第13回：応用事例と社会的インパクト

生成AI音声偽造・バイアス検証

音声処理と認識

小野原彩香 ／ 北星学園大学

この回は動画回です。Google Formsでテストを受けてください。

この回の学習目標

1. 生成AI音声技術の技術的仕組みを説明できる
2. 音声Deepfakeが社会に与えるリスクを複数の観点から考察できる
3. 音声AIにおけるバイアスの問題を具体的に指摘できる
4. 技術者・研究者として倫理的責任を自分の言葉で表現できる

配慮事項

この回は動画視聴とCBTが中心です。
字幕付き動画を利用するなど、やれる範囲で自分のペースで進めてください。
テストは持ち込み可（紙メモ）。
時間や方法について相談があればメールでご連絡ください。

なぜ音声の真正性を考えるのか

「電話で家族の声を聞いたとき、その声が本人のものか、どう確認しますか？」

- **2023年**：米FTCが、家族の声を模倣する詐欺について注意喚起
- **2024年1月**：米ニューハンプシャー州予備選前に、バイデン氏を模した自動音声電話が発信
- 個別事例の存在と、世界全体の被害件数の増減は分けて考える

この回の問い：技術で何ができるか。どんな影響があるか。利用・評価・本人確認をどう設計するか。

出典：FCC（2024）、FTCの資料は本文参照。詳しい条件・追加出典は講義本文を参照。

TTSの発展：方式と課題を区別する

代表例	主に扱うこと
規則・波形接続・統計的合成	規則、録音単位の接続、音響パラメータの生成
WaveNet（2016）	先行する波形サンプルに条件づけた生成
Tacotron 2（2018）、FastSpeech（2019）	メルスペクトログラム予測。構造・生成速度は異なる
VITS（2021）、NaturalSpeech（2022）	潜在変数などを使うテキストから波形への合成
VALL-E、Voicebox（2023）	参照音声を条件にした生成。内部方式は異なる

排他的な「第1〜第5世代」という標準分類ではありません。**波形生成器・TTS全体・話者への適応**は別の比較軸です。

出典：WaveNet、各モデルの原論文は本文参照。詳しい条件・追加出典は講義本文を参照。

Zero-shot TTSと声質変換（VC）

課題	入力と出力
TTS	テキスト＋参照音声など → 新しい内容の音声
VC	入力音声＋対象話者の条件 → 内容を保った変換音声

Zero-shot TTS：対象の新しい話者について追加のパラメータ学習をせず、学習済みモデルを利用する設定。

- VALL-E（2023）は約6万時間の英語音声で事前学習
- 約3秒の参照音声を推論時の音響プロンプトに使用
- 「3秒だけでモデル全体を学習」「事前学習データ不要」という意味ではない
- 話者類似性と、読み・内容の正確さは別々に評価する

出典：VALL-E原論文。詳しい条件・追加出典は講義本文を参照。

音声生成モデルの主な仕組み

方式	例・区別すること
GAN	生成器と識別器を使う。HiFi-GANなどは波形生成器
拡散モデル	雑音から波形や潜在表現を生成。DiffWave、AudioLDM
フローマッチング	分布間の連続的な変化を学習。**Voicebox**はこちら
潜在変数・コーデックの利用	VITSは潜在変数等、VALL-Eは音声コーデックの離散コードを使用

方式は組み合わせられます。TTS・VCという**課題の分類**とも区別します。AudioLDMは環境音等も扱い、話者クローン専用ではありません。

出典：Voicebox：非自己回帰フローマッチング。詳しい条件・追加出典は講義本文を参照。

公的資料で確認できる事件と注意喚起

時期	内容
2023年3月	米FTCが、家族の声を模倣した緊急の金銭要求に注意喚起
2024年1月21日	ニューハンプシャー州予備選前に、バイデン氏を模した自動電話
2024年1月に警察が受理	香港で幹部らを模した録画の会議映像を使い、**約2億香港ドル**の送金被害

香港の事件は**映像と音声**を使ったものです。発生年と公表年、金額と通貨、事件と注意喚起を分けて読みます。

出典：香港政府（2024年6月26日）、FCC・FTCは本文参照。詳しい条件・追加出典は講義本文を参照。

判別を難しくし得る条件

- **話者らしさ**：声が似ていることだけで本物と判定できない
- **短い参照音声**：利用できるモデルがあるが、品質は条件で変わる
- **文脈**：名前・用件などが聞き手の判断に影響する
- **通信・加工**：圧縮や帯域の変化で検出性能が変わる場合がある
- **未知の生成方式**：学習にない方式への汎化を検証する

検出スコアは「本人である確率」「詐欺である確率」と同じではありません。配信元や独立した確認経路も使います。

出典：ASVspoof 5評価計画。詳しい条件・追加出典は講義本文を参照。

社会的インパクト①：詐欺への対処

声の模倣を、緊急の金銭要求や、メール・会議映像によるなりすましと組み合わせる場合があります。

- いったん通話を切り、**以前から知っている番号や別の経路**で本人に確認する
- 着信表示や相手が提示した番号だけに頼らない
- 組織の送金では既存の承認手順や複数人での確認を守る
- 合言葉は補助策。漏えい・推測の可能性もある
- 検出器を使う場合も、見逃し・誤警報を実際の環境で評価する

出典：米FTCの本人確認の勧告。詳しい条件・追加出典は講義本文を参照。

社会的インパクト②：Liar’s Dividend

Liar’s Dividend（嘘つきの配当）：偽造技術が知られることで、本物の不都合な証拠にも「偽物だ」と疑いを向け、責任を逃れやすくなる問題。

Chesney & Citron（2019）がこの名称で論じました。単に詐欺師が金銭的利益を得ることではありません。

- 偽物を信じさせることと、本物の証拠を退けさせることの両方を考える
- 具体的な事例の発信時期・配信経路・訂正への反応を確認する
- 偽音声の存在だけで、特定の選挙結果への影響は実証できない

出典：Chesney & Citron (2019)。詳しい条件・追加出典は講義本文を参照。

声の権利：法律・規則・指針を分ける

声質、実演・録音物、氏名・肖像の利用、契約は別の検討対象です。（2026年10月19日時点）

地域	例と位置づけ
米国テネシー州	ELVIS Act（2024）：voiceを保護対象へ明記した州法。要件・例外あり
EU	AI Act第50条：2026年8月から適用する透明性義務
中国	2025年9月施行の表示規則：明示的表示・メタデータ等の表示
日本	AI事業者ガイドライン第1.2版（2026）：自主的取組の指針。既存法の義務は別

声質の模倣に著作権法上の保護が及ばない場合でも、他の権利・契約が問題になり得ます。文化庁の検討も参照。

Deepfake検出：表現と分類器

スペクトル解析と機械学習は排他的な二分類ではありません。音響特徴を分類器へ入力する方法もあります。

研究例	概要
RawNet2を用いた検出	波形から特徴を学習し、本物・偽造を分類
AASIST	時間・スペクトル情報をグラフ注意機構で扱う
wav2vec 2.0を用いた検出	自己教師あり学習の表現と分類器を組み合わせる

- すべての偽音声に共通する、高周波ノイズや不自然な位相があるとは限らない
- 未知の生成方式・話者・通信条件に対する見逃しと誤警報を測る

出典：AASIST原論文、その他は本文参照。詳しい条件・追加出典は講義本文を参照。

ASVspoof：課題と評価指標

版	主な評価設定
2015／2017	合成・変換／録音再生の攻撃
2019	LA（合成・変換）とPA（再生）を別タスクで評価
2021	LA・PA・DF。通信・圧縮等の条件も検討
ASVspoof 5（2024）	単独のDeepfake検出と、頑健な話者照合（SASV）の2トラック

EER：2種類の誤り率が等しくなる付近の値。**t-DCF**：検出器と照合器を組み合わせたコスト。

第5版の主指標はTrack 1が**minDCF**、Track 2が**min a-DCF**。実際の電話詐欺での性能は別に検証します。

出典：ASVspoof 5評価計画。詳しい条件・追加出典は講義本文を参照。

バイアス：精度差の観察と原因の検証

訓練データの代表性は重要な仮説です。平均WERの差だけで、原因を一つに確定することはできません。

確認する情報	比較の観点
話者と発話	言語・方言・年齢・発話特性、話者人数と音声時間
録音と内容	マイク、雑音、話速、語彙、読み上げ／会話
モデルと評価	版、設定、正解表記、分かち書き、話者ごとの分布
属性の扱い	ジェンダーの自己認識、データのラベル、F0を混同しない

Koeneckeほか（2020）の差は、研究対象の話者・音声と当時の5社のASRによる結果として読みます。

出典：Koenecke et al. (2020)。詳しい条件・追加出典は講義本文を参照。

精度差が利用に与え得る影響

- **アクセシビリティ**：やり直し・訂正に要する負担の差
- **医療**：記録支援の誤りと確認作業。患者の判断とは別に評価
- **教育**：発話内容より認識誤りが評価へ影響していないか
- **司法**：自動文字起こしを原音声等で確認する

取組の例：GoogleのProject Euphoniaによる非定型的な発話の研究、Mozilla Common Voiceによる参加型の音声収集。

取組の存在と、全利用者で十分な性能が得られたことは別です。

出典：Project Euphonia (2025)。詳しい条件・追加出典は講義本文を参照。

EU AI Act：音声AIの用途と義務

2024年発効・段階的適用。以下は2026年10月19日時点。

- 第50条2項：提供者による、生成物の機械可読な表示・検出可能性
- 第50条4項：Deepfake利用者の開示義務。芸術・風刺等も適切な方法で開示
- 第50条は2026年8月2日適用。同日前の既存システムの2項には12月2日までの経過措置
- 第5条：職場・教育での感情推定AIの禁止には医療・安全の例外。公共の場のリアルタイム遠隔生体識別の禁止には法執行目的等の要件

文字起こし・話者照合・感情推定は別用途です。「音声AIがすべて禁止／一律に高リスク」とは読めません。

技術者としての倫理的責任

> "Just because you can, doesn't mean you should."
> （できることと、すべきことは別の問いです）

技術者が直面する倫理的ジレンマの例

1. 依頼された研究だが悪用リスクが高い場合
2. 精度を上げるためにプライベートな音声データを使いたい場合
3. バイアスを知りながらシステムをリリースする場合

一緒に考えましょう

> 音声処理・音声認識の技術者として、あなたはどのような社会を作ることに加担したいですか？

音声AIの責任ある使用

場面	実践
研究	同意または適切な利用根拠、ライセンス、所属機関の手続を確認
開発	利用場面に合うデータと評価を整え、未評価の範囲を示す
展開	合成・加工の表示、利用目的・契約、悪用対策とその限界を確認
評価	集団別・話者別の性能と、誤りによる利用者の負担を検証

公開音声だから用途を問わず自由に使えるとは限りません。法律上の義務、契約、倫理上の指針を分けて確認します。

出典：IEEE Ethically Aligned Design（2019）。詳しい条件・追加出典は講義本文を参照。

まとめ：第13回

- **音声生成**：TTS・VCと、GAN・拡散・フローマッチング等の方式を区別
- **Zero-shot**：対象話者の追加学習なし。事前学習は必要
- **Deepfake**：偽音声を信じる問題、本物を否定するLiar’s Dividend
- **検出と認証**：目的と評価条件を分け、実運用で検証
- **バイアス**：精度差の観察と原因の特定は別
- **規制・倫理**：適用範囲、義務と指針、例外・開始日を確認

CBTテストの案内

全6問・各問4択・30分。紙メモを持ち込めます。時間や方法についての相談はメールで受け付けます。

- Zero-shot TTSと事前学習
- Liar’s Dividend
- 精度差の原因を調べる方法
- EU AI Actの開示義務
- 倫理的なリスクへの対処
- 声の類似性と読み誤りの分析

小課題・演習30%の範囲で扱います。**提出期限・提出先は授業で案内します。**

第14回（オンライン同期）：まとめと最終発表

- 授業全体の振り返り（14回の学習内容マップ）
- 期末プロジェクト・レポートの提出要件の最終確認と質疑
- 最終発表セッション

発表が困難な場合は、事前録画・スライド+テキスト・書面レポートでの代替が認められています。