5R-08

韻律特徴を考慮した音声仮名化 *

伊藤 葵 †,　伊藤 克亘 †

1　はじめに

近年, 音声認識や話者認識といった研究が進むにあたり, スマートスピーカーや自動議事録作成アプリなど, 音声に含まれる様々な情報 (発話内容, 話者, 感情) を活用したサービスが普及している. 音声内の情報を活用するためにも, 音声データをオリジナルのまま使用することが望ましい. しかし, 生の音声データは目的の情報 (言語情報など) 以外に, 話者自身の特徴に関する情報も保持しているため, プライバシーの侵害につながる恐れがある. General Data Protection Regulation でも音声情報の保護は着目されており, 今後音声を活用する場合は, 音声データ内の個人情報保護が不可欠である. 音声データに対するプライバシー保護について, VoicePrivacy Challenge [11] をはじめ, 音声匿名化 (仮名化)[2, 7] の研究が進められている. これらの音声匿名化 (仮名化) では, 発話の明瞭度を保ちつつ発話者の特定に繋がる情報の処理 (マスキング, 平均化, 除去など) が求められている. 発話の明瞭度を保つことで, 発話者のプライバシーを保護したまま, 音声データを音声認識などのシステムに活用できる.

本稿では, 従来研究において取り扱ってきた話者特徴量の埋込である x-vector[10] を用いた声色情報の変換に加え, 同じく発話者の特定の一助となる韻律情報を明示的に扱うことにより, マルチファクターに音声データを仮名化する手法を提案し, 概要を紹介する.

2　従来研究

2.1　x-vector を用いた音声匿名化

VoicePrivacy Challenge[11] で紹介されているベースラインシステムの一つに, x-vector を用いた手法 [2] がある.

はじめに, 発話から発話内容と話者識別に用いる話者特徴量 (x-vector) を抽出する. 抽出した x-vector に対し, [5] は事前訓練済みの変換モデルを用いて, 非識別化を試みた. これに対し [2] は, 一つの変換関数のみを学習し, かつ複数の x-vector をランダムに選択し平均を取ることで生成した疑似話者の特徴量を用いる匿名化を提案した.

[2] は, 音声の品質を保ったまま話者照合システムの等価誤り率 (Equal Error Rate: EER) を上昇, すなわち匿名化性能を向上させた.

2.2　発話リズムの埋込

[3] は, 音声合成において, 個人ごとの音素継続時間長のモデル化に適した話者埋め込み手法を提案した. 従来, 学習時に使用されてきた x-vector やメルスペクトログラムといった特徴量は, 発話リズムといった時間特徴量を明示的に扱っていない. そこで, [3] は, 音素とその継続時間長を用いて発話リズムを埋め込むことにより, 時間的特徴量に基づく話者埋め込みベクトルの生成手法を提案した. この時間的特徴量を用いることで, 音素継続時間長が似ている話者同士では, 類似した埋め込みベクトルを生成できるようになった.

3　提案手法

本稿では, 韻律情報を明に扱ったマルチファクターな音声仮名化を提案する. 提案手法の概要を図 1 に示す.

話者特徴量 x-vector を用いた仮名化は, [2] 同様, 入力音声 $\boldsymbol{X}$ から x-vector $\boldsymbol{xvec}$ を抽出し, 事前訓練済み x-vector のプールから複数の x-vector を選択・平均して生成された疑似話者の x-vector $\boldsymbol{xvec}_{\text{new}}$ と置換する.

韻律情報の埋め込み [3] を用いた仮名化手法として, はじめに, [3] で提案されている韻律情報の埋め込みプールを取得する. 音声認識モデルに訓練用音声データを入力し, 発話内容 $\boldsymbol{X}_{\text{train}}$ を取得する. 発話内容 $\boldsymbol{X}_{\text{train}}$ は, [2] と同様 x-vector による話者特徴量ならびに韻律情報の置換後, 音声合成にも使用する. アライメント $\boldsymbol{P}_{\text{train}} = [p_1, \ldots, p_T]$ は音声認識モデルで認識した発話内容 $\boldsymbol{X}$ を発話内容から各単語の読み, そして音素に変換した形で取得する. ここで, $p_t \in \mathbb{R}^{K+1}$ とし, K は音素の種類数とする. 取得したアライメント $\boldsymbol{P}_{\text{train}}$ は, 各音素の発話内での開始時刻とともに 1 次元のベクトルの形で, [3] にならい Transformer に入力する.

$$\boldsymbol{H} = \text{TransformerEncoder}(\boldsymbol{P}_{\text{train}};\ \theta) \qquad (1)$$

θ はモデルパラメータ, $\boldsymbol{H}$ は出力シーケンスである. この Transformer を, 話者照合モデルとして学習する. 入力音声に対し話者照合ができるよう訓練したモデルの全結合層から, 韻律情報の埋込として $\boldsymbol{P}_{\text{train\_embedding}}$ を取得し, プールする.

つづいて, 仮名化対象である音声について, 入力音声に対してアライメントを取得し, 韻律情報を加工する. 仮名化のため, $\boldsymbol{X}_{\text{train}}$ と同様, 仮名化対象である入力音声から認識した発話内容 $\boldsymbol{X}$ に対し, アライメント $\boldsymbol{P}$ を取得し, Transformer モデルの全結合層から入力音声の韻律埋め込み $\boldsymbol{P}_{\text{embedding}}$ を得る. そして, $\boldsymbol{P}_{\text{train\_embedding}}$ の一つ $\boldsymbol{P}_{\text{train\_new}}$ を $\boldsymbol{P}_{\text{embedding}}$ と置換する.

最後に, 取得した発話内容 $\boldsymbol{X}$ と置換後の $\boldsymbol{xvec}_{\text{new}}$, $\boldsymbol{P}_{\text{train\_new}}$ を合成し, 仮名化済み音声を取得する.

4　実験条件

4.1　データセット

本実験では, 入力に CommonVoice 14.0 日本語データセットを用いる. アライメントの単位は、日本語話し言葉コーパス (CSJ) で採用された音韻計 148 個とする. Transformer で学習する際は, 発話内に登場する音素を 3 つずつ組み合わせ, 音素セットに加え始めの音素の開始時刻を入力ベクトルとして渡す. x-vector

---

* Speaker Pseudonymization Considering Prosodic Features *Aoi Ito (Hosei Univ.) et al.*

† 法政大学情報科学部

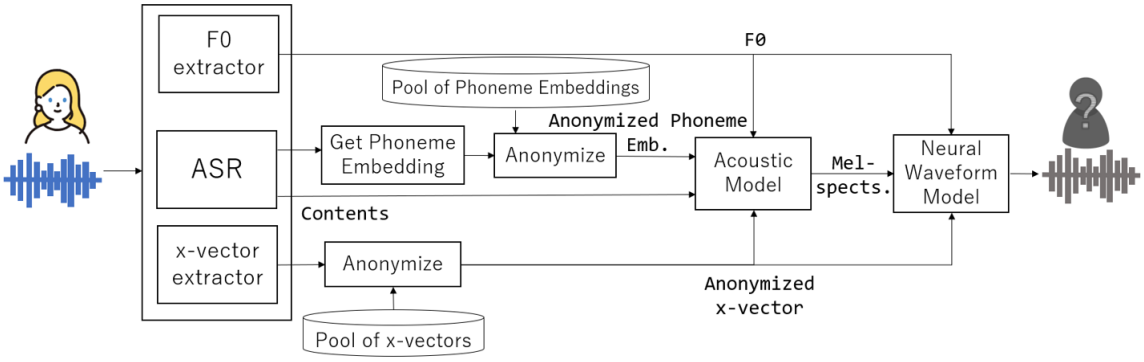

図 1. Overview of Proposed Speaker Pseudonymization Using Phoneme Embedding

は, VoicePrivacy Challenge[11] に準拠し、Kaldi を使い VoxCeleb [6] を用いて学習された x-vector 抽出器を利用する. また, 置換先の x-vector を生成するため, VoxCeleb から抽出された x-vector 計 7325 個で構成されたプールを用いる.

4.2 モデル構造

アライメントは, Common Voice 14.0 日本語データセット (Train, Dev, Validation) を用いて日本語音声認識用にファインチューンした wav2vec 2.0 [1] モデルを用いて取得した. 各話者の発話から取得したアライメントに基づいて話者照合をするモデルは, [3] でも使用されている Transformer を採用した. 音声合成は, Tacotron2 [9] を用いてメルスペクトログラムを生成し, HiFi-GAN [4] を用いて新たな音声を合成した.

4.3 評価手法

評価では, VoicePrivacy Challenge の評価基準に準拠し, 話者が仮名化されているかどうか, 仮名化後の音声が自然であり発話内容が取得可能かどうか評価する. 仮名化性能は, VoicePrivacy Challenge で使用されている x-vector, PLDA [8] に基づいた話者照合モデルと, 発話から埋め込んだ話者の韻律情報 [3] を基に話者照合を行うモデル (ASV) の 2 種を利用する. 音声認識には, アライメント取得時に利用した Common Voice 14.0 日本語データセット (Train, Dev, Validation) を用いて wav2vec 2.0 [1] をファインチューンした日本語用音声認識モデル (ASR) を用いる. ASV モデルの等価誤り率 (EER) が高いほど発話者が特定できない, すなわち仮名化性能が高く, ASR モデルの単語誤り率 (WER) が低いほど発話内容が取得できるほど仮名化済み音声が明瞭であると評価する.

5 おわりに

本稿では, 従来使用されてきた x-vector を用いた話者特徴量の変換に加え, 話者の韻律情報も変換するマルチファクターな音声仮名化を紹介した. 今後の課題として, 韻律情報の観点に着目した音声仮名化評価法の議論と, 利用目的に応じたレベル別の仮名化が挙げられる.

参考文献

[1] A. Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

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

[3] K. Fujita et al. Phoneme Duration Modeling Using Speech Rhythm-Based Speaker Embeddings for Multi-Speaker Speech Synthesis. In *Proc. Interspeech 2021*, pages 3141–3145, 2021.

[4] J. Kong et al. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

[5] C. Magariños et al. Reversible speaker de-identification using pre-trained transformation functions. *Computer Speech Language*, 46:36–52, 2017.

[6] A. Nagrani et al. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech Language*, 60:101027, 2020.

[7] J. Patino et al. Speaker anonymisation using the mcadams coefficient, 2021.

[8] S. J. Prince et al. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[9] J. Shen et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018.

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

[11] N. Tomashenko et al. The VoicePrivacy 2020 challenge: Results and findings. *Computer Speech &amp Language*, 74:101362, jul 2022.