CL-003

態度分析タスクのための音声対話データ仮名化

伊藤 葵 *,　伊藤 克亘 †
Aoi Ito,　Katsunobu Itou

1　はじめに

近年，機械学習やデータ分析の目的で音声データを収集・保持する機会が増加している．音声データは，内容の有用性に加えて，話者の個人情報を含む点で機微性が高く，EU 一般データ保護規則 (GDPR) においても，その保護の重要性が明示されている．このような背景から，音声データを安全に保持・共有するための技術的対策が求められている．従来は，暗号化によってデータそのものを保護する方法が取られてきたが，音声データは通常浮動小数点形式で記録されるのに対し，多くの暗号システムは整数データを前提として設計されているため，実用上はデータ形式の変換を伴うという課題があった．

これに対し，音声データの特徴をふまえて話者情報の保護を目指す手法が音声仮名化である．音声仮名化とは，話者のプライバシーを保護するため，音声中の個人性が表れる特徴を変換し，話者情報を秘匿する技術である．音声匿名化とは異なり，個人同定にはつながらないが発話内の有用な情報 (発話内容，パラ言語情報) は仮名化処理後も保持されるため，音声データの安全な保持や共有が可能となる．特に，機械学習においては大量の学習データが必要とされることが多く，音声仮名化技術により個人情報を適切に保護することで，音声データの収集・利活用に関する倫理的・法的ハードルを低減し，データ収集の円滑化が期待される．従来の音声仮名化手法では，声色や韻律に対するピッチシフトや話速変換，あるいは機械学習による擬似話者音声の生成などが行われてきた．しかし，発話内容から話者が特定されるリスクは依然として残されている．また，従来の音声仮名化手法は，一発話・一話者単位での処理を前提としており，複数話者間の相互作用や文脈情報が重要となる態度・感情認識の観点からの対話音声データへの応用は，まだ初期的な段階にある．自然な対話を含む音声コーパスとして，CHiba3Party [1] のような自由発話型のデータセットがある．このコーパスは，友人同士による日常的な会話を対象として収録されており，音声・テキストデータとして，自然な対話に見られるパラ言語情報や会話の間が残存した自由発話の数少ないコーパスの一つである．このコーパスではプライバシー保護の観点から仮名化処理が施されており，例えば収録時は互いを「A さん」「B さん」といった仮名で呼ぶようにしている．しかし，実際のやり取りでは，B さん役の人に対して C さんと声をかけるなど，名前の呼び間違いが頻発しており，本来見知った人同士での自然な対話では起こらないであろうやり取りが生じている．また，研究室名や組織名など，特定の人物・団体に結びつく語彙が発現された場合は，テキストデータ上では適当な名称に置き換えられ，音声データ上では該当箇所にビープ音によるマスキングが施されている．その結果，当該箇所において発話本来の連続性が損なわれ，聞き取り困難な箇所が生じるなど，データの可用性・自然性の両面に影響を及ぼしている．このように，音声とテキスト間の関係性を無視した異なる仮名化処理が施され，対話音声の自然性が損なわれるという課題は，今後話者のプライバシー保護環境下で対話音声データを取り扱うにあたり大きな問題となる．

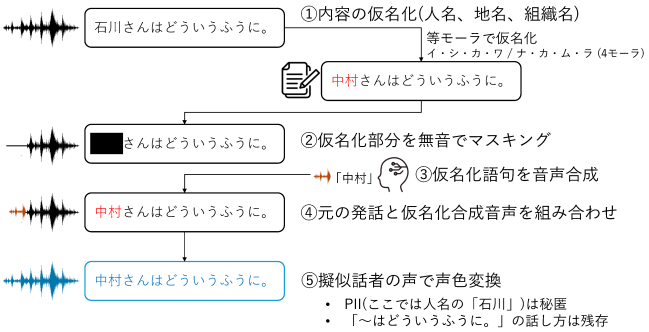

図 1. 提案手法の概要

本稿では，対話音声データに対し，声色と発話内容の両面における仮名化手法を提案する．特に，仮名化された内容を音声として生成するうえで，本来の発話の自然性が損失しすぎないように，発話内容の仮名化では元の単語とモーラ数が等しい単語に置き換え音声合成し，元の発話と同じ表現の箇所に対しては声色変換を組み合わせることで，対話音声データに含まれるパラ言語情報を保持した仮名化を実現する．

2　関連研究

2.1　音声データにおけるプライバシー保護の動向

音声データの個人情報保護には依然として多くの課題があり、GDPR においても音声は保護すべき情報として明記されているが、どの部分が個人情報に該当するかの明確な定義はされていない。このような不確実性を背景に、音声データに含まれる個人情報の保護に関する方向性や適用場面を整理した研究 [2] では、プライバシー概念の整理を通じて、生体認証データとしての音声の可能性や各国におけるセンシティブデータの扱いを分析している。また、音声データの収録・保持過程におけるプライバシー侵害のリスクやその対策についても検討されているが、現時点では音声データに対する保護技術は十分に体系化されておらず、技術的・運用的に解決すべき課題が多く残されている。

そのような背景のもと，Voice Privacy Challenge (以下，VPC) に代表される音声匿名化・仮名化の手法に関する国際的なコンペティションが開催されるようになり，話者を保護するさまざまな技術が提案されている．現時点では，話者照合モデルから得られる話者特徴量を基に，入力された発話を擬似話者の声色へと変換する手法 [3] がベースライン手法の一つとして採用されている．また，

*法政大学大学院 情報科学研究科
†法政大学 情報科学部

VPC2024 の課題では，音声データに含まれる感情表現をどの程度保持できるかが新たな評価指標として加えられ，個人情報保護ならびに音声の幅広い有用性の両立がより重要視されるようになっている．

音声データは，Common Voice [4] のような単一話者による読み上げ音声をはじめ，複数話者が登場する対話音声データがあり，対話音声データの保護には，従来の一発話・一話者に対する音声仮名化とは異なる課題がある．対話音声では，従来の音声仮名化手法をそのまま適用すると，誰がどの部分を話したのかという情報が損失したり，仮名化後に一部話者の声色が似通うことで聞き分けがしづらくなったりと，対話の構造把握および分析に悪影響を及ぼす可能性がある．このような対話音声に対する仮名化は，話者ごとに異なる声色へ変換するだけでなく，各話者に割り当てられる擬似話者の声色を特徴量の類似度に基づき選択することで，話者識別性も担保する手法 [5] が検討され始めている．

2.2 自然言語処理におけるテキストデータの仮名化

自然言語処理の分野においては，テキストデータに含まれる個人情報を保護しつつ，機械学習に必要なデータの有用性を維持するための仮名化手法が多数提案されてきた．例えば，機械学習の学習データとして利用想定のテキストデータについて，テキスト中の個人識別情報（Personally Identifiable Information: PII）のうち，人名・地名・組織名などの固有表現に対し，Sanitized（例：PERSON_1 などの汎用表現への置換）や Pseudonymized（同一属性内で異なる名称への置換）といった手法を用いて，仮名化処理後のデータで学習したモデルの性能への影響を評価した研究がある [6]．また，医療や司法といった個人情報が密接に関わるセンシティブな領域では，新たなサービスの開発や法的安全性を満たすうえで個々の事例が非常に参考になるのに対し，個人情報や秘匿情報への配慮が不可欠であり，データの仮名化方法や仮名化済みデータの質を検証する研究が行われている．医療分野では，スウェーデンの clinical BERT モデルを用い，5 つの自然言語処理タスクに対してファインチューニングを施した上で，学習データの仮名化による予測性能の変化を分析した研究 [7] がある．この研究は，プライバシーリスク低減のために仮名化処理を施しても，学習に必要なデータの有用性を保持しつつ，個人情報の保護が可能であることを示している．さらに，司法分野においては，民事判決文の公開を可能にするための仮名化技術の概要や各国の取り組みが紹介し，判例データを機械的に仮名化した際の固有表現抽出に関する問題（仮名化の過不足など）を，属性別に分析する研究 [8] が報告されている．

3 提案手法

提案手法の概要を図 1 に示す．提案手法では，対話音声データに対し，各話者の態度・意図に関するパラ言語情報は残しつつ，個人同定に繋がる可能性の高い発話内容を仮名化し，仮名化語句の音声合成と発話全体の擬似話者の声による声色変換から，声色・発話内容双方の仮名化を音声データ上で実現する．

手順 1．発話内容の仮名化
はじめに，入力音声に対して音声認識や文字起こしを行い，発話内容を取得する．得られた発話内容に，大規模言語モデルを用いて仮名化処理を施す．本稿では，従来研究 [6] に倣い，PII として人名・地名・組織名を固有表現として扱う．合成音声による置換時に，元の単語とモーラ数が大きく異なると，元の発話に含まれていた話速やスペクトル包絡といったパラ言語的特徴を反映できず，発話の自然さが損なわれる可能性がある．そのため，仮名化を行う際は，仮名化前の単語と等しいモーラ数の単語で置換するようにする．これにより，仮名化した音声を元の発話に組み合わせた際の全体の発話長や話し方への影響を抑える．

手順 2．仮名化部分の無音マスキング
次に，仮名化対象となる語句に対応する音声区間を無音でマスキングする．手順 1 では，テキスト上での仮名化を行った．これに対し，置換した語句が入力音声上のどの部分にあたるか，音声とテキストのアライメントを取り，該当する区間を無音でマスキングする．

手順 3．仮名化語句の音声合成
手順 1 で仮名化した語群を後段の処理で元の発話と組み合わせられるよう，音声合成を行う．ここで，仮名化した語句を含む一文全体を音声合成すると，話者の話し方や態度・意図といったパラ言語情報が損なわれるため，この処理では仮名化した語句のみを音声合成する．

手順 4．元の発話と仮名化合成音声の接続
手順 2 でマスキングした仮名化区間に対し，手順 3 で合成した音声を組み合わせる．

手順 5．擬似話者の声による声色変換
手順 4 までで生成された音声は，声色が元の話者のままであるため，話者同定に繋がる情報が残っているといえる．そのため，従来と同様，手順 4 で生成された音声に対し，擬似話者の声色特徴量を用いて声色変換を行う．これにより，手順 1 から 4 の音声上における発話内容の仮名化に加え，声色に含まれる個人同定に繋がる情報を秘匿する．

4 評価

本稿では，生成された仮名化済み音声に対し，話者同定に繋がるリスクが軽減され話者のプライバシーが保護されているか，そして対話音声データとして重要な態度に関わる情報が仮名化後も残存しているかの二点を評価する．前者については，従来研究と同様に，話者照合モデルの等価誤り率 (Equal Error Rate: EER) が仮名化前よりも高いほど，話者の特定が困難，すなわち話者の情報が仮名化されていると判断する．後者の態度に関する情報の残存度合については，仮名化前後で音声を入力とする態度認識結果の差分の有無で評価をする．差分が小さいほど，仮名化前と変わらない精度で仮名化後の対話音声データを用いて態度に関する分析が可能だと判断し，提案手法による仮名化済みデータの有用性を認める．

5 実験条件

本実験では，RWCP 会議音声データベース [9] を対象に，声色・発話内容の仮名化を行った．RWCP は，3 名以上の複数の参加者が，旅行会社における企画会議，お客様との意見交換会について会議形式で話し合う様子を収録したデータである．発話内容の仮名化では，GPT-4o [10] を使用し，プロンプトでは人名・組織名・地名に対しモーラ数の等しい単語電仮名化を指示した．この

GPT-4o への仮名化指示に関するプロンプトとその出力例

プロンプト:
発話内容を示す「かな漢字記述」列において、人名・組織名・地名に対して仮名化を行ってください。仮名化の際には、元の単語と同じモーラ数の自然な仮名語に置換してください。
また、発話内で同一の対象を指している語については、仮名化後も一貫性が保たれるよう、同一の仮名語に変換してください。
なお、話の流れ上、元の単語の要素 (「大阪」→関西であること等) であることが重要な場合には、文脈に応じてその要素を反映させた仮名化を施してください。

元の発話 1: ［え］西鉄旅行四十周年記念の、キャンペーンを

仮名化発話 1: ［え］関空観光四十周年記念の、キャンペーンを

元の発話 2: 今度，大阪に、ユニバーサルスタジオ、できる訳ですから、

仮名化発話 2: 今度，神戸に、ユニバーサルスタジオ、できる訳ですから、

(事実と異なり，仮名化する必要がないため，人手によって「神戸」を「大阪」に修正)

際，出力されたデータの数点は，事実と異なる変換や文脈に沿わない仮名化となっていたため，このような語句は人手によってチェックし修正した．仮名化語句の音声合成には XTTS[1] [11] を，声色変換には Seed-VC[12] を使用した．声色変換時には，擬似話者の声色を用意するため x-vector を複数準備する必要がある．本実験では Common Voice 21.0 の内，20 発話以上登録されている話者 2418 名を対象に x-vector を抽出し，それらの中からランダムに 20 個を選択・平均して生成される x-vector を擬似話者の声色特徴量とした．そして，擬似話者の声色特徴量それぞれで適度な長さの発話 (本実験では「おはようございます」) を音声合成によって得られる擬似話者の発話データを用いて，Seed-VC で声色変換を行った．仮名化対象の音声データ上にて，どの区間が GPT-4o で仮名化された語句か検出するため，本実験では Montreal Forced Aligner [13] を用いて入力音声と仮名化対象語句のアライメントを取った．仮名化語句は単に音声合成するのではなく，WORLD [14] を用いて元の発話から抽出した Fo（基本周波数），スペクトル包絡，非周期成分，話速を模倣して反映することで，元発話に含まれるパラ言語的特徴が大きく損なわれないよう配慮した．アライメント結果に基づき，元音声の仮名化対象区間を無音でマスキングし，合成音声を元音声と正規化処理を行った後に結合することで，自然な仮名化発話を生成した．評価では，仮名化性能評価のために CNN ベースの話者照合モデル[2]を，態度認識の性能評価のために mimi AIR [3] を使用した．話者照合モデルは，Common Voice 日本語版 ver.19.0 の validated セットから条件を満たす約 720 話者・43,000 発話を抽出して学習に用いた．評価では各話者につき登録・テスト各 5 発話を用い，仮名化前後の音声で EER を算出した．mimi AIR では，肯定・平叙，否定，考え中，疑問の 4 項目に対し，それぞれスコアを出力する．今回の検証では，これら 4 項目のスコアをベクトルとして捉え，発話内容も含め仮名化された音声データそれぞれに対し，仮名化前後のベクトルのコサイン類似度を算出し，各音声データから得られるコサイン類似度の平均をとった．仮名化前後のベクトル間の類似度が高いほど，仮名化による態度認識に関わる情報の欠落が少ないとみなした．

[1] https://huggingface.co/coqui/XTTS-v2

[2] https://github.com/sp-au-mu-nl/SpeechComm/blob/main/notebook/chap10_speaker_verification.ipynb

[3] https://mimi.fairydevices.jp/technology/cloud/air/

表 1. 仮名化前後の等価誤り率 [%]

	元の音声	仮名化音声
EER	2.45	評価不能

6 結果

GPT-4o による発話内容の仮名化について，プロンプト上でモーラ数が保たれるよう指示をしたが，一部の仮名化語句は，元の単語と比較しモーラ数が 1 増減しているケースがあった．

仮名化前後の話者照合モデルの EER を表 1 に示す．仮名化後の音声では、登録・照合に用いる発話間で話者情報が一致せず、本人拒否率と他人受入率の交点が存在しない状態となった。このため、EER の閾値が無限大（∞）となり、EER は 0.00% と出力されたが、これは話者照合が全く成立しないことを意味する。本稿ではこの状態を「評価不能」と記述した。

MIMI AIR を用いて得られる態度を表す 4 項目の数値ベクトルについて，仮名化前後のコサイン類似度は 0.937 という値を得た．また，仮名化の一例として，RWCP に含まれる話者 f02 の音声「特に東京とかで、出発んなったら、関東近辺で」を仮名化し,「特に名古屋とかで、出発んなったら、関東近辺で」という発話内容に仮名化した際の声色・発話内容を仮名化した音声データの仮名化前後のスペクトログラムを図 2，態度認識の 4 項目のスコアを表 2 に示す．

仮名化前後で，態度認識 4 項目の値が大幅に増減することはなかった．ただし，一部仮名化音声には，仮名化後，態度認識 4 項目のうちの「否定」や「考え中」のスコアが増加し，他 2 項目が減少するケースが見られた．

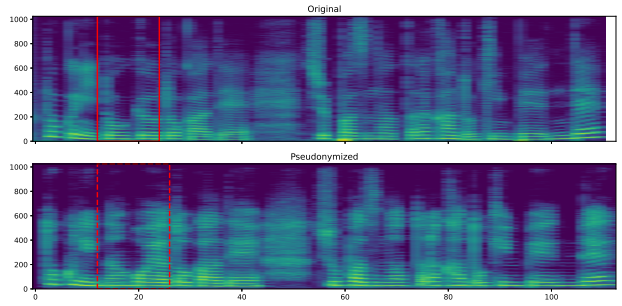

図 2. RWCP 話者 f02 の音声仮名化前後のスペクトログラム．赤枠は元の発話において仮名化対象となった「東京」の発話区間を，赤の破線枠は仮名化済み発話における仮名化した区間「名古屋」を示す．

表 2. RWCP 話者 f02 の音声仮名化前後の態度認識結果

	肯定・平叙	否定	考え中	疑問
仮名化前	0.00848	0.0261	$5.26e-5$	0.965
仮名化後	0.000399	0.00711	$3.98e-6$	0.992

6.1 考察

発話内容を仮名化する際，音声に表れている態度・意図といった情報が損なわれないよう，元の単語のモーラ数に基づく単語に置換したことで，元の音声データに表れているスペクトル包絡や話速を揃えることができた．これにより，態度認識モデルの 4 項目における仮名化前後のコサイン類似度の値，および表 2 の結果に見られるように，提案手法による仮名化で態度の情報への影響が抑えられた．一方，一部の仮名化音声に見られた「否定」「考え中」のスコアが上昇した原因について，これは元の単語よりもモーラ数の多い単語や元の単語にはない濁音ありの単語に仮名化された場合にスコアの上昇が見られた．今回の実験では，仮名化合成音声を元の発話の仮名化区間に結合する際，モーラ数が元の単語よりも多い仮名化語句の場合は，無音区間をわずかに足してから合成するといった微調整を機械的に行った．そのため，元の単語と仮名化語句にモーラ数のずれがあると，わずかだが結合時に音の切れ目が発生する場合があるため，新たに生まれた無音区間がポーズとして認識され，「考え中」のスコアが上昇した可能性がある．また，元の単語に含まれていなかった濁音が仮名化語句に含まれると，音声の抑揚や発音の強さが変化し，否定のように聞こえる可能性があるため，「否定」のスコアに影響を及ぼしたといえる．

7 おわりに

本研究では，対話音声データに対し声色および発話内容に対する仮名化手法を提案した．提案手法では，音声に見られる話者の態度情報が損なわれないよう，発話内容の仮名化時に元々の単語とモーラ数が一致する単語に置換するといった制限を設けることで，元発話のパラ言語情報を仮名化語句に反映することができ，その結果話者の情報を秘匿しつつ，態度認識の分析に利用可能な情報を保持できることを示した．今後の課題として，仮名化によって置換される語句の滑らかな結合方法の提案が挙げられる．

参考文献

[1] 伝康晴，榎本美香． 千葉大学 3 人会話コーパス（chiba3party），July 2014.

[2] Andreas Nautsch, et al. The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding. In *Proceedings of Interspeech*, pp. 3695–3699, September 2019.

[3] Fuming Fang, et al. Speaker anonymization using x-vector and neural waveform models. In *10th ISCA Workshop on Speech Synthesis (SSW10)*, pp. 155–160, 2019.

[4] Rosana Ardila, et al. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[5] 伊藤葵, 伊藤克亘. 話者のプライバシー保護下における対話音声データ活用のための音声仮名化. 情報処理学会第 87 回全国大会講演論文集, pp. 3–557–3–558, March 2025.

[6] Oleksandr Yermilov, et al. Privacy- and utility-preserving nlp with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pp. 232–241, Toronto, Canada, July 2023.

[7] Tahereh 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, June 2024.

[8] 久本空海ほか. 民事判決のオープンデータ化へ向けた機械処理による判例仮名化の検証. 言語処理学会第 28 回年次大会発表論文集, pp. 1406–1410, March 2022.

[9] RWCP (Real World Computing Partnership) 知的資源 WG. Rwcp 会議音声データベース（rwcp-sp01），September 2006.

[10] Aaron Hurst, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[11] Edresson Casanova, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

[12] Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

[13] Michael McAuliffe, et al. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pp. 498–502, 2017.

[14] Masanori Morise, et al. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884, 2016.