周波数伸縮に基づく話者匿名化のための クラウドソーシングに基づくパラメータ最適化*

高道 慎之介 (東大院・情報理工), ○小沼 海, 金田 卓, 金田 隆志 (HOYA),
齋藤 佑樹, 郡山 知樹, 猿渡 洋 (東大院・情報理工)

1 はじめに

クラウドソーシングサービスの普及により，群衆の力を利用した大規模なアノテーションが期待される．音声データに対するアノテーションは従来，in-houseの熟練したアノテータにより実施されていたが，クラウドソーシングサービスの台頭により，大量のクラウドワーカ（評価者）による実施も可能となった．そのような音声アノテーションでは，クライアント（例えば研究者）が音声データをウェブサーバにアップロードした後，評価者がそのサーバの音声を受聴してタグ等を付与する．このアノテーション法は，アノテーション品質制御の問題を持つものの，群衆の力を容易に利用できる点で強力な可能性を持つ．

しかしながら，クラウドソーシングを用いた音声アノテーションは，音声に含まれる個人情報（例えば，発話内容や話者性）漏えいの危険性を孕む [1]．例えば，in-house の音声を評価者に提示した場合，悪意のある評価者によりその音声がダウンロードされ，その個人情報が分析・取得されてしまう可能性がある．本稿では，元音声の発話内容と音質を保持しながら話者性を隠匿する話者匿名化技術に着目する．この方法の確立により，音声アノテーションの精度を落とさずに音声を話者匿名化し個人情報漏洩を防止できると期待される．音声の話者匿名化は音声から音声への変換と見なされるため，統計的話者変換（声質変換 [2] とも呼ばれる）が主要なアプローチとなりうる [3]．この統計的話者変換は，統計モデリングに恩恵を受けた強力な話者匿名化を可能にするが，音声の音質劣化と発話内容欠落をしばしば発生させる [4]．加えて，統計的手法に基づく話者匿名化を利用できるアノテーションタスクは，非常に限定される．例えば，複数話者混合音声（複数の話者が同時に発話している音声）の発話者数アノテーションは，話者ダイアライゼーション [5] や音声分離 [6] の学習データ作成に必要だが，統計的手法は，複数話者混合音声を入力にとらないため本アノテーションタスクにおいて実用的ではない．

より広範囲のアノテーションタスクに利用可能，かつ高い音質・話者匿名度をもつ話者匿名化を目指し，本稿では，周波数伸縮に基づく話者匿名化を提案する．従来の統計的手法と提案法の比較を**図 1** に示す．提案法は，区分線形関数を用いて対数スペクトル包絡を変形する．そのモデルパラメータは，大規模（約1,000 人）の評価者とパブリック音声コーパスによる主観評価により，匿名化音声の音質と話者匿名度の両方を最大化するように決定される．提案法は複数話者混合音声も入力として扱えるため，発話者数アノテーションにおける話者匿名化にも利用できる．実験的評価から，提案法は，音質・話者匿名度に加えて明瞭度においても従来法を超える性能を達成できることを示す．また，提案法を発話者数アノテーションに応用した結果，提案法はアノテーション精度の僅かな劣化のみで音声を話者匿名化できることを示す．

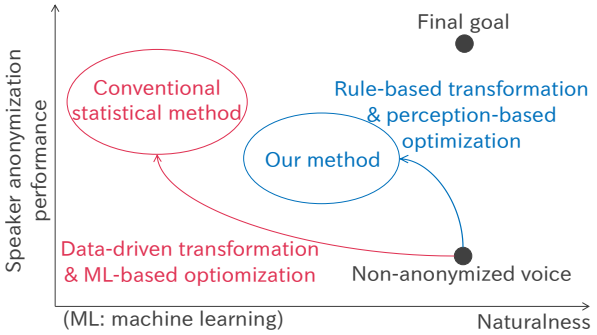

Fig. 1 従来法と提案法の比較

2 従来の統計的手法

統計的話者変換を用いた従来の話者匿名化を概説する．この手法では，学習済みの統計モデル（例えば，deep neural network (DNN)）を用いて，元話者の音声を匿名話者の音声に変換する．ここで，匿名話者とは，その音声の話者性が特定個人と紐付いていない話者を指す．典型的な一対一話者変換 [2] を用いた話者匿名化では，パラレル音声（2 話者が同じ発話内容を別々に発話した音声）を用いる．$\boldsymbol{x}$ と $\boldsymbol{y}$ をそれぞれ，元話者と匿名話者の音声特徴量（ボコーダ特徴量）とすると，統計モデル $\boldsymbol{G}(\cdot)$ は，この $\boldsymbol{x}$ と $\boldsymbol{y}$ を用いて事前に学習される．話者匿名化時には，ボコーダを用いて，変換後の特徴量 $\boldsymbol{G}(\boldsymbol{x})$ から匿名化音声を生成する．一方で，一対多話者変換 [7] を用いた話者匿名化では，x-vector [8] のような話者ベクトルで匿名話者を制御する．統計モデルは，事前収録した多人数話者の音声を用いて学習される．

これらの統計的手法は，高い話者匿名度を達成するが，音声の音質を劣化させ発話内容を欠落させる．ニューラルボコーダ [9] やテキスト補助情報 [10] は，この問題の緩和に多少有効だがその性能は十分でない．また，統計的話者変換は単一話者の音声を入力にとるため，従来の統計的手法を利用できるアノテーションタスクは非常に限定される．音声分離 [6] は，従来の統計的手法を複数話者混合音声に適用する際の前処理に利用できるが，そのようなパイプライン

*Crowdsourcing-based parameter optimization for frequency warping-based speaker anonymization, by TAKAMICHI, Shinnosuke (The University of Tokyo), ONUMA, Kai, KANEDA, Taku, KANEDA, Takashi (HOYA Corporation), SAITO, Yuki, KORIYAMA, Tomoki, and SARUWATARI, Hiroshi (The University of Tokyo)

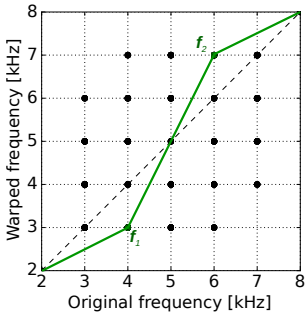

Fig. 2　$\boldsymbol{f}_1 = (4000, 3000)\,, \boldsymbol{f}_2 = (6000, 7000)$ をもつ区分線形関数に基づく周波数伸縮．対角の破線は伸縮なしに対応する．

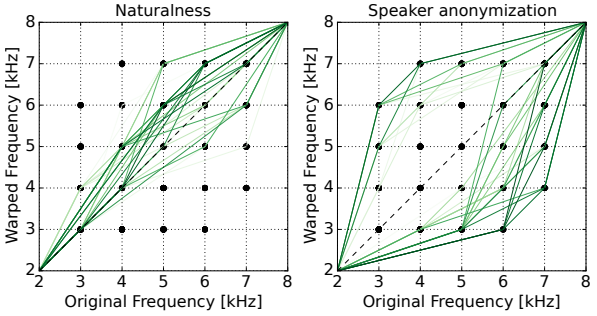

Fig. 3　音質 (左) と話者匿名度 (右) のそれぞれに関してスコアの高い 30 個の関数．濃い色ほど高いスコアを表す．

処理は，最終的な音質を著しく劣化させる [11]．

3　提案法

周波数伸縮に基づく話者匿名化法を提案する（**3.1 節**）．伸縮に用いるモデルパラメータは，クラウドソーシングを用いて，音質と話者匿名度の両方を最大化するように最適化される（**3.2 節**）．提案法はボコーダ特徴量を用いて匿名化するが，広範囲のアノテーションタスクへの応用を見据え，提案法をボコーダフリーの枠組みに拡張する（**3.3 節**）．

3.1　周波数伸縮に基づく話者匿名化

周波数伸縮は，フォルマントに基づく統計的話者変換 [12] において提案されたものであり，本稿はこれを話者匿名化に利用する．ボコーダを用いて入力音声のスペクトル包絡を抽出した後，事前に決定した区分線形関数に基づいてスペクトル包絡の周波数を伸縮する．匿名化音声は，周波数伸縮したスペクトル包絡と非加工の音源特徴量から合成される．**図 2** に，2 つの伸縮点を用いた区分線形関数の例を示す．この関数は 2 つのモデルパラメータ $\boldsymbol{f}_1, \boldsymbol{f}_2$ を有する．モデルパラメータの要素は，話者知覚の主要成分 [13] である 2 kHz から 8 kHz の値を取る．本稿では，当該周波数帯域をグリッド状（**図 2** の灰色の点に相当）に区切り，$\boldsymbol{f}_1$ と $\boldsymbol{f}_2$ の取りうる全ての組み合わせを，モデルパラメータ候補とする．

3.2　クラウドソーシングを用いたパラメータ最適化

パブリック音声コーパスを用いて，モデルパラメータ候補のうち最適な 1 つを選択する．まず，全てのモデルパラメータ候補を用いて，コーパスに含まれる多人数話者の音声を話者匿名化する．次に，匿名化音声の音質と話者匿名度に関する大規模主観評価を実施する．音質評価では，匿名化後音声を評価者に提示してその音質を評価させる．話者匿名度の評価では，元音声（すなわち，匿名化されていない音声）と匿名化後音声を提示して，その話者匿名度（すなわち，2 つの音声の話者性がどの程度異なるか）を評価させる．各モデルパラメータ候補について音質スコアと匿名化スコアを平均した後，そのスコアの二乗和が最大のモデルパラメータを，最適パラメータとして用いる．

3.3　ボコーダフリーへの拡張

3.1 節の手法をボコーダフリーの枠組みに拡張するために，ボコーダ分析の代わりに離散フーリエ変換を使用する．**3.2 節**で最適化したモデルパラメータを用いて，短時間フーリエ変換の対数振幅スペクトルを周波数伸縮する．匿名化後音声は，伸縮後の対数振幅スペクトルと元音声の位相を用いて，逆短時間フーリエ変換により生成する．この拡張により，複数話者混合音声のようなボコーダ分析困難な音声に対しても提案法を適用できる．

3.4　考察

規則ベースの話者匿名化法はいくつか提案されている [14]．このモデルパラメータ最適化には，非常に小規模（典型的には，in-house の評価者）の主観評価により決定されてきた．一方で本研究では，クラウドソーシングにより大量の人間を容易に雇用できること [15] を利用して，群衆の力によりモデルパラメータを最適化する．

4　実験的評価

4.1　実験条件

パブリックコーパスの音声として女性 12 名による音声データを用い，匿名化する話者を JVS コーパス [16] の女性 51 名から選択する．この 12 名と 51 名の間で話者の重複はない．サンプリング周波数は 16 kHz，音声特徴量は WORLD [17, 18] で抽出した 513 次元の対数スペクトル包絡である．クラウドソーシングサービスとして Lancers[1]を使用する．提案法の伸縮点は，帯域を 1 kHz 毎に分割した $3, 4, 5, 6, 7$ kHz のいずれかである．提案法のモデルパラメータの候補数は 92 である．

4.2　クラウドソーシングを用いた最適化

まず，クラウドソーシングを用いて提案法のモデルパラメータを最適化する．ここでは，音質に関する MOS (mean opinion score) テストと，話者匿名度に関する DMOS (Degradation MOS) テストを実施する．MOS テストでは，評価者に匿名化音声を提示し，その音質を 1（非常に不自然）から 5（非常に自然）の 5 段階で回答させる．DMOS テストでは，評価者に元音声と匿名化音声を提示し，その話者類似度を

[1] https://www.lancers.jp/

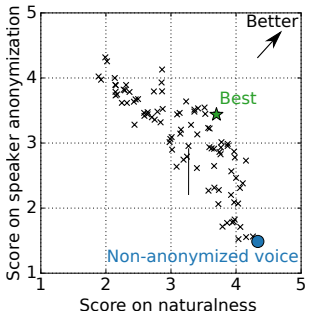

Fig. 4　音質 (x 軸) と話者匿名度 (y 軸) の散布図．各点が各モデルパラメータに対応する．

Table 1　音質に関するプリファレンススコア．高いほど良い．**太字**は p-value < 0.05 で優れた手法を表す．

A	Scores	p-value	B
Statistical (sim)	0.345 vs. **0.655**	$< 10^{-10}$	Proposed
Statistical (med)	0.485 vs. 0.515	0.397	Proposed
Statistical (dis)	0.430 vs. **0.570**	7.09×10^{-5}	Proposed

1（非常に似ていない）から 5（非常に似ている）の 5 段階で回答させる．DMOS テストに関しては，得られたスコアを 5 から引いたものを話者匿名度のスコアとして扱う．各評価者は，25 問（MOS テスト）もしくは 13 問（DMOS テスト）に対して回答する．各評価者は，女性 12 名の匿名化音声と，1 名の元音声を受聴する．MOS テストと DMOS テストの評価者数はそれぞれ，385 名と 607 名である．

図 3 に，音質と話者匿名度のそれぞれに関して高いスコアを獲得した 30 個の関数を示す．音質に関して，対角に近い関数が高いスコアを獲得している．しかしながら，高周波数帯域は，低周波数帯域と比べて大きく伸縮される傾向にある．高周波数帯域はフォルマント位置を含まないため，この結果は合理的である．一方で，話者匿名度に関して，対角から離れた関数が高いスコアを獲得している．

図 4 に，音質と話者匿名度のスコアの散布図を示す．両スコアは負に相関しているが，いくつかのモデルパラメータは音質と話者匿名度の両方で高いスコアを獲得している．以降の評価では，両スコアの二乗和が最大のモデルパラメータ（図中の “Best”）を使用する．

4.3　従来法との比較

次に，音質と話者匿名度に関して従来法と提案法を比較する．さらに本節では，両手法の明瞭性（元音声の発話内容をどの程度保持しているか）に関しても比較する．使用する音声は以下の 3 つである．

Statistical (sim, med, dis): 従来法 (**2 節**)
Proposed: 提案法 (**3 節**)
Non-anonymized: 元音声 (話者匿名化なし)

匿名化される話者は，JVS コーパスの *jvs019* である．従来法の匿名話者は，同コーパスの *jvs096* (sim)・*jvs008* (med)・*jvs093* (dis) である．これらの話者は，匿名化される話者との主観的話者類似度スコア [16] が最高値 (すなわち，主観的に非常に似ている)，中央値，最低値の話者である．なお，*jvs019* と *jvs096* は，コーパス中で最高の主観的話者類似度スコアをもつ話者対である．話者変換モデルは，59 ユニットの入力層，2×256 ユニットの gated linear [19] 隠れ層，59 ユニットの線形出力層を持つ feed-forward DNN である．従来法の音声特徴量は，1 次から 59 次のメルケプストラム係数である．DNN パラメータの最適化には学習係数 0.01 の AdaGrad [20] を使用する．DNN の学習には 90 発話の音声を使用する．

Table 2　話者匿名性に関するプリファレンススコア．低いほど良い．**太字**は p-value < 0.05 で優れた手法を表す．

A	Scores	p-value	B
Statistical (sim)	0.535 vs. **0.465**	0.047	Proposed
Statistical (med)	**0.440** vs. 0.560	6.71×10^{-4}	Proposed
Statistical (dis)	**0.250** vs. 0.750	$< 10^{-10}$	Proposed

Table 3　CER [%]．低いほど良い．

Statistical (sim)	Proposed	Non-anonymized
5.25	**4.20**	4.21

クラウドソーシングを用いて，音質に関するプリファレンス AB テストと，話者匿名度に関するプリファレンス XAB テストを実施する．AB テストでは，従来法と提案法の匿名化音声を評価者にランダムに提示し，高音質の音声を選択させる．XAB テストでは，リファレンスとして元音声を提示したのちに従来法と提案法の匿名化音声を提示し，元音声に類似した音声を選択させる．XAB テストに関しては，スコアの低い手法が優れていることに注意する．各評価者は 10 問に対して回答する．評価に用いる音声データは，従来法の DNN 学習及び提案法の最適化に使用していない 10 発話である．各テストにおける評価者数は 40 名である．さらに，明瞭度評価のために，クラウドソーシングを用いた書き起こしテストを実施する．評価者に元音声もしくは匿名化音声を提示し，その発話テキストをひらがなで回答させる．評価には 40 文を使用し，各評価者は 9 文 (従来法・提案法・元音声を 3 発話ずつ) のテキストを書き起こす．書き起こしテストの評価者数は 184 名である．

表 1 と**表 2** にそれぞれ，音質と話者匿名度の評価結果を示す．**表 1** より，提案法のスコアは，従来法のスコアよりも常に良いことが分かる．一方で，元話者に類似した匿名話者 (sim) に従来法で変換する場合よりも，提案法は良いスコアを得ている．これらの結果より，提案法は，高い音質を達成しながら，元話者に類似した話者に変換するよりも良い話者匿名度を達成できることが示される．

表 3 に，書き起こしテストにおける character error rate (CER) の結果を示す．提案法の CER は元音声と同程度かつ従来法よりも低いため，本実験により，提案法は明瞭性劣化なしで話者匿名化できることが示される．

4.4　アノテーションにおける評価

最後に，提案法の話者匿名化がアノテーション精度に与える影響を調査する．ここでは，クラウドソーシングを用いて，複数話者混合音声に対する話者数

Table 4　元音声を用いたアノテーション数 (Accuracy = 0.624)．**太字**は正しくアノテートされた数を表す．

		Annotated					Recall
		1mix	2mix	3mix	4mix	5mix	
Ref.	2mix	3	**710**	54	7	0	0.917
	3mix	4	217	**478**	68	7	0.618
	4mix	2	57	395	**260**	60	0.336
Precision		-	0.722	0.516	0.776	-	

Table 5　匿名化音声を用いたアノテーション数 (Accuracy = 0.555)．**太字**は正しくアノテートされた数を表す．

		Annotated					Recall
		1mix	2mix	3mix	4mix	5mix	
Ref.	2mix	3	**636**	121	14	0	0.822
	3mix	1	231	**441**	88	13	0.570
	4mix	1	96	415	**212**	50	0.274
Precision		-	0.660	0.451	0.675	-	

アノテーションを対象とする．アノテーションに用いる音声として，2, 3, 4 名が同時に発話する音声を用意する．この音声は，JVS コーパスの女性 51 名の音声データのうち，発話内容は異なるが発話区間の近い 2, 3, 4 発話を選択してミックスして作成される．すなわち，アノテーションに用いる音声は，複数話者が同一の音声区間で同時に発話している音声である．提案法として **3.3 節**のボコーダフリー話者匿名化を使用する．アノテーション時には，評価者に音声を提示し，その発話者数を 1 人から 5 人で選択させる．評価者数は 387 人である．

表 4 と**表 5** にそれぞれ，元音声と匿名化音声を用いたアノテーションの confusion matrix を示す．匿名化による accuracy の劣化は 10%以下であるため，この実験より，提案法はアノテーション精度の僅かな劣化のみで音声を話者匿名化できることが示される．

5　まとめ

本稿では，クラウドソーシングでモデルパラメータを最適化する，周波数伸縮に基づく話者匿名化法を提案した．実験的評価から，従来の統計的話者変換を用いる方法よりも，提案法は高い音質・明瞭性・話者匿名度を達成できることを示した．また，発話者数アノテーションに提案法を適用した結果，提案法はアノテーション精度の数%の劣化のみで音声を話者匿名化できることを明らかにした．今後は，他のアノテーションタスクにおいて評価を行う．

謝辞：本研究は，東京大学工学部計数工学科の仮屋郷佑氏の協力のもとで実施した．

参考文献

[1] P. P. Zarazaga et al., “Sound privacy: A conversational speech corpus for quantifying the experience of privacy,” in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 3720–3724.

[2] T. Toda et al., “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] F. Fang et al., “Speaker anonymization using X-vector and neural waveform models,” in *Proc. SSW10*, Vienna, Austria, Sep. 2019.

[4] J. L.-Trueba et al., “The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, Rhodes, Greece,, Jun. 2018, pp. 195–202.

[5] S. E. Tranter, D. A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[6] J. R. Hershey et al., “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 31–35.

[7] Y. Saito et al., “Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[8] D. Snyder et al., “X-Vectors: robust DNN embeddings for speaker recognition,” in *Proc. ICASSP*, Calgary, Canada, May 2018, pp. 5329–5333.

[9] A. Tamamori et al., “Speaker-dependent WaveNet vocoder,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1118–1122.

[10] L. Sun et al., “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[11] Y. Tajiri, T. Toda, “Nonaudible murmur enhancement based on statistical voice conversion and noise suppression with external noise monitoring,” in *Proc. SSW9*, Sunnyvale, CA, U.S.A., Sep. 2016, pp. 54–60.

[12] Y. Qian et al., “A frame mapping based HMM approach to cross-lingual voice transformation,” in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5120–5123.

[13] T. Kitamura, M. Akagi, “Speaker individualities in speech spectral envelopes,” *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283–289, 1995.

[14] P. Jiří et al., “Evaluation of speaker de-identification based on voice gender and age conversion,” *Journal of Electrical Engineering*, vol. 69, pp. 138–147, Mar. 2018.

[15] Y. Kawase et al., “Mob scene filter: Conversion of facial appearance by changing position and size of facial regions,” *Transactions of the Virtual Reality Society of Japan*, vol. 21, no. 3, pp. 483–492, Mar. 2016.

[16] S. Takamichi et al., “JVS corpus: free japanese multi-speaker voice corpus,” *arXiv preprint, 1908.06248*, Aug. 2015.

[17] M. Morise et al., “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[18] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

[19] Y. N. Dauphin et al., “Language modeling with gated convolutional networks,” *arXiv preprint, 1908.06248*, 2016.

[20] J. Duchi et al., “Adaptive subgradient methods for online learning and stochastic optimization,” *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121—-2159, Jul. 2011.