2ZE-08

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*

伊藤 葵†, 伊藤 克亘‡

1 はじめに

音声データには, パラ言語情報や発話内容といった有用な情報が含まれる. 特に対話音声データに対し, 音声認識や話者分離, 感情認識といった音声処理を施すことで, 音声データからの自動議事録作成やコールセンターでの感情分析が可能となる. これらのタスクは, 機械学習や信号処理を用いた手法が多く提案されており, 特に音声情報処理に関する機械学習においては, 学習データや検証データの不足が問題となる. そこで, 様々な研究機関が機械学習・音声分析用に収録したデータをクラウド上で公開し, 世界中の人が使用できるようにしている. しかし, CommonVoice[1] などのコーパスは生データのまま公開されており, 悪意のある第三者によるプライバシー侵害の恐れ (コーパスに含まれる音声を利用した話者推定, ディープフェイクの生成など) が残存している. そのため, データの暗号化のみならず, 音声データ自体に対し, 個人同定に繋がる可能性のある情報を適切に処理することで, 音声データの有用性を保ちつつ話者のプライバシーが保護されることが望ましい. 音声仮名化は, 声色や韻律情報を加工することで, 音声から読み取れる有益な情報を保持しつつ, 話者のプライバシーを保護する手段の一つである. 従来, 音声仮名化では一発話もしくは話者単位での処理がされてきた. けれども, 話者分離や感情分析で求められる対話音声データでは, ターンテイキングの情報や複数人のやり取りで発生するパラ言語情報が重要であり, 一発話単位や話者間の関係性を無視した音声仮名化では, 同一話者による発話の繋がりが損失され, 対話内容及び対話データに含まれるパラ言語情報を失う.

本稿では, 対話音声データを想定した複数話者間の声色の関係性を考慮する日本語音声仮名化を提案する. 対話音声データを仮名化する場合, 話者毎の特徴を秘匿しつつ, 各話者の発話内容や複数人でのやり取りで発生するパラ言語情報を分析できるよう話者の区別が出来なければならない. この話者区別の可不可を, 話者毎の発話区間を推定する話者ダイアライゼーションを用いて評価する.

2 関連研究

2.1 声色の特徴量変換による音声匿名化

音声の特徴量抽出の一つに, メル周波数ケプストラム係数を入力とする話者認識モデルに対し, 分類層の直前を特徴量として抽出する x-vector[8] がある. この x-vector を変換することで, 話者の同定に繋がる情報の一つである声色情報の秘匿を試みたのが [2] である.

この手法では, 入力音声に対し, 個人同定に繋がる話者特徴量 x-vector と, 匿名化後も有益な情報として保持しておきたい発話内容および基本周波数 (Fo) を抽出する. この内 x-vector は, 新たに生成される擬似話者 x-vector と置換される. 擬似話者 x-vector は, 別途プールした x-vector 群の中からランダムに抽出された任意の数の x-vector を平均して得られる 512 次元のベクトルである. 置換した擬似話者 x-vector と, 元の入力音声から抽出した発話内容および Fo を音声合成することで, 声色が変換された匿名化音声を得られる.

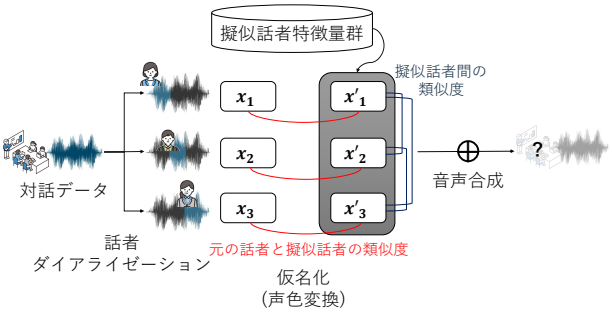

図 1: 提案手法の概要. 元の話者と擬似話者間, 擬似話者間同士の類似度を考慮する.

2.2 話者ダイアライゼーション

話者ダイアライゼーションとは, 「誰が, いつ」話したかを識別する音声処理タスクである [5]. 深層学習の登場をうけ, 話者ダイアライゼーションでは深層学習を用いて取得可能な高次元の話者特徴量を利用した手法が提案されるようになった. 話者ダイアライゼーションでは, まず発話区間を推定する. 次に, 推定された発話区間に対し, セグメント化を施す. 得られた各セグメントに対して特徴量抽出を行い, 特徴量をクラスタリングすることで, 音声内の類似領域をグループ化する. クラスタリングの結果より, 入力音声内について誰が, どの区間で話しているかを推定する.

3 提案手法

本稿では, 一発話または一話者を対象としていた従来の音声仮名化プロセスを, 複数人が登場する対話音声データ用に拡張する. 対話音声データを仮名化する場合, 各話者の声色を本来の話者とかけ離れた特徴に変換しつつ, 話者同士の声が聞き分けられるよう, 声色変換後の空間において各擬似話者の声色特徴量が離れた位置に分布するよう擬似話者特徴量を選択する必要がある. まず, 入力対話音声データに対し, 話者ダイアライゼーションを用いて各話者の発話区間を推定する. 次に, 話者毎に声色の音声仮名化 [2] を施す. ここで, 各話者の元の音声から抽出された声色特徴量をそれぞれ $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}$, 各話者の仮名化後の声色となる擬似話者の声色特徴量をそれぞれ $\boldsymbol{x'_1}, \boldsymbol{x'_2}, ..., \boldsymbol{x'_n}$ と表現する. 元の話者の声色特徴とのコサイン類似度の最小化を目指しつつ, 擬似話者同士のコサイン類似度も最小となるよう $x'_1, x'_2, ..., x'_n$ を選ぶとし, 目的関数 L を式 1 のように定義する.

$$L = \sum_{i=1}^{n} \left(1 - \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}'_i}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}'_i\|}\right) + \lambda \sum_{1 \le i < j \le n} \left(1 - \frac{\boldsymbol{x}'_i \cdot \boldsymbol{x}'_j}{\|\boldsymbol{x}'_i\|\|\boldsymbol{x}'_j\|}\right) \tag{1}$$

λ は重みづけパラメーター, n は話者数である. 特に,

*Speech Pseudonymization for Utilizing Dialogue Speech Data While Protecting Speaker Privacy, *Aoi Ito (Hosei Univ.) et al.*

†法政大学大学院情報科学研究科
‡法政大学情報科学部

λ の変化によって, 元の話者および擬似話者間の距離を優先して考慮するか, 擬似話者同士の区別を強調する方を優先させるか調整できる. この目的関数 L を満たす擬似話者声色特徴量を, 別途プールした擬似話者用声色特徴量群の中から選択する. 擬似話者用声色特徴量群は, 入力音声とは関係のないコーパスから抽出された複数人の声色特徴量に対し, 任意の数の特徴量を選択・平均して得られる特徴量の集合である. 対話音声データに登場する話者数分の擬似話者声色特徴量を選択した後, 各話者の音声データに対し, 元の音声から抽出した発話内容・Fo と合わせて音声合成する. 最後に, その他の話者の仮名化音声と組み合わせることで, 対話音声データを仮名化する.

ここで, 複数話者間を考慮した音声匿名化手法として [4] がある. [4] では, 声色特徴量として x-vector を対象に, 複数人の話者間の関係性を考慮した擬似 x-vector の選択を検討しているが, あらかじめ対話音声データに表れる話者数を定義する必要がある. 一方, 本稿での提案手法は, 擬似話者間の声色特徴量を検討する際, 過去に選択した声色特徴量の情報を保持・参照する. そのため, 新たに擬似話者特徴量を選択する際は, これまで選ばれてきた擬似話者の声色と比較した際に類似していない特徴量を選択するため, 計算効率を無視すると理論上では無限人まで拡張可能である.

4 実験条件

4.1 データセット

本実験では, RWCP 会議音声データベース [7] を使用した. このコーパスは, 4 名以上が参加する模擬会議の音声データである. 参加者の職業に関係した, 企画・立案に関するテーマを設定しており, 話者ごとの接話マイクと全体用のマイクで同時録音している.

4.2 評価手法

本実験では, 仮名化した対話音声データに対し, 音声の有用性が残存しているかを評価した. 音声の有用性を確認するため, 本実験ではダイアライゼーション誤り率 (Diarization Error Rate: DER) を使用する. ダイアライゼーション誤り率は, 音声を非音声として分類した誤り率, 非音声を音声として分類した誤報率, ある話者の音声を別の話者の音声と混同した話者誤り率の合計から算出される値である. 今回対象としている対話音声データでは, 仮名化後も対話のやり取りが追えるよう話者別の発話区間が正しく推定できることが望ましい. そのため, ダイアライゼーション誤り率が低いほど, 仮名化後も対話音声データの有用性が保持できていると判断する. 話者ダイアライゼーション並びに誤り率の算出は, pyannote-audio[6] を利用した. そして, 仮名化の性能を比較するため, 一つの会議音声を対象に, 発言している 4 名の各発話の x-vector の分布を仮名化前後で可視化し, 話者の特徴量がどのように変化したかを分析する.

5 結果

元の音声, ならびに仮名化後の音声データそれぞれにおけるダイアライゼーション誤り率を表 1 に示す.

仮名化前後の x-vector の分布を Umap[3] を用いて可視化した図を図 2 に示す. 点同士の距離が近いほど, 似ている特徴量であることを示す.

話者 1 と 4 について, 図 2 左側の分布をみると, 元々は似た声質の話者同士であったことが分かる. 提案手法により, 図 2 右側に示すように, 擬似話者の特徴量空間上に置いて離れた位置に分布する, すなわち元の音声より類似度の小さい声色のペアになったため, 対話が追いやすくなったことが分かる. 一方, 図 2 右側の分布に示されているように, 話者 2,3 の声は近い声色に変換されており, これが表 1 に示す DER の劣化に繋がったと考えられる. 実際の音声を聞くと, 話者 2,3 の声色は同程度の低い声であった. 提案手法のアルゴリズムより, 擬似話者として選択される特徴量は類似度を考慮して選ばれるため, この問題は, 特徴量選択以前の箇所, すなわち擬似話者特徴量群の声色の偏りが原因と考えられる.

表 1: 元の対話音声と仮名化後の対話音声におけるダイアライゼーション誤り率

	DER ↓
Original	0.2200
Pseudonymized	0.6882

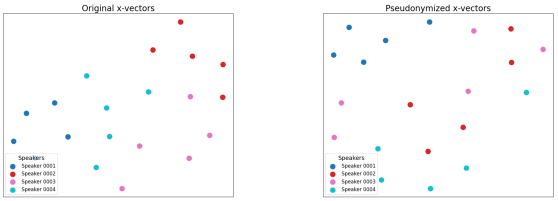

図 2: 1 会議音声 (話者 4 人) の x-vector の分布. 左: 仮名化前, 右: 仮名化後.

6 おわりに

本稿では, 対話音声データのための音声仮名化を提案した. 提案手法では, 入力音声と仮名化音声の声色特徴量同士の距離だけでなく, 仮名化後の空間における複数話者の特徴量同士の距離も考慮することで, 仮名化後も対話として聞き取り可能な音声データを生成した.

参考文献

[1] R. Ardila et al. Common voice: A massively-multilingual speech corpus. In *Proc. 12th Lang. Resources Eval. Conf. (LREC)*, pages 4218–4222, Marseille, France, May 2020. European Lang. Resources Assoc.

[2] F. Fang et al. Speaker anonymization using x-vector and neural waveform models. pages 155–160, 09 2019.

[3] L. McInnes et al. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018.

[4] X. Miao et al. A benchmark for multi-speaker anonymization. *arXiv preprint arXiv:2407.05608*, 2024.

[5] T. J. Park et al. A review of speaker diarization: Recent advances with deep learning. *Comput. Speech Lang.*, 72:101317, 2022.

[6] A. Plaquet et al. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

[7] RWCP 知的資源 WG. RWCP 会議音声データベース（RWCP-SP01）, sep 2006.

[8] D. Snyder et al. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 5329–5333, 2018.