話者プライバシー保護のための選択的話者情報抑制による タスク適応型フレームワーク

A Task-Adaptive Framework of Selective Speaker Information Suppression for Speaker Privacy Protection

伊藤葵 (Aoi Ito)

法政大学大学院 情報科学研究科 情報科学専攻 24T0002
aoi.ito.8q@stu.hosei.ac.jp

Abstract

With the growing demand for large-scale speech data sharing and secondary use, protecting speaker privacy has become a critical challenge. Existing approaches often operationalize privacy protection through mechanisms such as speaker pseudonymization and evaluate their effectiveness using threat-model-dependent metrics (e.g., Equal Error Rate against speaker verification systems). While these formulations provide practical benchmarks, privacy objectives and evaluation criteria are typically defined implicitly and remain tied to specific downstream tasks. In real-world scenarios, speech data are used for diverse purposes —including automatic speech recognition, conversational analysis, and attribute inference—where the information that must be preserved or suppressed varies according to task requirements. Conventional frameworks do not explicitly model this task-dependent structure of privacy and utility.

We propose a task-adaptive framework that formulates speaker privacy protection as selective speaker information suppression under explicit task requirements. Speech information is conceptualized as multiple domains, such as timbre, prosody, linguistic content, and speaker-related attributes. For each task, required domains are specified and non-essential domains are selectively transformed or suppressed. The framework is demonstrated through representative implementations and evaluated using dual criteria of task performance and privacy metrics, providing a principled and extensible foundation for task-dependent speaker privacy protection.

Supervisor: Prof. Katunobu Itou

1 はじめに

近年，大規模な音声データ共有と二次利用の需要が高まっている．音声には声色に代表される話者固有の声道特性に加え，韻律（話速・ポーズ配置・Fo など），発話内容，年齢・性別・方言といった話者属性に関する手掛かりが重畳して含まれる．これらを解析・モデル化することで，音声認識，話者認識・照合，話者ダイアライゼーション，態度認識・感情認識など多様なタスクが実現され，深層学習の発展により議事録作成や音声対話，コールセンター支援など実応用も拡大している．

一方で，音声データは個人情報を含み得るため，利活用に伴うプライバシー侵害リスクが大きい．声色や韻律，発話の癖から話者同定や属性推定が可能であり，自由発話・対話音声では人名・地名・所属などの個人識別情報（PII）が内容に含まれる場合も多い．音声データの保護は制度面でも重要視されており，GDPR では音声データを保護対象として挙げているが，具体的形態は曖昧であり技術的整備の必要性が指摘されている [1].

しかし，大規模音声の収集には高いコストが伴うため，研究では既存コーパスの共有利用が一般的である．代表例の Common Voice は音声認識研究の基盤となっているが [2]，利用規約が紳士協定に留まる場合もあり技術的保護は十分とは言い難い．また，自由発話・対話音声は価値が高い一方で PII 混入リスクが高く，一部コーパスでは仮名化が施されているものの，例えば Chiba3Party [3] のようなビープ音等の強いマスキングは理解や対話分析を阻害し得るほか，手動処理は高コストである．

音声プライバシー保護研究は主として話者照合に対する耐性向上を目的として発展してきた．VPC は匿名化手法の共通ベンチマークを整備し，EER 等の秘匿指標と下流タスク有用性を併記して評価する枠組みを提示している [4]．代表的には x-vector [5] 等の話者埋め込みを擬似埋め込みへ置換して再合成する方法が提案されている [6]．しかし，既存枠組みは話者照合中心の脅威モデルに偏りやすく，話し方（韻律・時間構造），内容中の PII，年齢などの属性といった他の漏洩経路が統合的に扱われにくい．実際，声色を変換しても話し方に基づく再識別が残り得る

ことが示唆されており [7]，継続長に基づく照合 [8] など別経路の攻撃可能性も指摘されている．さらに，多話者対話に対する匿名化では話者一貫性など固有要件があり，評価・枠組み整備も進んでいる [9]．内容中の PII 仮名化は NLP 分野で広く研究されているが [10, 11, 12]，音声では置換語の長さやリズム変化が時間的手掛かりを損ない得る．また属性推定研究も活発である一方 [13]，属性を秘匿しつつタスクに必要な情報を保持するという観点の統一枠組みは十分確立されていない．

以上より，安全な音声データ共有には，単一の匿名化操作を一律に適用するのではなく，タスク要件に応じて保持すべき情報と抑制すべき話者情報を区別し，話者情報の漏洩経路を選択的に制御する設計が必要である．本研究はこの観点から，音声を複数の情報ドメインに分解し，タスクごとに保持・秘匿を明示的に定義した上で，話者のプライバシー保護実現に向けた，タスクに不要な話者情報のみを選択的に抑制する特徴量変換型フレームワークを提案する．

2　問題設定

本節では，話者プライバシー保護をタスク依存の設計問題として再定式化するための前提を整理する．音声信号は一次元の時系列として観測されるが，実際には複数の情報が重畳した複合的媒体である．音声情報処理では，この信号から特定の情報のみを抽出して下流タスクを実現するが，タスクに必要な情報と不要な情報は同一信号内に共存している．安全な音声データ共有を考えるには，まず音声に含まれる情報ドメインを構造的に整理し，その上でタスク要件に基づく保持・秘匿の枠組みを定める必要がある．

2.1　音声データに含まれるドメイン

音声には少なくとも，(1) 声道特性に由来する音色的特徴（話者性），(2) 韻律（基本周波数推移，強勢，イントネーション，エネルギー変動），(3) 時間構造（音素長・モーラ長，ポーズ配置，ターン間の間合い），(4) 発話内容（言語情報），(5) 年齢・性別等の属性情報が含まれる．

まず話者性は，声道長や共鳴特性などに起因するスペクトル包絡の特徴として現れ，近年は深層学習により抽出される埋め込み表現が広く用いられている [5, 14]．これらの埋め込みは内容差を越えて話者固有性を安定的に表すため，再識別リスクの中心的要素となる．そのため従来の音声匿名化研究は主にこの成分の変換に焦点を当ててきた．

韻律情報は感情や態度を反映し，対話分析や態度認識に不可欠である．しかし話し方の癖やリズム傾向も個人差を持つため，再識別経路ともなり得る．声色のみを変換しても韻律が保持されれば個人性が残存し得ることが指摘されている [7]．

時間構造も独立した重要成分である．音素長，モーラ長，ポーズ位置や長さは発話リズムを規定し，対話においてはターン交替や重なりの構造にも関与する．日本語ではモーラが時間構造の基盤単位であり [15]，時間制御・音韻的長さ・産出単位・知覚単位として機能する [16, 17]．母語背景差や方言差がモーラ長分布に現れることも報告されており [18, 19, 20]，時間構造は体系的制約と社会言語学的要因の双方を反映する多層的特徴量である．したがって，時間構造は再識別リスクの源泉であると同時に，対話分析に不可欠な保持対象でもあるという両義性を持つ．

発話内容は多くの用途で中心的役割を果たすが，自由発話では人名や所属などの PII が自然に含まれ得る．単純なビープ音によるマスキング [3] は可用性を損ない得るため，時間・韻律を保った秘匿が求められる．

さらに属性情報は声色や韻律に分散して表出し，推定可能であること自体がリスクとなる．音声対話システムの普及により属性推定が容易になりつつある現在，属性制御は重要な設計課題である．

以上より，音声は複数の情報成分が重層的に構成する媒体であり，それらは部分的に相関しつつも設計上は区別可能である．安全な共有には，これらを一律に変換するのではなく，目的に応じて選択的に制御する枠組みが必要となる．

2.2　タスク依存性の問題定式化

従来の音声匿名化は，特定の攻撃モデル，とりわけ話者照合器への耐性向上を主目標としてきた．しかし実利用環境では用途は多様であり，音声認識，対話分析，態度推定，属性抑止など，タスクごとに保持すべき情報は異なる．

本研究では話者プライバシー保護を，特定攻撃への対抗操作ではなく，利用目的に応じて保持対象と秘匿対象を設計する問題として定式化する．入力音声を複数成分に分解し，タスク T に対して保持集合 $K(T)$ と秘匿集合 $S(T)$ を定義する．保持集合には下流タスクに必要な成分を，秘匿集合には再識別や属性推定につながり得る不要成分を含める．有用性と個人性は連続的信号空間内に重畳して存在するため，二値的な完全分離は困難である．そこで秘匿は完全消去ではなく，識別性能を所定水準以下に低減させる連続的制御として扱う．公開範囲やリスク許容度に応じて，個人性抑制を優先する設計と有用性保持を優先する設計を選択可能とする．この定式化により，音声匿名化は単一手法の適用ではなく，タスクに基づく設計問題として位置付けられる．保持・秘匿成分を明示し対応する特徴変換を施すことで，有用性と秘匿性を同一枠組み内で統制可能とする．本研究はこの枠組みに基づき，各成分を制御可能な構成要素として実装し，複数利用形態に対する有効性を検証する．

3　提案手法: タスク適応型フレームワーク

3.1　概要

本研究で提案するフレームワークは，話者プライバシー保護を特定の変換手法としてではなく，音声に含まれる話者関連情報を選択的に抑制する設計問題として整理するものである．第 2 章で述べたように，音声には声道特性に由来する話者性，韻律や時間構造に現れる話し方の特徴，発話内容としての言語情報，さらに年齢などの属性情報が同時に含まれている．これらは互いに独立ではなく相関を持つが，設計上は区別可能な情報単位として扱える．

従来の音声匿名化研究では，主として話者識別を想定した攻撃モデルに対抗することが目標とされ，声道特性に対応する埋

め込みの変換が中心となってきた．しかし実利用環境では，音声データの利用目的は単一ではない．音声認識，対話分析，態度推定，属性推定など，目的に応じて必要とされる情報は異なる．したがって，どの情報を保持し，どの情報を秘匿すべきかはタスクに依存する．

本フレームワークは，このタスク依存性を明示的に扱うことを出発点とする．入力音声 X を，以下の情報成分に分解して考える．

$$X \to (x, p, d, c, a)$$

ここで，x は声色に基づく話者性を主に表す埋め込み，p は基本周波数やエネルギーに基づく韻律系列，d は持続時間やポーズ配置などの時間構造，c は発話内容系列，a は属性に関する埋め込みを表す．

利用目的を表すタスクを T とする．本研究では，タスク T に対して，保持すべき情報成分の集合と秘匿すべき情報成分の集合を定義する．これをそれぞれ

$$K(T) \subseteq \{x, p, d, c, a\}, S(T) \subseteq \{x, p, d, c, a\}$$

と表す．設計上は

$$K(T) \cup S(T) = \{x, p, d, c, a\}$$

$$K(T) \cap S(T) = \emptyset$$

を満たすように定義する．すなわち，各情報成分は，話者プライバシー保護の観点をふまえ，タスクにとって保持対象または抑制対象のいずれかに分類される．話者情報抑制処理は，秘匿集合に含まれる成分のみを変換・抑制する操作として定義される．すなわち，

$$\Phi_T : (x, p, d, c, a) \to (x', p', d', c', a')$$

とし，

$$z' = \begin{cases} z & (z \in K(T)) \\ \phi_z(z) & (z \in S(T)) \end{cases}$$

とする．ここで ϕ_z は各成分に対する変換操作である．

この定式化の重要な点は，仮名化を単一のアルゴリズムとしてではなく，タスクから保持集合と秘匿集合への写像として整理する点にある．変換手法そのものは実装の問題であり，本フレームワークの本質は，どの情報を制御対象とするかを設計段階で明示することにある．

例えば，音声認識を目的とする場合には

$$K(T_{\text{ASR}}) = \{c\}, S(T_{\text{ASR}}) = \{x, p, d, a\}$$

と定義できる．対話分析を目的とする場合には

$$K(T_{\text{dialogue}}) = \{p, d\}, S(T_{\text{dialogue}}) = \{x, c, a\}$$

といった定義が考えられる．属性秘匿を目的とする場合には

$$K(T_{\text{attribute}}) = \{c, p, d\}, S(T_{\text{attribute}}) = \{a, x\}$$

のように定義することができる．

このように，タスクに応じて保持集合と秘匿集合を明示的に定義することで，話者プライバシー保護は攻撃モデル依存の変換技術から，利用目的依存の情報制御設計へと再整理される．

3.2 設計原理

本フレームワークは，以下の設計原理に基づいて構成される．

第一に，情報分解に基づく設計原理である．音声に含まれる複数の情報成分を制御単位として区別することで，保持と秘匿の境界を明示する．従来は声道特性の変換が中心であったが，本設計では韻律，時間構造，発話内容，属性も同列の設計対象とする．これにより，単一成分の変換に依存しない柔軟な設計が可能となる．

第二に，タスク駆動型設計原理である．設計の出発点を攻撃モデルではなく利用目的とする．従来の Voice Privacy Challenge 型の評価枠組みでは，話者照合器が固定された攻撃モデルとして設定され，秘匿対象は実質的に声色に基づく話者性に限定されてきた．その構造では秘匿集合はほぼ $\{x\}$ に固定される．一方，本フレームワークでは秘匿集合はタスクに依存して定義される．したがって，

$$S(T) \subseteq \{x, p, d, c, a\}$$

はタスクごとに変化する．この違いにより，本設計は単一攻撃モデルに依存せず，複数の利用形態を統一的に扱うことが可能となる．

第三に，多話者整合性原理である．実利用環境では対話音声が重要な対象となる．対話では，同一話者が一貫した擬似話者として表現される必要があり，異なる話者同士は区別可能でなければならない．そのため，話者性に対する変換は発話単位ではなく話者単位で定義される．さらに，対話構造が保持集合に含まれる場合，時間構造や相対韻律が変換によって不自然に崩れないように設計する必要がある．

第四に，評価整合原理である．保持集合に含まれる成分は下流タスク性能によって評価され，秘匿集合に含まれる成分は識別性能や推定性能によって評価される．すなわち，

$$\text{Utility}(T) \leftrightarrow K(T), \text{Privacy}(T) \leftrightarrow S(T)$$

という対応関係を持つ．この対応を明示することで，有用性と秘匿性を同一の設計空間内で議論できる．

また，ドメイン間の依存関係にも配慮する必要がある．例えば年齢は声道特性だけでなく話速や流暢性にも現れる．そのため属性秘匿を行う場合には，複数成分を組み合わせて制御する必要がある．一方で，保持対象となる成分に過度な変換を加えればタスク性能が低下する．したがって，設計は常に保持集合と秘匿集合の整合を意識して行われる．

以上により，本研究は話者プライバシー保護のアプローチを，単一の変換技術としてではなく，タスク依存の話者情報抑制設計として体系化する．次章では，この設計原理に基づく具体的な実装戦略を示す．

4 フレームワークの実装戦略

本節では，前章で定義した保持集合 $K(T)$ と秘匿集合 $S(T)$ に基づき，入力音声に含まれる各情報成分 $z \in \{x, p, d, c, a\}$ に対して，$z \in S(T)$ のときに適用する変換 ϕ_z を具体化する．本研究では音声 X を

$$X \to (x, p, d, c, a)$$

へ分解し，x は声道特性に基づく話者埋め込み，p は韻律系列，d は持続時間等の時間構造，c は発話内容系列，a は年齢等の属性埋め込みを表す．仮名化後を (x', p', d', c', a') とし，最終出力 $\hat{X}$ は再合成器 G により

$$\hat{X} = G(c', p', d', x', a')$$

として生成する．G には FastSpeech2 [21] を用い，必要に応じて Seed-VC [22] を併用して声色変換を行う．なお成分間の相関や再合成器の制約により，ある成分への操作が他成分へ波及し得るため，本節では制御点を明確化し，後段で保持・秘匿の両面から影響を検証できる形に整理する．

4.1 韻律制御

本手法は，韻律系列 p と時間構造 d を話し方由来の漏洩経路として明示し，統計量整合では残り得る話者手掛かりを，識別器表現空間で直接撹乱する [23]．入力 X から韻律特徴列 $u_{1:T} = \mathrm{ProsodyFeat}(X)$（Fo・エネルギー等）と，アライメントに基づく $d = \mathrm{Duration}(X)$（音素・モーラ長，ポーズ長等）を抽出し，$r = (u_{1:T}, d)$ とする．次に r から話者を識別する韻律話者照合器 D_{pros} を学習し，埋め込み $h = D_{\mathrm{pros,emb}}(r)$ が表す韻律漏洩空間を得る．変換は $r' = \phi_{p,d}(r)$ として，(i)h の算出，(ii) 擬似表現プール $\mathcal{H}_{\mathrm{pros}}$ による $h' = \psi_{\mathrm{pros}}(h; \mathcal{H}_{\mathrm{pros}})$ への置換，(iii)h' 条件での再生成 $(u'_{1:T}, d') = F_{\mathrm{pros}}(c, h', a)$ を行う．実装上は F_{pros} を FastSpeech2 の Fo・エネルギー・duration 予測に対応付け，p', d' を得る．この制御は声色埋め込み x の置換と独立に定義し，声色変換だけでは残り得る話し方経路を別の制御点として抑制する．

4.2 内容制御（PII 置換）

本手法は，話者プライバシー保護における再識別リスクの一因となる発話内容由来の識別情報として，成分 c に含まれる PII を扱い，話者情報抑制の経路の一つとして対話の自然性・時間構造への影響を抑えながら仮名化する．音声ではテキスト置換のみだと，置換語の長さ・韻律変化によりリズムやアクセント句境界，ポーズ配置が崩れ，対話分析・態度認識に必要な時間的手掛かりを損ない得る．そこで，(i) 置換語生成，(ii) 音声区間への整合的反映の二段に分ける．まず ASR 転記に対し LLM（GPT-4o [24] など）で候補集合 $\mathcal{C}(w)$ を生成し，(i) 同一モーラ数，(ii) 可能なら同一アクセント型，(iii) 文脈整合，(iv) 参照一貫性，(v) 必要最小限の意味要素保持を制約として与える．候補には不自然・事実改変が混在し得るため，最終的に軽量な人手チェックで逸脱候補を修正する．次に Montreal Forced Aligner [25] で該当区間を同定し，置換語のみ合成して局所的に置換する．合成時には元発話由来の Fo・話速を可能な範囲で模倣し，対話のパラ言語情報保持を補助する．

4.3 複数話者入力を想定した声色制御

本手法は，声道特性に基づく話者成分 x を秘匿対象とする際に，多話者対話で必要な (i) 同一実話者の一貫表現，(ii) 異話者の区別可能性（非重複）を同時に満たす割当規則を与える．入力 X を話者区間集合 $\{\mathcal{U}_s\}$ に分割し，各区間から $x_{s,u} = D_{\mathrm{spk,emb}}(X_{s,u})$ を抽出して代表埋め込み $\bar{x}_s = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s} x_{s,u}$ を得る．

擬似話者プール $\mathcal{P} = \{p_1, \ldots, p_M\}$ は Common Voice 21.0 から抽出した x-vector を用い，20 名分を平均して 1 擬似話者を生成する操作を繰り返して構成する（10–20 名規模の対話を想定）．この平均操作により特定個人への近接を弱め，実在話者を直接模倣することによる倫理的リスクを低減する．x-vector は話者識別に最適化された高次元の識別器空間に分布しており，距離の集中現象により単純な最遠点選択が知覚的差異を保証するとは限らない．また，識別空間上の類似度と生成音声としての知覚類似度が単調に対応する保証もない．そこで本研究では候補集合の分散構造を利用する．想定話者人数を K とし，$\mathcal{P}$ を K クラスタに分割して各クラスタ中心を求め，可能な限り直交方向に近い配置となる代表点を構成する．各実話者代表埋め込み $\bar{x}_s$ に対してはクラスタ中心との類似度を比較し，最も類似度の小さいクラスタを割当候補とする．当該クラスタ内の未使用擬似話者を選択することで，元話者との近接を抑制しつつ擬似話者間の分離を確保する．割当は PAS（Partition–Assignment–Selection）規則により写像 $\pi : \mathcal{S} \to \mathcal{P}$ を決定し，各話者に $x'_s = \pi(s)$ を付与する．本規則は距離最小化ではなく，対話内制約を満たす決定論的手順（例：ハッシュによる候補集合決定と未使用選択）として定義し，一貫性と非重複を保証する．再合成は

$$\hat{X}_{s,u} = G(c_{s,u}, p_{s,u}, d_{s,u}, x'_s, a_{s,u})$$

により行い，x 以外は保持仕様に応じて制御する．FastSpeech2 は外部埋め込みを条件入力として受ける形に改変して実装する．

4.4 属性制御

本手法は，年齢などの属性情報を成分 a として明示し，属性推定器の識別空間で識別可能性を低下させることで二次推定（悪用）を抑制する．年齢は声色だけでなく話速，duration 分布，韻律変動幅など複数成分に分散して現れるため，単一物理量の規則変換では秘匿が不十分となり得る．そこで年齢推定モデル D_{age} を学習し，$a_{s,u} = D_{\mathrm{age,emb}}(X_{s,u})$ として埋め込みを抽出，話者代表 $\bar{a}_s = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s} a_{s,u}$ を得る．学習データには Common Voice 21.0 [2]，CIAIR-VCV [26]，S-JNAS [27] を用い，child/senior/others で学習する．仮名化時は属性プール

$\mathcal{A} = \{q_1, \ldots, q_L\}$ から $k(=3)$ 個を選んで平均し，

$$a'_s = \frac{1}{k} \sum_{m=1}^{k} q_{i_m}$$

として推定器空間上の識別信号を希釈する．出力は

$$\hat{X}_{s,u} = G(c'_{s,u}, p'_{s,u}, d'_{s,u}, x'_s, a'_s)$$

で生成し，年齢制御は声色制御と独立に定義する．スマートスピーカー等ではアクセス権限に応じた情報開示制御として，ユーザー識別に必要な最小限の機能は維持しつつ，年齢推定などの二次推定を抑制できる．

5 評価設計

本章では，本研究の評価設計を述べる．本研究は，タスク要件に応じて保持成分と秘匿成分を分離する設計思想に基づくため，評価もまた秘匿性能と有用性の二軸で構成する．秘匿性能は再識別攻撃を想定し，声道特性 x に基づく照合と，韻律・時間構造 (p, d) に基づく照合を分離して検証する．指標には Equal Error Rate（EER）を用い，話者数 470 名，登録 5 発話，照合 5 発話の統一条件で比較する．VoicePrivacy Challenge 2022 の慣例を参照し，本研究では 20%，30% を目安値として解釈する．攻撃者は仮名化規則を知らず，登録音声は原音声とする保守的設定を採る．有用性は保持集合に対応させ，態度保持は mimiAIR による態度ベクトルのコサイン類似度，内容保持は PII 部分を除外した WER で評価する．さらに年齢推定 EER により属性秘匿も確認する．以上により，秘匿と有用性を仕様対応的に検証する．

6 ケーススタディ

本稿では，3 つのケースを考える．音声を

$$X \to (x, p, d, c, a)$$

と成分分解し，タスク要件に応じて保持集合 $K(T)$ と秘匿集合 $S(T)$ を定義する．各ケースでは，想定利用状況ごとに保持・秘匿仕様を与え，対応する制御点を組み合わせて実装する．秘匿性能は識別器性能の低下で，有用性は下流タスク性能で評価する．漏洩経路は声色 x と韻律・時間構造 (p, d) を分離して扱い，必要に応じてドメイン識別も補助的に用いる．

6.1 ケース 1：音声認識用学習データの共有

音声認識学習データの共有を想定する．発話内容系列 c を保持対象とし，声色変換後も残存し得る韻律 p と時間構造 d を秘匿対象とする．

$$K(T_1) = \{c\}, \qquad S(T_1) = \{p, d\}.$$

実装では，韻律側入力 $r = (u_{1:T}, d)$ を構成し，韻律話者照合器の埋め込み空間上で r を擬似表現へ写像した後，c を条件として (p', d') を再生成する．時間構造の制御点としてモーラ長系列も抽出し，モーラ長に基づく方言識別でドメイン情報の低減を確認する．評価は WER による内容保持，Prosody EER による秘匿性能，および方言識別正解率で行う．

6.2 ケース 2：自由発話・複数人対話データの共有

自由発話を含む複数話者対話音声の共有を想定する．内容 c，声色 x，韻律・時間構造 (p, d) を秘匿対象とする．

$$S(T_2) = \{x, p, d, c\}.$$

運用上は話者ラベルやターン境界などの構造メタ情報を保持する．実装は三つの制御点からなる．(1) 内容：PII スパンを検出し，モーラ数（可能ならアクセント型）一致を制約として LLM により置換語を生成する．(2) 声色：話者単位代表埋め込みを抽出し，擬似話者プールから対話内非重複を満たす規則で x を置換する．(3) 韻律・時間構造：韻律話者照合器の表現空間に基づき識別可能性を低下させる向きに変換し，再合成条件に与える．評価は Voice EER と Prosody EER を併記し，態度スコアのコサイン類似度および WER（PII 区間除外）で有用性を測る．

6.3 ケース 3：スマートスピーカーにおける年齢推定の抑止

スマートスピーカー環境での年齢推定悪用リスクを想定する．声色 x と内容 c を保持し，属性埋め込み a と韻律・時間構造 (p, d) を秘匿対象とする．

$$K(T_3) = \{x, c\}, \qquad S(T_3) = \{a, p, d\}.$$

年齢推定器の中間表現として Age-vector を学習し，擬似 Age-vector へ置換することで推定性能を低下させる．併せて p, d も変換し，再合成条件に反映する．評価は年齢推定の Accuracy，Macro-F1，EER を主要指標とし，必要に応じて内容保持や声色安定性も確認する．

7 結果

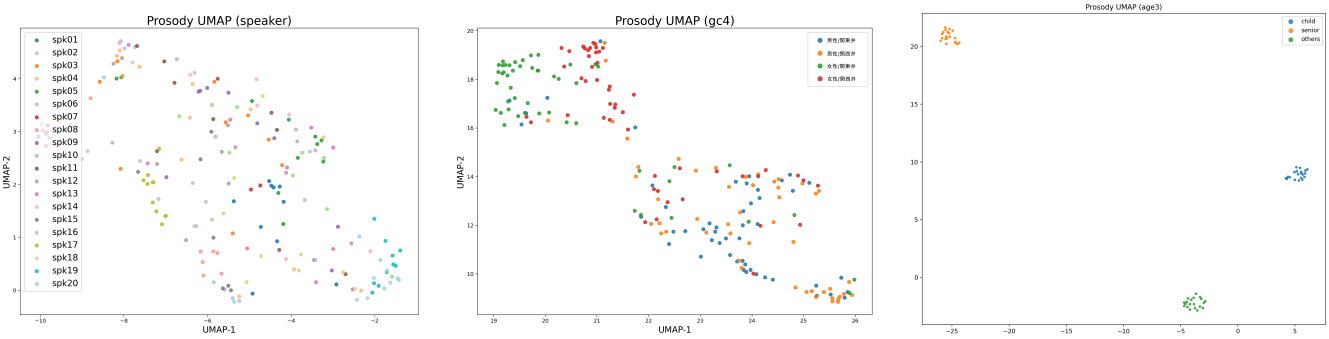
本章では，秘匿（EER ↑，Accuracy ↓）と保持を同一表内で整理した結果を示す（表 1）．本研究では，秘匿対象に含めた成分は識別・推定性能の低下で評価し，保持対象に含めた成分は下流タスク性能で評価するという設計原理に基づき，各ケースを解釈する．

ケース 1（ASR 学習データ共有）では，韻律・時間構造を秘匿対象とした結果，Prosody EER が 30.25% に上昇した．これは，韻律系列のみからの話者識別が実用的に困難な水準まで低下したことを意味する．また，モーラ長系列に基づく方言識別精度は 68.28% から 64.19% に低下し，時間構造がドメイン識別に利用され得ること，およびその寄与が一定程度抑制されたことを示している（図 2）．スペクトログラム上でも，声色に関わる周波数構造は概ね維持されつつ，ポーズ位置やモーラ長といった時間的配置に変化が確認でき，設計意図に沿った変換が行われていることが視覚的にも確認された．

ケース 2（多話者対話共有）では，声色・内容・韻律を同時に秘匿対象とした．その結果，仮名化後は Voice EER が定義不能となり，登録・照合の対応付けが成立しない水準まで話者照合が困難化した．一方で，対話利用において重要な態度情報は平均コサイン類似度 0.937 と高水準で保持された．UMAP 可視化（図 3）では，元話者と擬似話者が空間的に分離しつつ，擬似話者同士も区別可能である様子が確認できる．さらに，韻律分布（図

Table 1: 各ケースにおける統一評価結果

Case	Voice EER (%)	Prosody EER (%)	Dialect Acc. (%)	Age EER (%)	WER (%)
Original	3.87	25.53	68.28	0.00	35.25
Case1 (ASR share)	35.32	30.25	64.19	–	50.23
Case2 (Dialogue share, original)	2.45	27.44	–	–	30.14
Case2 (Dialogue share, pseudonymized)	53.47	29.67	–	–	29.55
Case3 (Smart speaker, age)	10.89	–	–	55.56	38.66



(a) 韻律話者照合モデル　(b) 性別 × 方言識別モデル　(c) 年齢推定モデル

Fig.1: 各識別モデルから抽出した埋め込みの分布を UMAP により二次元へ次元削減して可視化した結果．各点は発話を表し，色は各モデルにおけるラベル (話者，方言 × 性別の 4 クラス，年齢層の 3 クラス) に対応する．韻律に基づく埋め込み空間においても，話者・属性・ドメインに対応したクラスタ構造が観察され，韻律が再識別や属性推定の手掛かりとなり得ることが示唆される．

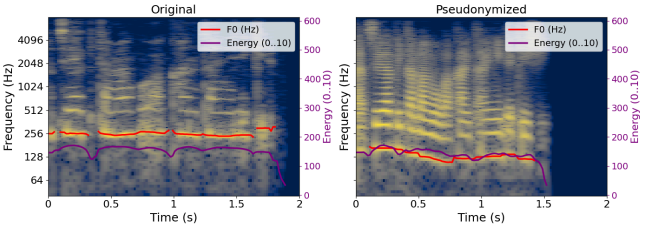


Fig.2: 韻律仮名化前後のスペクトログラム例（発話「厚焼き卵は簡単にできる」）．左図：元音声，右図：韻律仮名化後．周波数帯域のエネルギー分布（声色に関わる成分）は概ね維持されている一方で，格助詞「は」の後のポーズ位置やモーラ長など時間構造に変化が見られる．

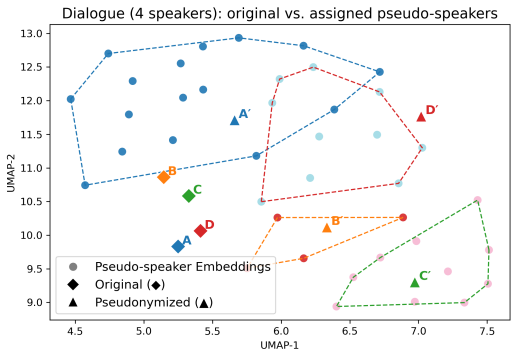


Fig.3: 対話セッションにおける埋め込み分布の可視化．円は擬似話者候補の分布，菱形は元話者（A–D），三角形は割り当てられた擬似話者（A'–D'）を表す．A'–D' が互いに分離し，かつ対応する元話者から離れた位置に配置されているかを観察する．

4）や PII 置換例（図 5）から，統計的韻律特性と時間的連続性が大きく崩れていないことが観察され，実際に聴取した場合にも対話の自然さが大きく損なわれないことが確認された．

ケース 3（年齢推定抑止）では，年齢を秘匿対象とした結果，Accuracy 0.100，Macro-F1 0.061，Age EER 55.56% となり，チャンスレベル近傍まで推定性能が低下した．UMAP 上でも年齢クラスの分離が縮小しており（図 1），韻律・時間構造に基づく年齢識別手掛かりが弱まったことが示唆される．以上より，各ケースにおいて，指定した漏洩経路の識別困難化と，指定した保持対象の維持が概ね両立していることが確認された．

8　考察

本研究は，音声仮名化を話者照合耐性の向上という単一目的から拡張し，タスク要件に応じて保持情報と秘匿情報を明示的に設計する情報制御問題として再定式化した．音声には内容，声色，韻律，時間構造，属性といった複数の情報成分が重畳して含まれており，単一の匿名化操作を一律に適用するだけでは，必要な情報を過剰に損なう一方で別の漏洩経路を残す可能性がある．本研究はこの状況を設計段階で扱うため，音声を (x, p, d, c, a) に

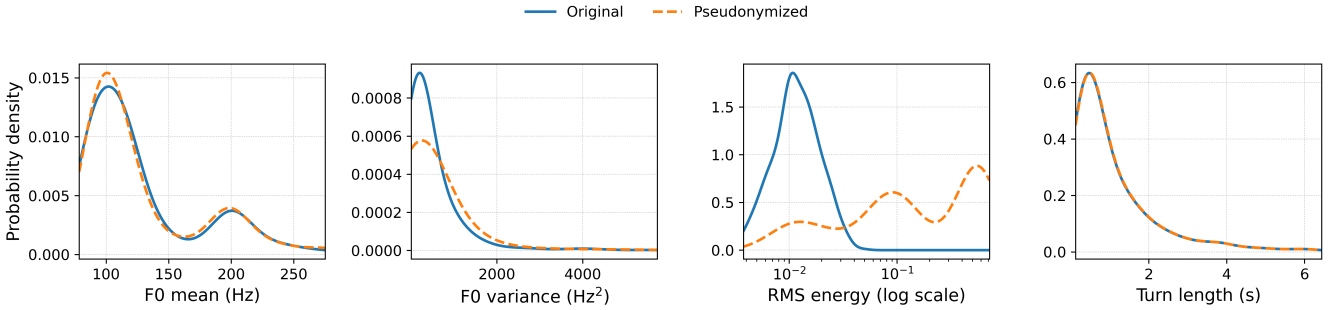


Fig.4: オリジナル音声と仮名化音声における韻律指標分布．F_0 平均・分散，RMS エネルギー，ターン長を比較している．両条件で分布形状が大きく重なるほど，統計的な韻律特性が保持されていることを示す．

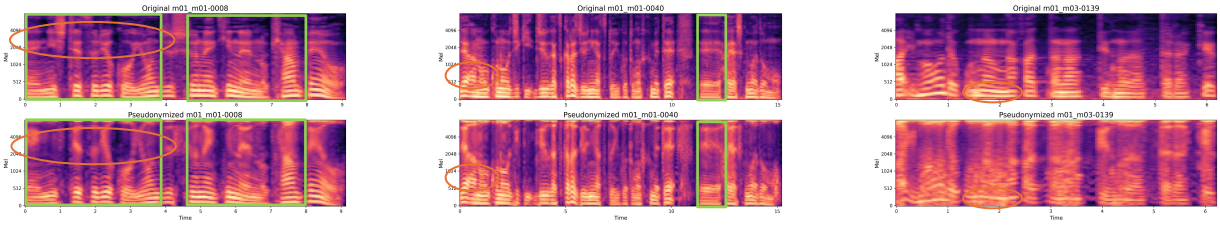


Fig.5: PII 置換例のスペクトログラム（上：元音声，下：仮名化音声）．薄緑の領域は PII 区間を表し，該当部分が擬似表現へ置換されていることを示す．オレンジの丸は声色やエネルギー分布が変化している領域であり，一方で時間構造や F_0 に対応する強度パターンが大きく崩れていないかを視覚的に確認できる．

分解し，タスク T に対して保持集合 $K(T)$ と秘匿集合 $S(T)$ を定義する枠組みを提示した．これにより，「何を守り，何を抑制するか」を仕様として記述し，秘匿性と有用性を同一空間で議論する基盤を与えた．

三つのケーススタディでは，指定した漏洩経路に対応する識別指標が低下し，保持対象に関わる有用性指標が一定程度維持されることが確認された．この結果は，成分ごとに制御対象を明示する設計が実際の指標変化として観測可能であることを示唆している．特に，声色のみならず韻律や時間構造を独立した制御対象とした点は重要である．Case1 では韻律経路の識別困難性が上昇し，時間構造に基づく方言識別精度も低下したことから，話し方由来の個人性が再識別やドメイン推定に寄与し得る一方，適切な変換によりその影響を弱められる可能性が示された．これは，声色変換のみでは十分でないという既往研究の指摘とも整合的である．さらに，多話者対話に特有の要件を設計に組み込んだ点も本研究の特徴である．対話データでは，同一話者の一貫性と異話者間の分離性を同時に満たす必要がある．擬似話者プールと決定論的割当規則を導入することで，対話内で非重複かつ安定した写像を実現し，声色および話し方双方の漏洩経路が抑制される傾向を確認した．同時に，態度表現の類似度が高水準で維持されたことは，強いマスキングに依存しない対話データ共有の可能性を示唆している．一方で，成分間の完全分離は保証されず，ある成分の変換が他成分へ影響を及ぼす可能性がある．また，識別指標は評価条件に依存するため，数値の解釈は設定に限定される．内容仮名化や擬似話者設計にも実装上の制約が残る．今後は，複合タスクに対応した保持・秘匿集合の統合設計，成分間干渉を低減する生成モデルの改良，および擬似話者・擬似属性プール構成の定量的指針の確立が課題である．以上より，本研究は音声仮名化を攻撃耐性最大化の問題から，タスク要件に基づく仕様駆動型設計へと再構築し，漏洩経路分離と多指標評価に基づく議論枠組みを提示したと位置付けられる．

9 結論

本研究は，音声仮名化を話者照合中心の設計から拡張し，利用目的に応じて保持情報と秘匿情報を仕様として定義できる枠組みを提示した．音声には内容，声色，韻律・時間構造，属性といった複数成分が重畳して含まれるため，単一の匿名化操作では用途に必要な情報を損なう一方，別の漏洩経路を残す可能性がある．本研究は，この問題を設計段階で扱う原理と評価体系を与えた．

提案フレームワークは，入力音声 X を (x, p, d, c, a) に分解し，タスク T に対して保持集合 $K(T)$ と秘匿集合 $S(T)$ を定義することで，仮名化を選択的情報制御問題として定式化する．秘匿対象のみを変換し，保持対象は下流タスク性能で，秘匿対象は識別・推定性能の低下で評価することで，設計と評価の整合を確立した．本枠組みの意義は，音声仮名化を用途横断で設計・比較可能にする共通基盤を与えた点にある．さらに，声色に限らない複数の漏洩経路を明示し，韻律・時間構造や属性推定リスクを同一設計空間で扱えることを示した．多話者対話においても，一貫した変換規則を設計要素として組み込み，利用可能性と秘匿

性の両立を図った．一方で，成分間の完全分離は保証されず，変換が保持対象へ影響する可能性や評価指標の条件依存性といった課題が残る．今後は，複合タスク仕様の統合，成分間干渉を抑える生成設計，および擬似話者・擬似属性プール設計の定量的指針の確立が必要である．以上より，本研究は音声仮名化を攻撃耐性最大化の問題から，タスク要件に基づく仕様駆動型設計へと再構成した．本成果は，安全な音声データ共有に向けた方法論的基盤を提供するものである．

10 謝辞

本研究は，JST CREST JPMJCR21M4 の支援を受けたものである．

参考文献

[1] A. Nautsch *et al.*, “The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding,” in *Proceedings of Interspeech*, Sep. 2019, pp. 3695–3699.

[2] R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *Proc. 12th Lang. Resources Eval. Conf. (LREC)*. Marseille, France: European Lang. Resources Assoc. (ELRA), 2020, pp. 4218–4222.

[3] 伝. 康晴 and 榎. 美香, “千葉大学 3 人会話コーパス（chiba3party）,” Jul. 2014.

[4] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, “The voiceprivacy 2022 challenge evaluation plan,” 2022.

[5] D. Snyder *et al.*, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[6] F. Fang, X. Wang, J. Yamagishi *et al.*, “Speaker anonymization using x-vector and neural waveform models,” in *Proc. 10th ISCA Speech Synthesis Workshop (SSW10)*, 2019, pp. 155–160.

[7] N. Tomashenko, E. Vincent, and M. Tommasi, “Exploiting context-dependent duration features for voice anonymization attack systems,” 08 2025, pp. 5128–5132.

[8] Y. I. Kenichi FUJITA, Atsushi ANDO, “Speech rhythm-based speaker embeddings extraction from phonemes and phoneme duration for multi-speaker speech synthesis,” *IEICE TRANSACTIONS on Information*, vol. E107-D, no. 1, pp. 93–104, January 2024.

[9] X. Miao, R. Tao, C. Zeng, and X. Wang, “A benchmark for multi-speaker anonymization,” *arXiv preprint arXiv:2407.05608*, 2024.

[10] K. Hisamoto *et al.*, “Anonymization techniques for publishing civil court judgments,” in *Proc. 28th Annual Meeting of the Association for Natural Language Processing*, 2022, pp. 1406–1410.

[11] O. Yermilov, M. Kartashov, and A. Panchenko, “Privacy- and utility-preserving nlp with anonymized data: A case study of pseudonymization,” in *Proc. 3rd Workshop on Trustworthy NLP (TrustNLP)*, 2023, pp. 232–241.

[12] T. Vakili *et al.*, “End-to-end pseudonymization of fine-tuned clinical bert models: Privacy preservation with maintained data utility,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 162, 2024.

[13] K. Fujii and R. Nishimura, “A study of age estimation task for young speakers using age-embedded features,” in *Proc. ASJ Spring Meeting*, 2025.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” 10 2020.

[15] N. Trubetzkoy, *Principles of Phonology*. University of California Press, 1969. [Online]. Available: https://books.google.co.jp/books?id=5Xd8yQEACAAJ

[16] H. Kubozono, *Mora and Syllable*. John Wiley & Sons, Ltd, 2017, ch. 2, pp. 31–61. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405166225.ch2

[17] 窪薗晴夫, “モーラと音節の普遍性 (特集 音節とモーラの理論),” **音声研究**, vol. 2, no. 1, pp. 5–15, 1998. [Online]. Available: https://cir.nii.ac.jp/crid/1390001204786897920

[18] K. Sugai, “Mental Representation of Japanese Mora; Focusing on its Intrinsic Duration,” in *Proc. Interspeech 2017*, 2017, pp. 2973–2977.

[19] 賈. 海平, 森. 大毅, and 粕. 英樹, “話速の変化に対する日本語の促音・長音の時間構造の分析に基づく日本語学習者の習熟度評価：中国語母語話者を例として,” **日本音響学会誌**, vol. 62, no. 6, pp. 433–442, 2006.

[20] 堀. 智子 and 森. 庸子, “日本語の自然発話における伸長の程度と生起環境—大学生の絵描写課題から—,” **音声研究**, vol. 20, no. 2, pp. 38–47, 2016.

[21] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.

[22] S. Liu, “Zero-shot voice conversion with diffusion transformers,” *arXiv preprint arXiv:2411.09943*, 2024.

[23] A. Ito and K. Itou, “Speaker pseudonymization for japanese speech using duration embeddings,” in *2024 International Symposium on Multimedia (ISM)*, 2024, pp. 41–48.

[24] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.

[25] M. McAuliffe *et al.*, “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in *Interspeech*, 2017, pp. 498–502.

[26] CIAIR, Nagoya University, “Ciair children voice speech corpus (ciair-vcv),” 2006, speech corpus dataset. [Online]. Available: https://doi.org/10.32130/src.CIAIR-VCV

[27] K. Shikano, “Japanese newspaper article sentences read speech corpus of the aged (s-jnas),” 2007, speech corpus dataset. [Online]. Available: https://doi.org/10.32130/src.S-JNAS