音声プライバシーのためのブラックボックス型音声加工法 *

☆甲斐優人 (都立大), 高道慎之介 (東大), 塩田さやか, △貴家仁志 (都立大)

1 はじめに

音声には発話内容以外にも発話者の年齢，性別，心理状態，人種などの豊富な情報が含まれており，個人を特定できるデータの 1 つと見なされる．そのため，機械との音声対話の機会が増えると同時に音声の個人情報としての価値についても着目されるようになり，音声データに含まれるプライバシー情報の保護を目的とした音声加工技術に関する研究が注目されてきている．音声データのプライバシーを守る手法として主に暗号化と匿名化の 2 種類が挙げられる．匿名化は暗号化に比べて，計算コストが低く，音声の分野以外の専門知識を必要としないという利点がある．そこで，本研究では音声データの匿名化を目的としたブラックボックス型のハイパーパラメータ最適化法を提案するとともに様々な音声加工法によって匿名化がどの程度可能なのかを検証する．匿名化した音声は原音声と比較して，話者情報が抑制されているとともに，音声の明瞭度が維持されていることが求められている．そこで本研究では，話者照合の照合率と音声認識の単語エラー率を用いた目的関数を定義し，様々な音声加工法において目的関数が最小となるような音声加工法のハイパーパラメータ最適化を行う．実験では VoicePrivacy 2020 で提供されている評価タスクを用いて評価を行った．実験結果からリサンプリングによって音声認識率をある程度維持したまま，話者照合率を低下できることを報告する．

2 VoicePrivacy 2020

2.1 音声プライバシーに関するコンペティション

音声データのプライバシーを守る必要性が高まっているが，音声プライバシーに関連する研究はまだ多くない．また，音声データの匿名化に求められる要件の一般的な定義や，共通の評価基準が存在しない．そこで，音声データを匿名化する手法の開発を促進するとともに，評価基準およびベンチマークの基準を設ける目的で VoicePrivacy 2020 というコンペティションが開催された．

第 1 回目のコンペティションでは，共通のデータセットと評価プロトコルおよび 3 つの条件が与えられている．条件の 1 つ目は，構築したシステムの出力が音声波形であること．2 つ目は，匿名化された音声においても明瞭性は維持されており，かつ話者情報は匿名化されていること．3 つ目は，匿名化された音声同士であれば同一話者かどうかの判定が可能であること．この 3 つの条件をどの程度満たしているかがコンペティション評価基準の一部となっている．

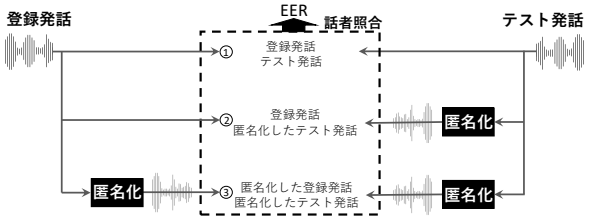

Fig. 1 匿名化した音声の話者照合による評価プロトコル

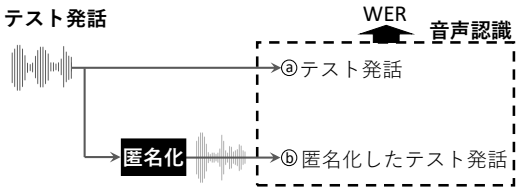

Fig. 2 匿名化した音声の音声認識による評価プロトコル

2.2 データベース

学習に使うデータベースは LibriSpeech [1]，LibriTTS [2]，VCTK [3] それぞれの開発用データセットおよび VoxCeleb-1，2 [4,5] の全データセットのみと指定されている．LibriSpeech はオーディオブックからおよそ 1000 時間分の読み上げられた音声を取ってきたコーパスである．LibriTTS は LibriSpeech を text-to-speech 用にデザインしたコーパスであり，およそ 585 時間分の音声やテキスト等が正規化されている．VCTK は 109 人の英語を母国語とする話者が読み上げた音声をおよそ 44 時間分収録したコーパスである．VoxCeleb-1，2 は Youtube にアップロードされた著名人のインタビュービデオから収集されている．VoxCeleb-1 は学習用のデータセットが話者数 1211，発話数は 148,642，照合用のデータセットが話者数 40，発話数が 4874，VoxCeleb-2 は話者数 5994，発話数は 1,092,009 含んでいる．評価には LibriSpeech，VCTK のそれぞれの評価用データセットを用いる．

2.3 評価基準

VoicePrivacy 2020 の評価タスクには，客観的評価と主観的評価が用意されているが本稿では，客観的評価についてのみ述べる．客観的評価では，与えられた話者照合システムを用い，Fig. 1 の①～③ の場合における等価エラー率 (EER) を求めると同時に，与えられた音声認識システムを用いて Fig. 2 のⓐ，ⓑ の場合における単語エラー率 (WER) をそれぞれ計算す

* Black-box voice modification method for voice privacy. by KAI, Hiroto (Tokyo Metropolitan University), TAKAMICHI, Shinnosuke (The University of Tokyo), SHIOTA, Sayaka, KIYA, Hitoshi (Tokyo Metropolitan University)

る．2.1 節で述べた条件 2，3 を満たすためには Fig. 1 の①と③ の登録発話とテスト発話の条件が揃っていること場合には登録者が正しく特定される必要があるため EER が低くなり，② の場合，つまり登録発話とテスト発話の条件が揃っていない場合には EER が高くなることが求められている．また，2.1 節の条件 2 を満たすためには，Fig. 2 に示すⓐ，ⓑ両方の場合において WER がほぼ同等の性能となることも求められる．

評価用に提供される話者照合には，state-of-the-art な手法の 1 つである x-vector に基づく話者照合システムが採用されている [6]．音声認識には，入力特徴量が i-vector と MFCC である TDNN-F 音響モデルと trigram 言語モデルを用いる音声認識システムが採用されている [7]．両システムの構築には LibriSpeech-clean-360 が用いられており，学習後のモデルパラメータが提供されているため，参加者はそれらを変更することなく用いて評価結果を算出することになっている．

3 音声プライバシーのための音声加工パラメータ最適化

2 章で述べた VoicePrivacy 2020 に準ずる音声の匿名化手法として，EER と WER を用いた目的関数を定義し，この目的関数を最小化することで匿名化に適した音声加工を行うことを考える．音声加工には，機械学習を用いたアプローチと信号処理を用いたアプローチが考えられる．深層学習に代表される前者は，無数のモデルパラメータが存在するが，学習データを用いて半自動的に最適化できる利点がある．一方，後者は最適化する必要のあるハイパーパラメータが少数しか存在しないという利点があるが，自動的な最適化は難しい．本章では，両方のアプローチの利点を両立する手法として，信号処理的アプローチをとる音声加工法のハイパーパラメータを機械学習によるアプローチで決定する手法を提案する．Figure 3 で示したフロー図のとおり，提案法では，3.1 節で述べる音声加工法を入力音声に施した加工音声と入力音声に対する正解テキストおよび話者ラベルを入力情報として用いる．そこから得られる WER および EER を組み合わせた目的関数の値に基づいて音声加工のハイパーパラメータが最適化される．

3.1 音声加工法

3.1.1 声道長正規化

声道長正規化 (vocal tract length normalization; VTLN) は，話者間の声道長などの違いを取り除く手法である [8]．この手法では，原音声の短時間フーリエ変換を通して得られる対数振幅スペクトルの周波数軸を周波数伸縮係数に基づいて伸縮する．以降では，この周波数伸縮係数を $\alpha_{\mathrm{vtln}} \in [-1, 1]$ とする．

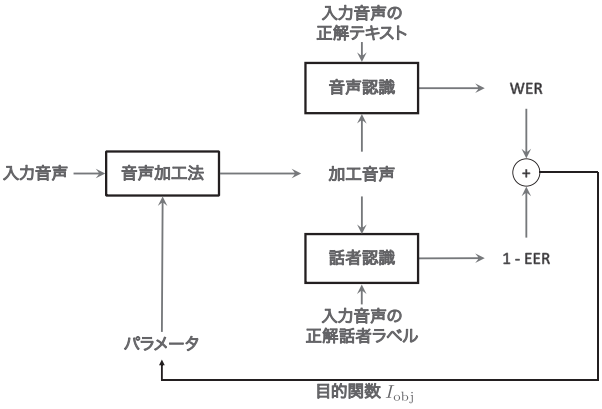

Fig. 3　話者照合と音声認識の評価スコアを目的関数として用いた音声加工パラメータ探索のフロー図

3.1.2 リサンプリング

リサンプリングは，原音声のサンプリング周波数を変化させる手法である．リサンプリングにより原音声の音声波形長が変わってしまうことを防ぐため，本稿では，サンプリング周波数の変化率の逆数だけ原音声を時間伸縮させたのち，原音声の波形長になるように伸縮後音声をリサンプリングする．以降では，このサンプリング周波数の変化率を $\alpha_{\mathrm{resample}} \in \mathbb{R}^{+}$ とする．

3.1.3 McAdams 変換

McAdams 変換は，原音声の共振周波数を変化させる手法である [9]．この手法では原音声の線形予測符号 (linear predictive coding; LPC) [10] を通して得られた共振周波数と共振強度のうち共振周波数をパラメータで累乗し，変化後の共振周波数と元の共振強度から得られる LPC 係数のフィルタを LPC 残差信号に畳み込むことで音声加工を行う．以降では，共振周波数の値の累乗パラメータを $\alpha_{\mathrm{mcadams}} \in \mathbb{R}_{0}^{+}$ とする．

3.1.4 変調スペクトルスムージング

変調スペクトル (modulation spectrum; MS) スムージングは，音声データの音声特徴量時系列の高域変調周波数成分を除去する手法である [11]．この手法では，原音声の短時間フーリエ変換を通して得られる各周波数の対数振幅スペクトル時系列に対し，ローパスフィルタを施す．以降では，ローパスフィルタのカットオフ変調周波数を $\alpha_{\mathrm{ms}} \in [0, 1]$ とする．ただし，ナイキスト変調周波数は 1.0 である．

3.1.5 クリッピング

クリッピングは，原音声の振幅値が所望の範囲を超える箇所を揃える方法である．本稿では，音声の振幅値の絶対値の累積ヒストグラムを計算し，累積値に対して設ける閾値を超える振幅をクリッピング対象とする．以降では，累積値に対して設ける閾値を $\alpha_{\mathrm{clip}} \in [0, 1]$ とする．

Table 1　各音声加工法の音声加工パラメータとその探索範囲

音声加工法	パラメータ	探索範囲
声道長正規化	α_{vtln}	[-0.2, 0.2]
リサンプリング	$\alpha_{\mathrm{resample}}$	[0.3, 0.9]
McAdams 変換	$\alpha_{\mathrm{mcadams}}$	[0.7, 1.3]
MS スムージング	α_{ms}	[0.05, 0.3]
クリッピング	α_{clip}	[0.3, 1.0]
コーラス	α_{chorus}	[0.0, 0.2]

3.1.6　コーラス

コーラスは，原音声のピッチなどを微小に変更させた音声波形を，原音声に重畳する方法である．本稿では，原音声を α_{vtln} の絶対値の正と負の値でそれぞれ別々に加工した音声と原音声を重畳することで音声加工を行う．以降では，コーラスのハイパーパラメータを $\alpha_{\mathrm{chorus}} \in [0, 1]$ とする．

3.2　ハイパーパラメータ最適化の目的関数

3.1 節で述べた音声加工法のハイパーパラメータを最適化するための目的関数と最適化の手段について述べる．匿名化の条件を満たすためには，WER は小さく，EER は大きくなるような音声加工パラメータを選択する必要がある．ゆえに本稿では，最小化する目的関数 I_{obj} を式 (1) で定義する．

$$I_{\mathrm{obj}} = \mathrm{WER} + \omega(1 - \mathrm{EER}) \qquad (1)$$

ここで，ω は話者照合の正解率スコアを加算する際のバランスを調整するための重みとする．話者照合および音声認識，音声加工の微分可能性を保証していないため，式 (1) を最小化するハイパーパラメータの最適化には文献 [12,13] のようなブラックボックス型のアルゴリズムを適用する必要がある．

4　実験

4.1　実験条件

提案法であるハイパーパラメータの最適化法の実行条件として，ハイパーパラメータの探索回数を 50 回，目的関数の重み $\omega = 1.0$ として探索を行った．これは予備実験の結果から単独の音声加工法のハイパーパラメータ探索においては重み ω の影響があまりないことが確認されたためである．各音声加工法のハイパーパラメータ探索範囲を Table 1 にまとめる．最適なハイパーパラメータの探索には，Optuna [14] を用いた．VoicePrivacy 2020 の評価データベースとして VCTK と LibriSpeech が用意されているが，本報告では VCTK の女性話者のみの結果を示す．

VoicePrivacy 2020 では，音声匿名化のベースラインシステムとして深層学習や音声合成など様々な機械学習技術を取り入れた匿名化手法 (ベースライン 1) と 3.1.3 節でも述べた McAdams 変換に基づく音声加工法 ($\alpha_{\mathrm{mcadams}} = 0.8$) による匿名化手法 (ベースライン 2) が公開されている．提案法は機械学習的アプローチと信号処理的アプローチを混ぜた手法となっており，ベースライン 2 よりもベースライン 1 に近づくような精度を得ることが目標の 1 つとなる．

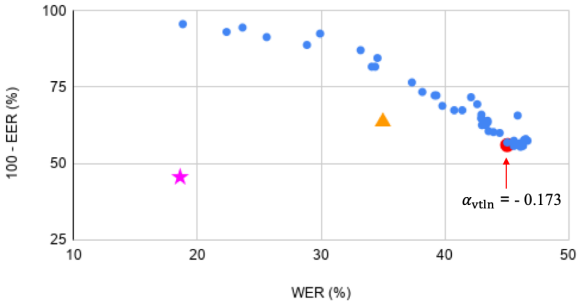

Fig. 4　α_{vtln} の探索による WER と EER の推移

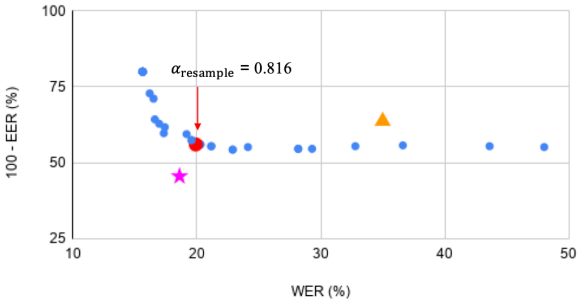

Fig. 5　$\alpha_{\mathrm{resample}}$ の探索による WER と EER の推移

4.2　実験結果

Fig. 4 から Fig. 9 に各音声加工法とそのハイパーパラメータ探索時の WER(%) と匿名化の達成度 (1 - EER)(%) の推移を示した．各図の赤い点が目的関数が最小化された時の点，紫色の星印がベースライン 1，橙色の三角がベースライン 2 の結果をそれぞれ示している．各図の点は左下に近づくほど匿名化の性能が高いことを示している．各音声加工法は非常に簡単な手法を用いていることから全体的に性能が高くないが，Fig. 5 のリサンプリングを行った際には赤い点がベースライン 1 に非常に近くなっており，WER を維持しながら匿名化できていることが確認できる．Figure 4 の VTLN の結果においても，匿名化の達成度としては，リサンプリングと同程度の値が得られたが，WER が大幅に低下してしまっているため，匿名化の条件を満たせていないことがわかる．他の手法では逆に，WER はあまり低下させないものの匿名化があまりできていない結果になっていることが確認できる．これらの結果は VCTK の男性話者を用いた場合にも近い傾向だった．しかし，LibriSpeech の場合は結果の傾向が変わってた部分もあるので，今後さらに実験結果を考察する必要がある．

5　まとめ

本報告では，音声データのプライバシーを保護するための音声加工法のための，ブラックボックス型の

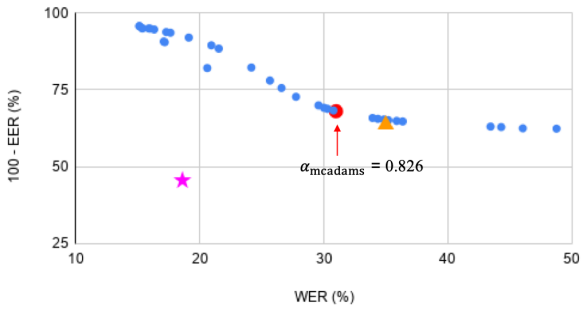

Fig. 6　α_{mcadams} の探索による WER と EER の推移

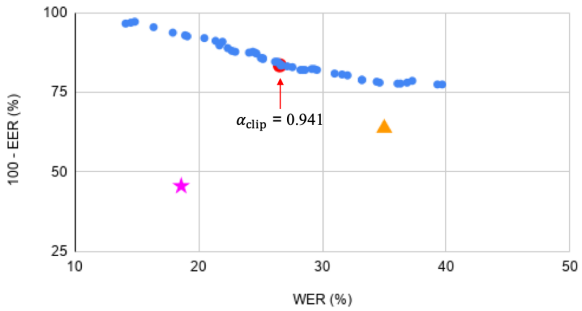

Fig. 8　α_{clip} の探索による WER と EER の推移

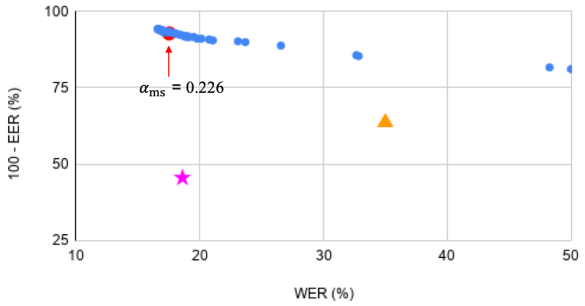

Fig. 7　α_{ms} の探索による WER と EER の推移

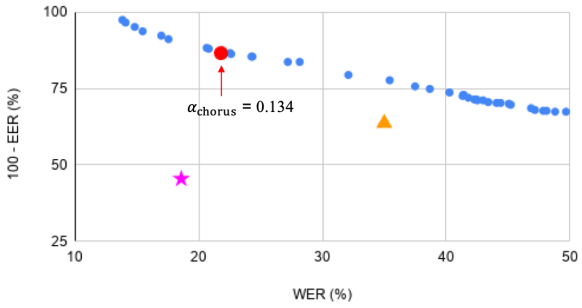

Fig. 9　α_{chorus} の探索による WER と EER の推移

パラメータ最適化法を提案する．VoicePrivacy 2020 で公開されている評価用の話者照合および音声認識を用い，それぞれの結果の EER と WER を目的関数に用いることで音声の匿名化と明瞭性の維持を可能とするハイパーパラメータの探索を行った．実験結果より，リサンプリングによって音声を加工する場合，EER を 50%弱低下させつつ，WER の低下は女性は小規模な低下に留めることに成功した．今後の課題として，データベースへの依存性の調査や，複数の音声加工法の組み合わせによるハイパーパラメータの最適化などが挙げられる．

謝辞

本研究は，JSPS 科研費若手研究 JP19K20271 と ROIS-DS-JOINT(023RP2020)、セコム財団挑戦的研究助成の助成を受けたものである。

参考文献

[1] V. Panayotov et al., “Librispeech: an ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.

[2] H. Zen et al., “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” *arXiv preprint arXiv:1904.02882*, 2019.

[3] J. Yamagishi et al., “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” 2019.

[4] A. Nagrani et al., “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.

[5] J. Chung et al., “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

[6] D. Garcia-Romero et al., “x-vector DNN refinement with full-length recordings for speaker recognition.” in *Proc. Interspeech*, 2019, pp. 1493–1496.

[7] V. Peddinti et al., “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. Interspeech*, 2015, pp. 3214–3218.

[8] L. Lee, R. Rose, “A frequency warping approach to speaker normalization,” *IEEE Transactions on speech and audio processing*, vol. 6, no. 1, pp. 49–60, 1998.

[9] S. E. McAdams, “Spectral fusion, spectral parsing and the formation of auditory images,” *Ph.D dissertation, Stanford University*, 1985.

[10] F. Itakura, “Analysis synthesis telephony based on the maximum likelihood method,” in *The 6th international congress on acoustics*, 1968, pp. 280–292.

[11] S. Takamichi et al., “The NAIST text-to-speech system for the Blizzard Challenge 2015,” in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[12] J. Snoek et al., “Practical bayesian optimization of machine learning algorithms,” in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[13] J. Bergstra et al., “Hyperopt: a python library for model selection and hyperparameter optimization,” *Computational Science & Discovery*, vol. 8, no. 1, p. 014008, 2015.

[14] T. Akiba et al., “Optuna: A next-generation hyperparameter optimization framework,” in *Proc. ACM SIGKDD*, 2019, pp. 2623–2631.