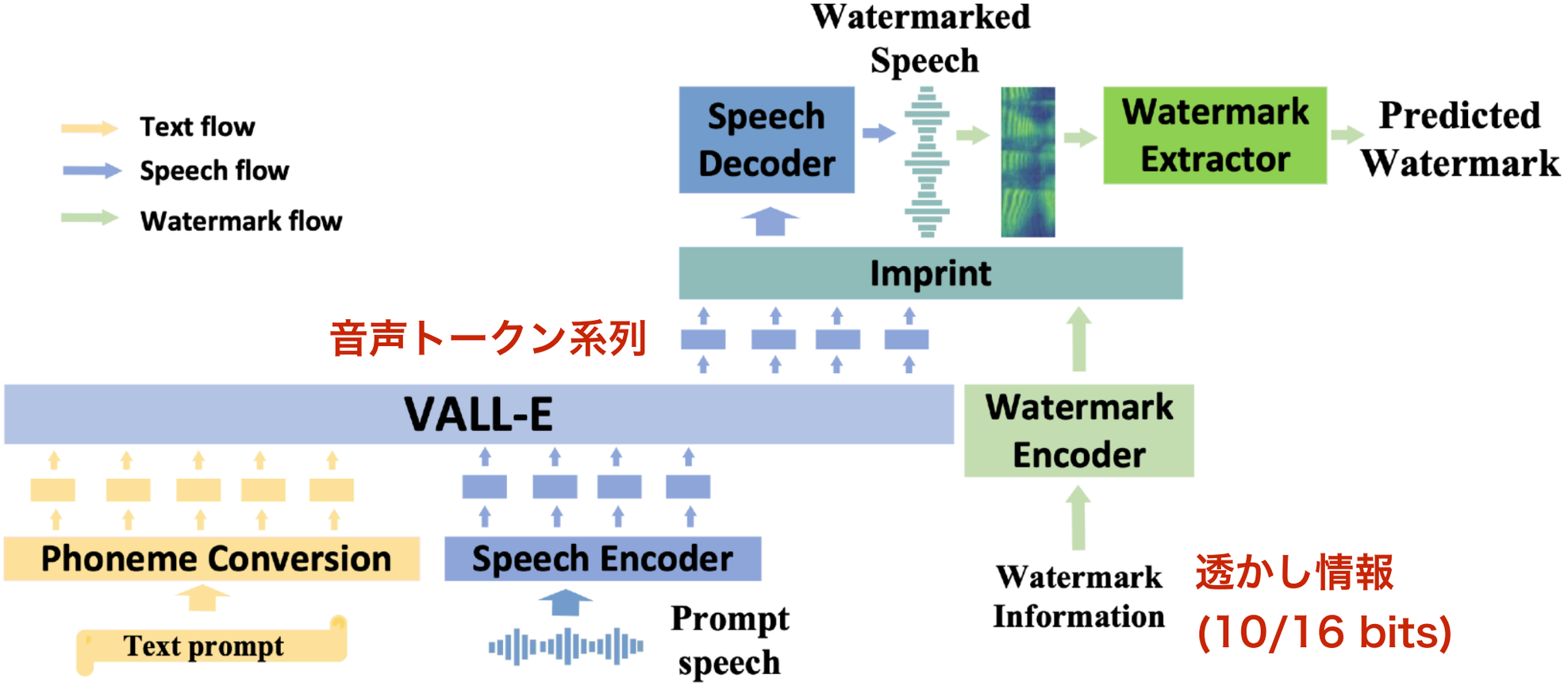
	Transmission Manipulation	EER (%)↓ of ASVspoof 2019 LA			
		Passive Models		Proactive Models	
		AASIST	SSL-AASIST	Timbre	AudioSeal
	None from § 3.3	0.83	0.23	0.00	0.00
Partially seen	Gaussian-noise	18.06	1.95 ∗	17.60	15.83 ∗
	DAC	1.66	0.27	0.01	97.40 ∗
	WavTokenizer	17.84	15.92	50.12	60.95 ∗
	Random-trimming	19.56 ∗	8.15	0.00	37.50
	Time-stretch	66.53	44.42	0.00	0.03 ∗
	Pitch-shift	66.12	48.36	52.62	47.30 ∗
Unseen	MUSAN	17.84	1.73	1.31	2.91
	RIR	35.49	4.41	0.00	57.08
	Quantization	26.15	3.31	8.66	19.59
	Compressor	9.30	1.02	0.00	0.00
	Opus	36.27	27.55	17.35	47.38
	Clipping	1.22	0.23	0.00	0.00
	Overdrive	15.30	6.19	0.11	0.00
	Equalizer	1.75	0.23	0.00	0.03
	Frequency-masking	43.32	33.11	2.94	24.40
	Noise-gate	10.56	2.56	0.13	2.56
	Noise-reduction	17.18	11.61	0.00	0.05
	Average w/o None	23.77	12.41	8.87	24.29

テキスト音声合成＋透かし

透かしのない言語モデル型
テキスト音声合成



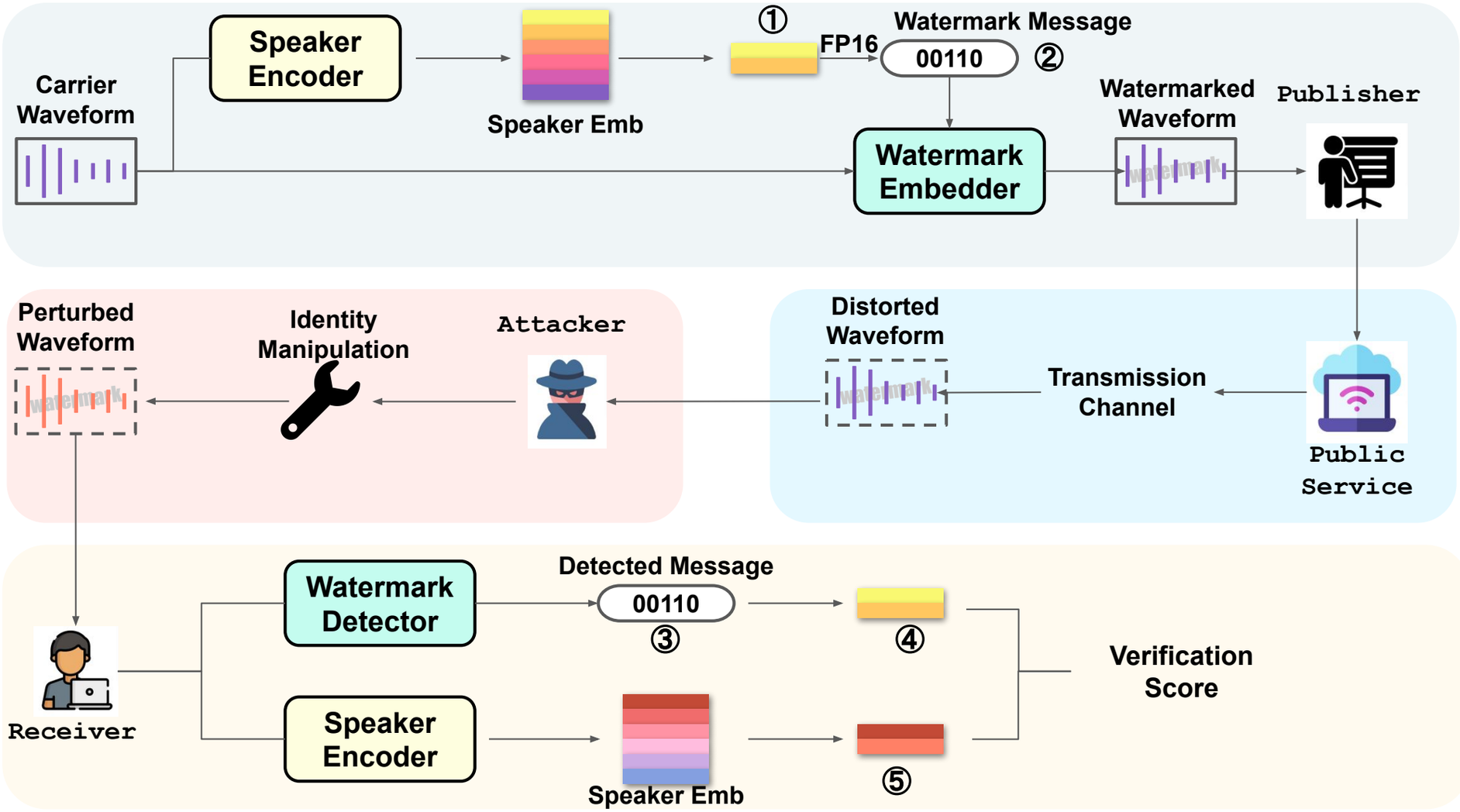
出力音声は自然音声に近づける様に学習すると同時に、埋め込んだ10/16bitの透かしを抽出させる！



TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking Junzuo Zhou, Jiangyan Yi, Tao Wang, Jianhua Tao, Ye Bai, Chu Yuan Zhang, Yong Ren, Zhengqi Wen, Interspeech 2024

透かしにはより多くの情報を埋め込める

- 自然・合成音声の情報ではなく、**発話者の話者ベクトルを埋め込み**
- 話者性の加工時には、透かし情報と音声の話者性が不一致になる
- →話者性の加工処理を自己検知！



Wanying Ge, Xin Wang, Junichi Yamagishi, “Proactive Detection of Speaker Identity Manipulation with Neural Watermarking”, The 1st workshop on GenAI Watermarking, collocated with ICLR 2025

補足とまとめ

- 透かしも完璧ではない

- 現状だとディープフェイク検知と透かしは同じくらいの精度

- 透かしのエンコーダとデコーダはペアになって利用することが仮定されているので、A社の透かしを検出できるのはA社のデコーダのみ

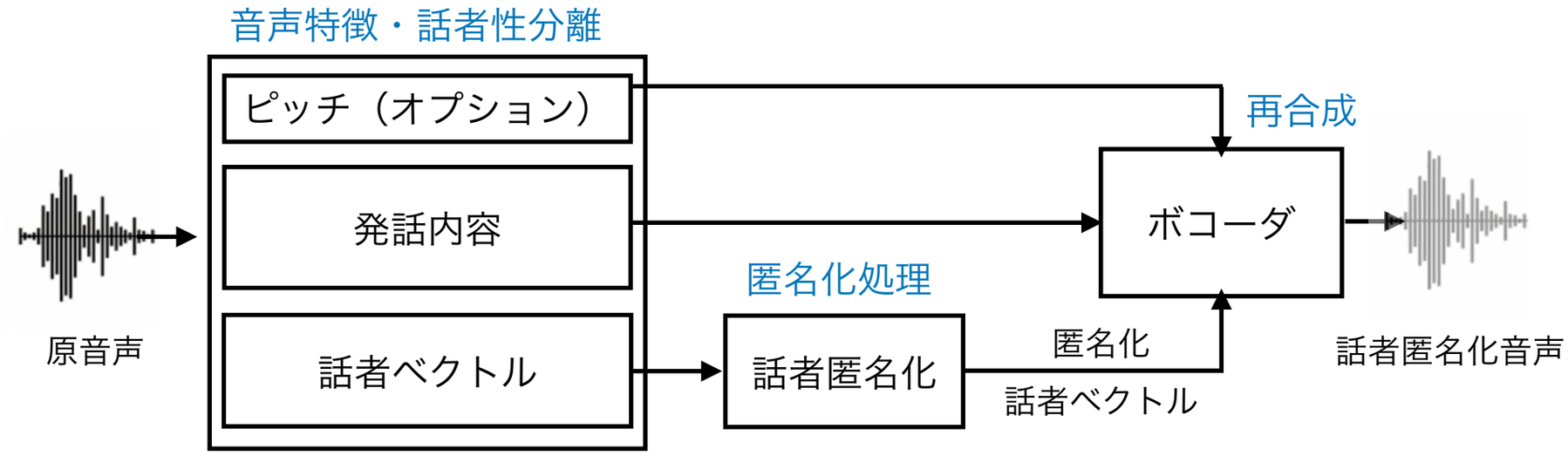
- 全ての生成AIの悪用防止ではなく、自社の製品が悪用されるのを防ぐ（自社の製品が悪用された事を証明する）用途

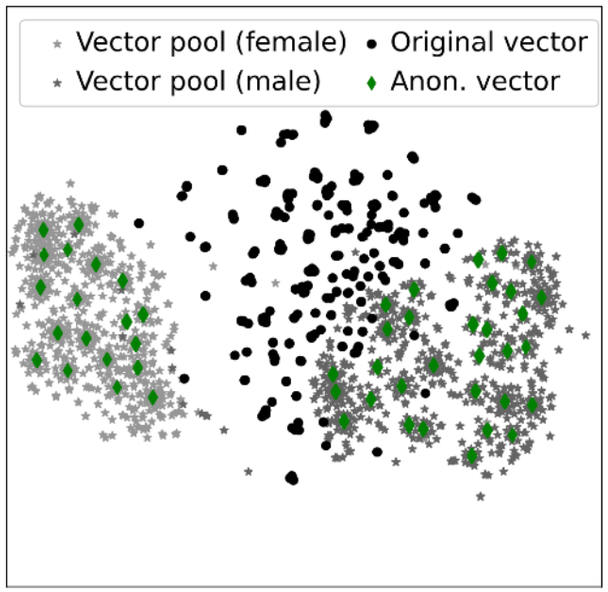
パート2-2

プロアクティブディフェンス：話者匿名化

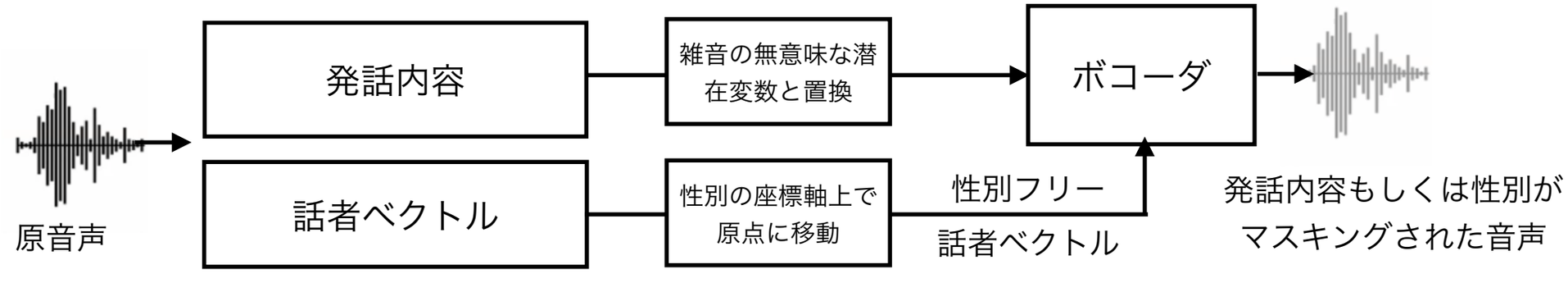
話者匿名化手法の提案

- Web上の音声データから個人特定およびディープフェイク作成が容易
- Web公開前にユーザ自身が、自分自身と紐づかない様に音声を予め変換する技術「話者匿名化」も有効
- 音声を抑揚、音素情報、および、話者ベクトルの３つの情報に分解し、話者ベクトルのみをK人の話者と平均化（K匿名化）する不可逆変換を提案





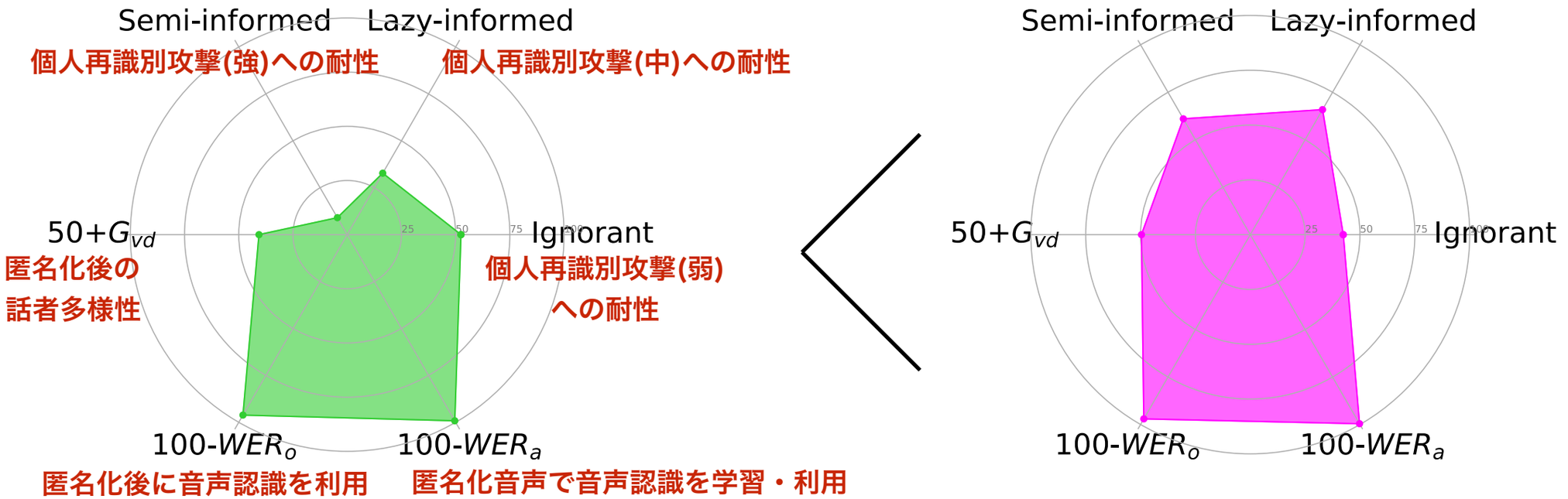
- 発話内容抽出には音声認識モデルの隠れ層[23]もしくはSSL特徴[24]を利用
- 確率密度分布(GMM)を利用した匿名化処理も提案[25]
- 同様の方法で話者性ではなく、発話内容[26]もしくは性別[27]をマスキングする枠組みも提案



[23] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, J-F Bonastre, “Speaker Anonymization Using X-vector and Neural Waveform Models,” 10th ISCA Speech Synthesis Workshop (SSW10), 2019
[24] X. Miao, X. Wang, E. Cooper, J. Yamagishi, N. Tomashenko, “Language-independent speaker anonymization approach using self-supervised pre-trained models,” ISCA Speaker Odyssey 2022, 2022
[25] B. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, M. Tommasi, “Design Choices for X-vector Based Speaker Anonymization,” Interspeech 2020, 2020
[26] J. Williams, J. Fong, E. Cooper, J. Yamagishi, “Exploring Disentanglement with Multilingual and Monolingual VQ-VAE,” ISCA Speech Synthesis Workshop 2021,2021
[27] P-G. Noé, X. Miao, X. Wang, J. Yamagishi, J-F. Bonastre, D. Matrouf, “Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline,” ICASSP 2023, 2023

話者匿名化手法の評価指標の提案

- 話者匿名化の評価指標も当然存在していなかった
- 以下の複数の指標で総合的に評価する事を提案
- 個人再識別攻撃に対する耐性[28]
    - 話者匿名化処理された音声が、悪意のある攻撃者の話者認識技術により本人と再識別されない指標
    - 内部犯行かどうか等レベルの異なる複数の攻撃者を想定し、それぞれの条件でどの程度の再識別リスクがあるか評価
- ダウンストリームタスクでの有用性[28]
    - 話者匿名化後にユーザ・企業が行いたい別のタスクにおける性能評価
        - 匿名化した状態で音声認識を利用する事を希望する場合、音声認識の性能
        - 匿名化音声DBで音声認識の学習を利用する事を希望する場合、学習された音声認識モデルの性能
- 匿名化後の話者多様性[29]
    - 複数人の会話においては、匿名化後の個人を識別できなくても、話者交代は理解できないといけない

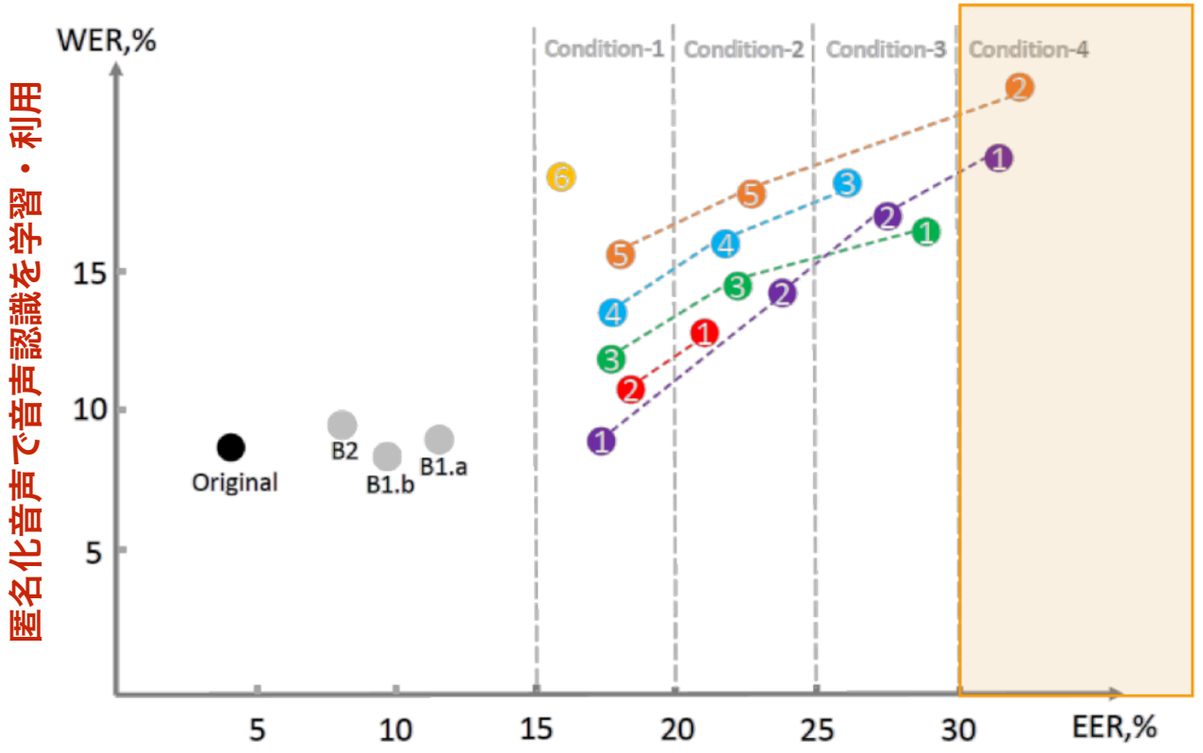


[28] Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, Massimiliano Todisco, “Introducing the VoicePrivacy Initiative”, Interspeech 2020, Oct. 2020

[29] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, Driss Matrouf, “Towards a unified assessment framework of speech pseudonymization,” Computer Speech & Language, Sept. 2021

VoicePrivacy Initiative

- 話者匿名化に関するチャレンジを2020年と2022年に運営：VoicePrivacy Challenge
- 山岸日仏共同CRESTメンバーと仏Inria研究所が協力
- 参加者が利用する音声データベース、評価セット、評価指標を規定[30]
- 大学・企業・研究組織が提案した話者匿名化手法を統一評価[30]
- 用途により求められるプライバシー保護のレベルは異なるので、プライバシー保護レベルを４つに分け、レベル毎に匿名化システムを評価
- どのシステムも各レベルである程度プライバシーを保護しているが、プライバシー保護のレベルを高く要求すると、話者匿名化後の音声の有用性が低下するトレードオフが存在
- セキュアな用途に適した匿名化手法とカジュアルな用途に適した匿名化手法を使い分ける事が望ましい



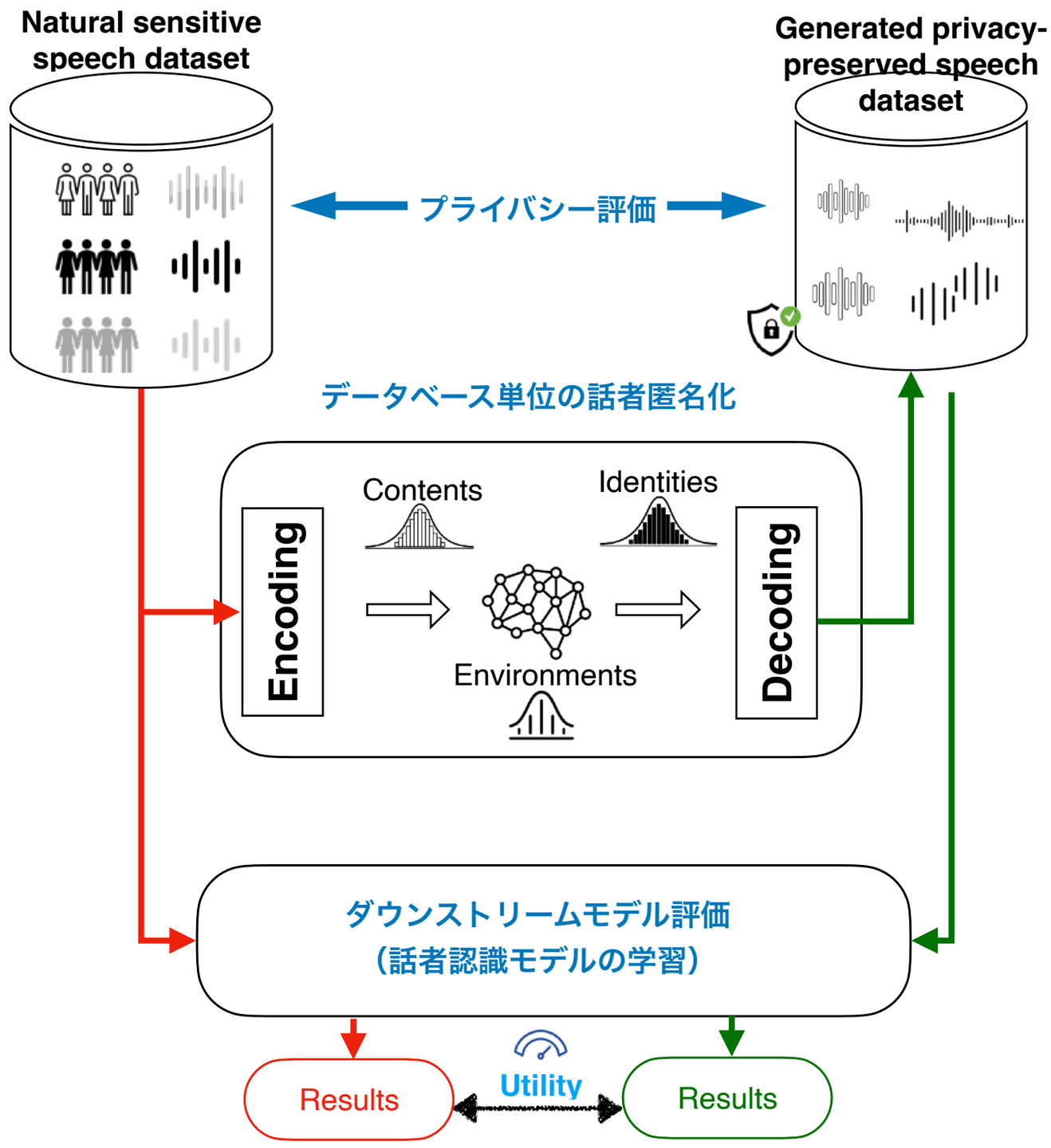
1. EER ≥15%
2. EER ≥ 20%
3. EER ≥ 25%
4. EER ≥ 30%.

プライバシー保護：個人再識別攻撃(強)への耐性

[30] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, Anaïs Chanclu, Jean-François Bonastre, Massimiliano Todisco, Mohamed Maouche, "The VoicePrivacy 2020 challenge: Results and findings," Computer Speech & Language, July 2022

発展的話題：話者匿名化から話者データベース匿名化へ

- 音声ファイル単位ではなく、データベース単位の話者匿名化手法も提案[31]
  - データベース内の話者多様性・有用性の著しい劣化を避ける様に、データベース内の話者ベクトル分布を動的に回転させる
- VoxCelebというWeb収集した実音声データベース全体を話者匿名化[32]
  - 雑音等の環境音は残す処理を追加
- 生成されたデータベースのプライバシーは保護されている(攻撃者による再識別可能性は低い)事を確認
- 生成されたデータベースの有用性には改善の余地あり
  - 生成データベース上で構築されたニューラル話者認識モデルのEERは7.3%
  - モデル学習は十分行えるが元データベースで構築された話者認識モデルのEER(1.3%)よりも高い



[31] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, Natalia Tomashenko, “Speaker Anonymization using Orthogonal Householder Neural Network,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Sept 2023

[32] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, Nicholas Evans, Massimiliano Todisco, Jean-François Bonastre, Mickael Rouvier, “SynVox2: Towards a privacy-friendly VoxCeleb2 dataset,” Submitted to ICASSP 2024

本講演の構成

- **パート１：ディープフェイク検知**
  - データベースと指標
  - （検知モデルの）何が重要？
  - 実環境での評価
  - 未知の生成手法の検知
  - 安定運用継続のための学習データ自動選択
- **パート２：プロアクティブディフェンス**
  - 2-1 音声の透かし技術
  - 2-2 話者匿名化

Q & A