CE-006

モーラ長に基づく話者モデルを用いた音声仮名化

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

1　はじめに

音声情報処理の分野では, 音声認識や話者照合, 音声からの感情認識といった技術が高度になるにつれ, それらの技術を活用したサービスの提供が盛んとなってきている. スマートスピーカーをはじめ, このような各ユーザーの音声を処理するようなシステムは企業や研究機関だけでなく, 一般家庭にも広まっており, ユーザーの増加に伴い, 音声情報を用いたサービス・メディアにおけるプライバシー保護に重点が置かれ始めている. 例えば, Facebook ではマイクを使って収集された周囲の音を解析するサービスを提供しており, Apple が提供する Siri [1] では, ユーザーの Siri への信頼性向上のため, グレーディング (ユーザーが Siri へ入力した発話とそれに対して Siri が生成した書き起こしを見直し Siri の性能を評価する) といった作業が過去に実施されていた (現在は, ユーザーが Siri の性能改善のため, 自身の声を提供してよいか選択可能になっている). このように, サービスの性能向上のためには多くの人による発話サンプルが必要である反面, データ収集の困難さといった問題や, ユーザーの特定に繋がるといったプライバシー侵害のリスクが付き纏う. さらに, EU 一般データ保護規則 (General Data Protection Regulation) [2] をはじめとし, 各国が個人データやプライバシーの保護を厳格に規定し, 従来よりデータの扱いに注意を払っている.

このようなプライバシー保護の観点から, 音声情報処理の分野においては音声匿名化・音声仮名化といった発話者の保護に重点を置いた研究が近年進められている. Voice Privacy Challenge [3] は, 音声匿名化など音声技術のプライバシー保護に関する手法の開発・取り組みを先導しており, 多くの研究者が競争的に性能の高い音声匿名化手法を提案している. ベースラインシステムの形式で主な音声匿名化手法が公開されており, その中には `x-vector` [4] を用いた声色変換による話者の秘匿化手法 [5] がある. この手法により, 話者照合モデルの誤り率が上昇しており, 話者の秘匿化を実現している. この手法では, `x-vector` の変換, すなわち音声に含まれる声色の要素のみ秘匿化しているといえる. つまり, その他の話し方 (韻律, イントネーション, 方言など) の要素は考慮されておらず, 聞いた人あるいは `x-vector` を用いていない話者照合モデルに, 声色以外の要因から発話者を推定される可能性がある.

本稿では, 声色以外に個人性が表れる要素として, 日本語発話者にとっての言語単位である「モーラ」に着目する. そして, 平均モーラ長と関係のあるドメイン「方言」の要素の秘匿化を目指し, モーラ長の変換による音声仮名化手法を提案する.

*法政大学大学院 情報科学研究科
†法政大学 情報科学部

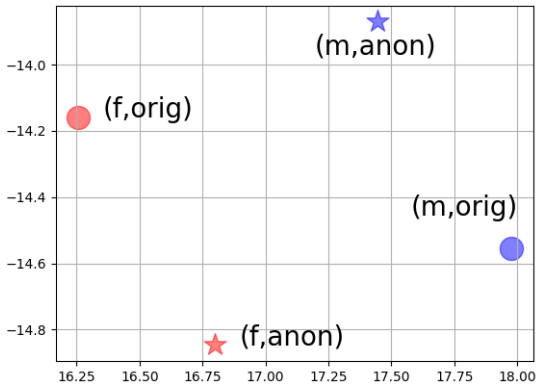

図 1. x-vector を用いた匿名化前後の x-vector の分布図. (●: 元の音声, ★: 匿名化後の音声, m: 男性, f: 女性)

2　従来手法

音声情報処理分野では, 音声匿名化・音声仮名化といった話者の秘匿化に焦点を当てた研究が進められてきている. 様々な手法があるが, ここでは Voice Privacy Challenge でベースラインシステムの一つとなっている `x-vector` [4] を用いた声色変換による手法について述べる. また, 本稿で着目した日本語発話において重要なモーラに関する従来研究についてまとめる.

2.1　x-vector を用いた音声匿名化

Fang らは, 入力された音声の `x-vector` を別に生成した `x-vector` と置換, 合成することで声色の変換を行い話者の秘匿化を実現した. この手法では, まず, 入力音声から F0(基本周波数), 音声認識の結果 (発話内容), `x-vector` をそれぞれ抽出する. ここで, 抽出した `x-vector` $\mathbf{x}_{\text{orig}}$ を別の `x-vector` $\mathbf{x}_{\text{new}}$ と置き換える. 置換先の `x-vector` $\mathbf{x}_{\text{new}}$ は, 複数話者から抽出した `x-vector` 群 $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ から任意の数の `x-vector` を選択しそれらの平均を取ったものである. 最後に, 80 次元のメルスペクトログラムと初めに抽出した F0, そして新たに生成した `x-vector` $\mathbf{x}_{\text{new}}$ を用いて音声合成をする. 実験で使用されているコーパスは VoxCeleb [6] であり, 発話の言語は英語である. 図 1 に Fang らの提案手法による匿名化前後の `x-vector` の分布を, 図 2 に男性の音声に対する匿名化前後のスペクトログラムを示す. Fang らが提案したこの手法は, 話者の秘匿化を実現しつつ, 後段の音声認識タスクの性能を考慮することで元の音声と同様な自然な音声を目指している. 図 1 からは, 匿名化前後で話者の声色に関する特徴量が変わっていることが分かる. 図 2 を見ると, 音声合成後に話速の変動はないことが見受けられる. これは, 提案手法内で, 話速に対して何らかの処理をしていないこと

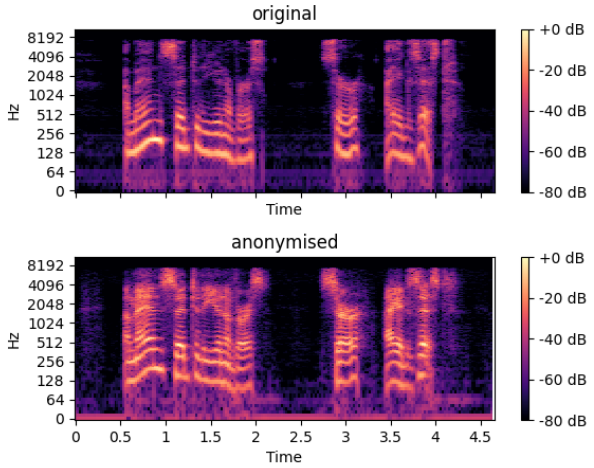

図 2. x-vector を用いた音声匿名化を適用した際のスペクトログラム. (上: 元の音声, 下: 匿名化後の音声)

からも自明であり, このことが後段の音声認識タスクの性能や音声自体の自然性に対し影響を与えていると考えられる反面, 話速から発話者を推定することを容易にしているとも捉えられる.

2.2 日本語におけるモーラ

本稿では, 従来の研究で扱われてきた英語ではなく, 日本語に焦点を当てている. 日本語には, 「モーラ」と呼ばれる言語単位が存在している. 自然言語は, ある言語内で使用される最小の韻律単位に応じて, 「モーラ言語」と「音節言語」の 2 つのグループに分類されており [7], 日本語はモーラを基調とする言語という議論がある. そして, 窪薗ら [8, 9] によると, 日本語におけるモーラは大別すると (1) 時間制御の単位 (2) 音韻的長さを測る単位 (3) 発話産出の分節単位 (4) 発話知覚の分節単位といった役割がある. モーラに関する研究 [10] では, 母音の継続長がモーラ数の知覚に影響を与えることが判明している. また, 日本語母語話者かそうでないかによって, 単語内の平均モーラ長に違う傾向が見られることが分かっている. 賈ら [11] によると, 日本語母語話者は, 単語内の平均モーラ長と促音長・長音長の間に直線関係があるのに対し, 中国語母語話者には直線からの変動及び個人差も大きく表れることが分かっている. これらのことより, モーラの継続長と個人性には関係性があることが分かる. そして, 堀ら [12] は, 東京方言と関西方言の話者に対し, 平均モーラ長は東京方言話者の方が有意に短いということを示している.

3 提案手法

本稿では, モーラ長に着目した音声仮名化手法のための特徴量抽出を提案する. 従来手法では, 話者と声色の情報が紐づいた特徴量である `x-vector` を変換することにより, 匿名化が行われていた. 提案手法では, 日本語において重要な言語単位であるモーラの継続長と各話者の情報 (ここでは方言) を紐づけたモデルを学習し, 本来の発話者の傾向とは異なるモーラ継続長を故意に出力させることで, モーラ長における音声仮名化を試みる. 提案手法の概要を図 3 に示す.

3.1 モーラ長に基づく話者の方言識別モデル

各話者のモーラ長に基づき, 発話者の方言を識別するモデルを学習する. 学習方法は, 藤田ら [13] が考案した方法に倣う. まず, 発話内の各モーラの継続長を示すベクトルを入力する. 入力されたモーラ長の情報は, Transformer Encoder [14] でその特徴量が抽出される. ここで, モーラは時間制御をしているものであり, 日本語発話内での拍の役割を担っている. そのため, モーラの個人性は発話全体に表れると考え, 個々人が持つモーラの傾向と発話に含まれる各モーラの継続長を学習するため, Attention 機構を持ち他のモーラとの依存関係を発話全体にわたって学習できると推測される Transformer を採用し, その Encoder を利用した. そして, 抽出された特徴量に基づき, 入力話者が東京方言もしくは関西方言のどちらであるかを判定するよう学習する.

3.2 話者別モーラ長出力モデルの学習

つづいて, 3.1 節で述べた話者の方言識別モデルを用いて抽出された 256 次元の特徴量と発話内容 (モーラ単位) を入力とし, 本来の話者の方言 (東京もしくは関西) に適したモーラ長を出力するモデルを学習する. まず, 発話内容に対して, モーラ単位でアライメントを取得する. モーラ区切りのルールとして, 捨て仮名は直前の仮名と一組で一拍として数えられる. また, 音節とは異なり, 長音, 促音, 撥音は一拍としてカウントされる. 次に, 話者の方言識別モデルの分類層の直前の層を, 話者の方言とモーラ長間の情報を持つ特徴量として抽出する. 特徴量抽出方法は,`x-vector` に倣っており, `x-vector` は MFCC を用いて学習された話者識別モデルから抽出しているのに対し, ここでは, モーラ長を用いて学習された 3.1 節のモデルを使用している. この特徴量の次元は 256 次元である. モーラ単位のアライメントは, 組み合わされた形で Transformer Encoder へ入力される. 正解ラベルは, 入力された話者の方言特徴量と発話内容に対応した実際の発話におけるモーラ長とする.

3.3 話者別モーラ長出力モデルによる音声仮名化

音声の仮名化時には, 3.2 節で述べた話者の方言特徴量と発話内容に応じてモーラ長を出力するモデルを用いる. 3.1 節で提案した話者の方言識別モデルから, 方言に関するモーラ長特徴量群を作成する. そして元発話の音声からモーラ長単位のアライメントを取得し, 方言特徴量群から抽出して新たに生成した特徴量ベクトルとともに, 3.2 節のモーラ長出力モデルへ入力する. ここで, 新たに生成した特徴量とは, 方言特徴量群から任意の数の特徴量を抽出し, 平均して生成することとする. このように生成されたモーラ長は, 元の話者の方言情報を参照していないため, モーラ長に基づいた仮名化が行えると推定する.

4 評価

提案手法で用いるモーラ長に基づく話者の方言識別モデルに対し, 仮名化前後のモデルの性能を比較する. 方言識別モデルは, 方言判定に関する正答率で評価する. 提案手法により仮名化されたモーラ長を用いて上手く仮名化されているのであれば, 正答率が下がると予想される.

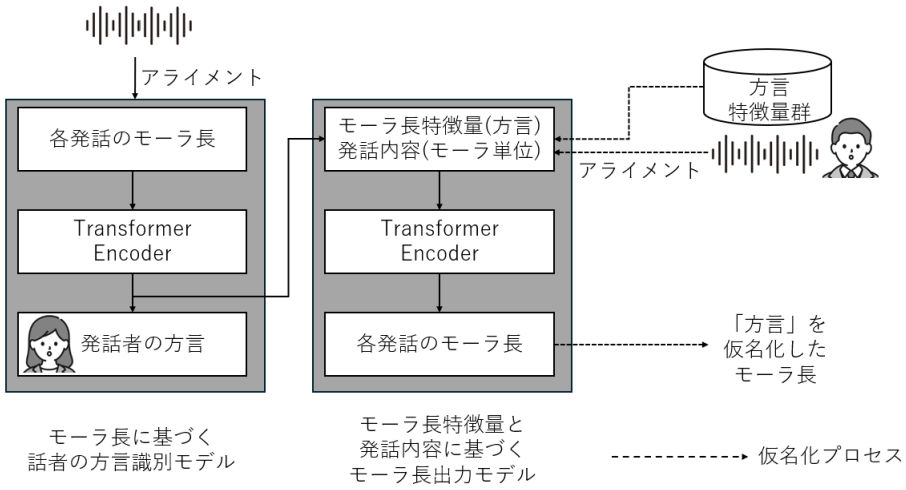

図 3. 提案手法の概要. モーラ長の変換により, 話速に基づく日本語音声の仮名化を目指す.

また, モーラ長に基づく話者の方言識別モデルから抽出した 256 次元の特徴量を Umap [15] にて次元削減を施し, その分布を可視化する. 可視化の結果から, 方言というドメインに基づく特徴量が正確に得られているか評価する.

4.1 使用データ

本実験では, Mozilla が提供しているコーパス CommonVoice 16 [16] の日本語データセットを使用した. このデータセットには, 一発話しか登録されていない話者や方言に関する情報が含まれていない話者が多く存在するため, 話者性の学習時には最低 40 発話含まれており, 話者の方言を申告している話者計 19 名を対象に実験を行った. データセットに登録されている方言に関するラベルは自由記述であり, 本実験では「東京式」「東京弁」のように答えている話者を東京方言, 「大阪弁」「関西弁」のように答えている話者を関西方言とした. アライメントの取得には, Montreal Forced Aligner [17] を用いた. モーラ単位に区切った継続長を取得するため, 母音の定義はアライメント取得時に使用した辞書 Japanese MFA dictionary に従った. 実験では, アライメントが取得できた計 929 発話を使い, その内 556 発話を訓練用, 186 発話を検証用, 187 発話をテスト用とした. また, 仮名化用に 186 個の特徴量を抽出し, 仮名化時にはそれらの中から無作為に摘出した特徴量 $n(n = 5, 10)$ 個を平均した特徴量を用いて, モーラ長の仮名化を行った.

4.2 実験条件

本実験で学習したモデルの各構造を表 1, 2 に示す. 話者の方言とモーラ長間の情報を持つ特徴量は, 表 1 に示すモデルの全結合層から抽出する. 学習率は 1×10^{-3} から 1×10^{-5} の範囲で学習をした結果, もっとも性能が高かった 1×10^{-4} を採用した. エポック数は, 300 である. 提案手法では, 仮名化の際, 特徴量群から無作為に特徴量を摘出するという過程があるため, 性能を測る際は乱数シードを 5 回変えた結果の平均を算出した. 方言識別モデルの損失関数には交差エントロピー誤差を, モーラ長出力モデルの損失関数には平均二乗誤差を使用した.

表 1. 話者の方言識別モデルの構造

Layer	Output Shape
TransformerEncoder	128
Fully-connected	256
Classification	2

表 2. モーラ長出力モデルの構造

Layer	Output Shape
TransformerEncoder	384
Fully-connected	256
Duration Output	128

5 結果

5.1 話者の方言識別モデルから抽出した特徴量

話者の方言識別モデルから抽出した 256 次元の特徴量の分布を図 4 に示す. 丸印が東京方言の話者から抽出された特徴量, バツ印が関西方言の話者から抽出された特徴量である. 次元削減の結果, 東京方言の話者と関西方言の話者それぞれの特徴量分布は大きく離れた箇所に位置することがわかる.

5.2 話者の方言識別モデルの性能結果

仮名化前後の話者の方言識別モデルの正解率を表 3 に示す. Original が話者の方言識別モデル本来の性能であり, 列 Pseudonymize は仮名化後の話者の方言識別モデルの性能を表す. 仮名化後は, モデルの正解率が落ちていることが分かる.

5.3 考察

図 4 より, 東京方言と関西方言それぞれの特徴量分布が離れていることから, 本実験で学習した話者の方言識別モデルは, 発話者の方言とモーラ長の情報を十分学習できており, 特徴量として抽出するには問題のない性能であるといえる. また表 3 より, 仮名化用に抽出する特徴量が多いほど仮名化性能が高くなると推測できる. た

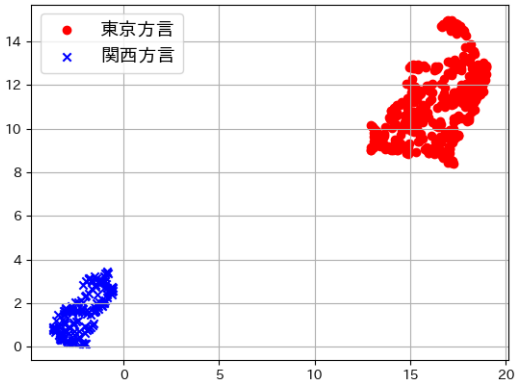

図 4. 話者の方言識別モデルから抽出した特徴量の分布図. ●が東京方言, ×が関西方言である.

表 3. 話者の方言識別モデルの正解率 (%). n は無作為に選択した特徴量の個数を示す.

Original	Pseudonymize(n=5)	Pseudonymize(n=10)
68.28	64.19	63.76

だし, 本実験で扱っているドメイン「方言」は, 東京と関西の 2 つしかラベルを持たないこと, 仮名化用の特徴量は無作為に選択し平均していることから, 選択された特徴量によっては, 仮名化後の継続長から推定される方言と, 本来の話者の方言が一致する可能性があるといえる.

6 おわりに

従来の `x-vector` を用いた声色変換による音声匿名化に対し, 本稿では日本語発話におけるモーラ長の変換に基づく音声仮名化手法を提案した. 提案手法では, まず話者の方言識別モデルを利用した方言とモーラ長の情報を結び付けた特徴量および発話内容を入力とし, Transformer Encoder を用いて方言に基づく適切なモーラ長を出力するモデルを学習した. そして, 元音声の話者とは異なる話者の傾向を持ったモーラ長を無作為に組み合わせた特徴量を入力することで生成し, 音声仮名化を目指した. 実験の結果, 話者の方言識別モデルの性能は 6.2%劣化し, 提案手法は方言というドメインにおける音声仮名化に繋がったといえる. 今後の課題として, より高性能な仮名化に向けた他ドメインに適応したモーラ長出力モデルの学習や方言に関する特徴量の使用法の検討が挙げられる.

参考文献

[1] Apple. Siri のプライバシー保護機能を強化, 2019. `https://www.apple.com/jp/newsroom/2019/08/improving-siris-privacy-protections/`.

[2] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

[3] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean-François Bonastre. The voiceprivacy 2022 challenge evaluation plan, 2022.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[5] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. Speaker anonymization using x-vector and neural waveform models, 2019.

[6] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *CoRR*, Vol. abs/1706.08612, , 2017.

[7] N.S. Trubetzkoy. *Principles of Phonology*. University of California Press, 1969.

[8] Haruo Kubozono. *Mora and Syllable*, chapter 2, pp. 31–61. John Wiley Sons, Ltd, 2017.

[9] 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.

[10] Kosuke Sugai. Mental Representation of Japanese Mora; Focusing on its Intrinsic Duration. In *Proc. Interspeech 2017*, pp. 2973–2977, 2017.

[11] 賈海平, 森大毅, 粕谷英樹. 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.

[12] 堀智子, 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.

[13] Kenichi FUJITA, Atsushi ANDO, and Yusuke IJIMA. Speech rhythm-based speaker embeddings extraction from phonemes and phoneme duration for multi-speaker speech synthesis. *IEICE Transactions on Information and Systems*, Vol. E107.D, No. 1, pp. 93–104, 2024.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, Vol. abs/1706.03762, , 2017.

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[16] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020.

[17] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, 2017.