Vol.2019-SLP-129 No.1
2019/10/11

音声認識のためのプライバシー保護音響モデル学習法

太刀岡 勇気[1,a)]

概要：ドメイン内の音声データを使うことで音響モデルの性能を顕著に向上させることができる．しかしながら，ドメイン内のデータは個人情報を含むので，書き起こしが残ると話者のプライバシーを侵害する．これに加えて，ある集団に属していることが知られたくない場合には話者同定も問題となる．ゆえに，ドメイン内データは利用期間が過ぎたら破棄される．ただ，ひとたびデータが破棄されてしまうと，より効果的なモデル構造が将来的に提案されたとしても再学習することができない．このため，音声データのプライバシー保護には価値がある．この際に求められるのは，書き起こしが再生されないことと，プライバシー保護されたデータセットから話者が特定されないことである．本報では，これらの要求を満たすプライバシー保護音響モデル学習を提案する．また 3 種類の特徴量 (n-gram，音素ラベル，音響特徴量) の提案の学習法に対する影響の受けやすさを調査する．影響の受けやすさの解析により，音素ラベルと音響特徴量は n-gram よりも影響が受けにくいことが分かった．これは音響モデルの学習の際に高精度な音素ラベルと音響特徴量が必要なことから考えると，良い性質である．音声認識実験により検証したところ，この良い性質のおかげで，提案法による単語誤り率の低下は 0.6%未満であった．

1. はじめに

音響モデルを学習する際にはたとえ少量であっても，ドメイン内のデータが有効である [1], [2]．しかしながら，ドメイン内データの書き起こしは，個人情報が含まれることが多く，ある集団に属していることを話者が秘匿したい場合には話者が特定されることも問題になりうる．このような事情から，ドメイン内データは利用期間が終了した後は破棄されなければならない．一旦学習データが削除されてしまうと，将来よりよいモデル構造が提案された際に，モデルを再学習することができない．よってデータを削除することなく，個人情報を保護できる技術が求められている．これはプライバシー保護データマイニング (privacy preserving data mining; PPDM) [3], [4] と呼ばれる問題の一つで，特定・暴露リスクを低減することを目的としている [5]．音声データの PPDM を考える際には，攻撃者に話者が何を話したか，話者が誰かということを知られないようにする必要がある．

音声処理の分野では，PPDM に関する研究はほとんど見られない．そのうちの一つでは，計算方法の手順を秘匿化する方法が提案されている [6], [7]．しかしながらこれには，非秘匿な場合に比べて多くの計算量を必要とし，モデルを変更した際には操作のプロトコルを変更する必要があるという問題がある．またこれはデータ保護には使えない．

そのほかの手法はデータ攪乱である．これによって個人情報を消し去ることができるが，特徴量の時系列が完全に失われてしまっているので，識別学習 [8], [9] や end-to-end の手法 [10] を使うことはできない．これを可能にするためには，特徴量の時系列が保存されていなければならない．

これに加えて，近年，学習済みモデルから学習データを再生する手法が提案されている [11]．このような方法が進展すれば，ドメイン内データで学習した深層神経回路網 (Deep neural network; DNN) は攻撃の危険にさらされるので，プライバシー保護したデータセットで学習する必要がある．

よって音声データの PPDM には，元のドメイン内データセットとそれから学習した DNN モデルを削除し，匿名化されたデータセットから個人情報が再生できないようにしつつ，プライバシー保護された時系列データセットを音響モデル学習に使うことができることが求められる．また実データを利用することも重要である．学習データを深層オートエンコーダで生成する方法も提案されているが [12]，生成モデルでは十分に音声のダイナミクスを表すことができない．本報では，これらの要求を満たすプライバシー保護音響モデル学習 (privacy preserving acoustic model training; PPAMT) を提案する．PPDM の調査論文 [13] では，様々な PPDM 手法を分類している (文献 [13] の表 1)．これによれば，我々の方法は “perturbation”，“randomization”，

[1] デンソーアイティーラボラトリ
東京都渋谷区渋谷 2-15-1 渋谷クロスタワー 28F 150-0002
[a)] ytachioka@d-itlab.co.jp

ⓒ 2019 Information Processing Society of Japan

“anonymization” に分類される．

プライバシーは入力プライバシーと出力プライバシーに分類される．入力プライバシーでは，例えばプライバシー保護データ公開 (privacy preserving data publishing; PPDP) のように，データセットにノイズを加える．PPAMT ではこれを利用する．発話は文節に分割され，文節をランダムに結合することで，書き起こしを匿名化する．PPAMT が 3 種の特徴量 (n-gram，音素ラベル，音響特徴量) に与える確率を定式化する．, これに加えて，話者を秘匿するため，話者クラスタリングに基づく k 匿名化 [14], [15] を使う．これにより，攻撃者に話者を k 人の候補者より絞り込ませないことができる．

出力プライバシーは，DNN の出力を事後確率に変換するために使われるトライフォン状態の事前分布に表れる．PPAMT によるそれらの違いは差分プライバシー (differential privacy; DP)[16]) での “perturbation” [17] やランダムサンプリング [18] の考察と同様の方法により評価される．

本報の残りは以下のようになっている．2 節では，“perturbation” と “ randomization” を用いた PPAMT を提案する．その際に，上述の 3 種の特徴量に対する PPAMT の敏感さ (sensitivity) を解析する．加えて，話者を秘匿するため，3 節に述べる方法により，話者クラスタリングを使う．4 節での実験により，大語彙連続音声認識 (large-vocabulary continuous speech recognition; LVCSR) タスクに対する提案の PPAMT の有効性を示す．

2. プライバシー保護音響モデル学習 (PPAMT)

2.1 概要

図 1 に，「はじめに」に記した要求条件を満足する提案の PPAMT のフレームワークを示す．PPAMT の基本の操作は “perturbation” と “randomization”[13] である．まず，文を短いポーズや文節境界で，文節に分割する．次に，これらの文節をランダムに結合することで新しい文を構築する．

今，s 番目の話者 ($1 \leq s \leq S$) に対して，$N(s)$ 発話がある場合を考える．元の $N(s)$ 発話を $D(s)$ 回分割し，$N'(s)(= N(s) + D(s))$ の数単語からなる文節に分割する．分割後にランダムに選んだ $W(s)$ 文節を結合し，$\lfloor N'(s)/W(s) \rfloor$ 文を作る．ここで，$\lfloor \cdot \rfloor$ はフロアリング関数である．$W(s)$ 文節は $N'(s)$ 文節中から選ばれ，組み合わせ数 N_c は，

$$N_c = {}_{N'}C_W \times {}_{N'-W}C_W \ldots = \prod_{i=0}^{\lfloor \frac{N'-W}{W} \rfloor} {}_{N'-Wi}C_W, \quad (1)$$

のようになる．ここで可読性のため，s を省略した．少なくとも，元の一文が再生されてしまう確率は $p_R = \frac{N'}{N_c}$ であり，これは，通常成り立つ $N' \gg W$ の条件下では，ほぼ零である．.

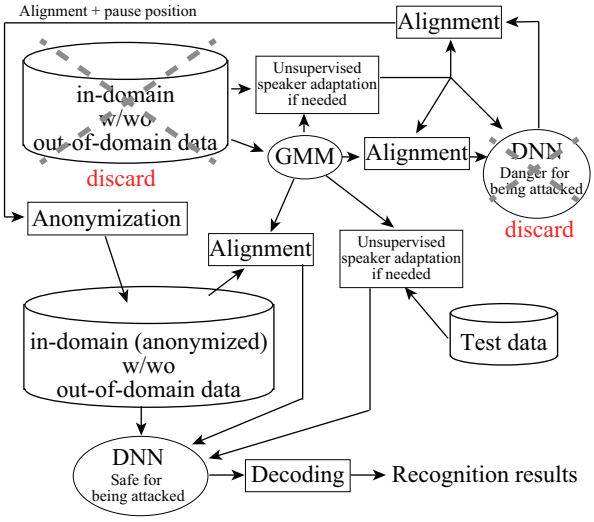

図 1 PPAMT のシステムの概要．GMM と DNN はそれぞれガウス混合モデル (Gaussian mixture model) と深層神経回路網 (deep neural network) のことである．

Fig. 1 System overview of PPAMT where GMM and DNN are the abbreviations of Gaussian mixture model and deep neural network, respectively.

以下の節では，上述の 3 種の特徴量に対する PPAMT の敏感さを分析する．

2.2 n-gram

まず，uni-gram は，全く変わらない．この枠組みで，uni-gram のプライバシーを保護したい場合には，該当の文節を削除する必要がある．bi-gram は，文節のはじめの部分と，分割部の右手側で 1 分割ごとに変化する．全部で $\sum_s D(s)$ 回の分割後に，影響を受ける確率は

$$p_{L_2} = \frac{2}{N_w} \sum_s D(s), \quad (2)$$

である．ここで，N_w は学習セットに含まれる総単語数である．この確率が PPAMT に対する特徴量の敏感さを示している．

tri-gram は，はじめの部分と分割部の右手側のそれぞれ 2 箇所が変化する．この時の確率は，

$$p_{L_3} = \frac{4}{N_w} \sum_s D(s), \quad (3)$$

である．

2.3 音素ラベル

mono-phone のラベルは変化しない．tri-phone のラベルは，4 箇所で変化する．すなわち，文節の冒頭，分割部の両側，そして文節の最後である．各ラベルが同じ長さであると仮定すると，影響を受ける部分の確率は，

$$p_{\pi_3} = \frac{4}{N_{\pi_3}} \sum_s D(s), \quad (4)$$

ⓒ 2019 Information Processing Society of Japan

である．ここで，tri-phone ラベルの数は，N_{π_3} である．

2.4 音響特徴量

音響特徴量は連続 $\pm\phi$ フレーム結合されて使われる．すなわち，各フレームに対して，$(2\phi+1)$ フレームにまたがる特徴量が使われる．学習データ中に全 N_F フレームある場合，使われる特徴量は $N_F(2\phi+1)$ フレームとなる．

音響特徴量は以下の 4 箇所で変化する．冒頭部の左側，分割部の両側，末尾の右側である．各箇所につき，$\sum_{\varphi=1}^{\phi}\varphi=\frac{(\phi+1)\phi}{2}$ フレームにまたがる特徴量が変化する．ゆえに，すべての分割により，$2(\phi+1)\phi\sum_s D(s)$ フレーム分の特徴量が変化する．確率は，

$$p_F=\frac{2(\phi+1)\phi}{N_F(2\phi+1)}\sum_s D(s). \tag{5}$$

である．

2.5 3 種の特徴量の関連

一般的に，式 (3)，(4)，(5) で表される影響の受けやすさには，$N_F > N_{\pi_3} \gg N_w$ の順序が成り立つため，$p_{L_3} \gg p_{\pi_3} > p_F$ の関係性がある．この関係性により，音素ラベルや音響特徴量は，PPAMT に対して，n-gram よりも影響を受けにくいといえる．これは音響モデルを学習するのによい性質である．なぜならば，これは n-gram は音素ラベルや音響特徴量よりも，よりランダム化されているといえるからである．攻撃者に元の書き起こしを再生させないためには，n-gram は十分にランダム化されている必要がある一方で，正確なモデルを学習するためには，音素ラベルや音響特徴量は正確でなければならないからである．

2.6 DNN の事前分布の出力プライバシー

tri-phone 状態 t の事前分布も PPAMT により変化する．元の事前分布 P と PPAMT を施した後の事前分布 P' の差異は，DP での場合 [16] と同様にして以下のように測ることができる．

$$\epsilon(t)=|\log(P(t))-\log(P'(t))|. \tag{6}$$

3. 話者秘匿化

2 節での要件に加えて，話者秘匿化は話者クラスタリングにより達成可能である．

3.1 i-vector に基づく話者クラスタリング

話者を秘匿するため，PPDP の手法の一つである k 匿名化を利用する．この手法では，学習話者や学習話者数を秘匿化することができる．異なる話者を同一のクラスターに混合し，すべての発話が複数話者の発話からなるようにすることで，話者特定手法に基づく攻撃に対して頑健性を持たせられる．まず，話者クラスタを i-vectors [19] に基づき構築する．i-vector は因子分析から導出されるもので，発話を話者/チャネルに不変の部分と可変の部分に $\boldsymbol{V}^n=\boldsymbol{v}+\boldsymbol{T}\boldsymbol{z}^n$ のようにわける．ここで，$\boldsymbol{V}^n$ はガウス混合モデル (Gaussian mixture model; GMM) のスーパーベクトルであり，発話 n に適応することで話者とチャネルに依存する．$\boldsymbol{v}$ も同じく GMM のスーパーベクトルであるが，話者とチャネルに非依存で，汎用背景モデルから得られる．$\boldsymbol{T}$ は低ランクの長方行列であり，全変数空間を張る基底からなる．$\boldsymbol{z}^n$ が発話 n に対する i-vector である．全 N 発話を k-means アルゴリズムにより，$\boldsymbol{z}^n$ のコサイン類似度に基づきクラスタリングし，話者を秘匿化する．

3.2 ランダム結合

クラスタリング後に，c 番目の話者クラスタには，$\sum_{s\in\mathcal{S}(c)}N(s)/C$ 発話，すなわち $\sum_{s\in\mathcal{S}(c)}N'(s)/C$ 文節が存在する．ここで $\mathcal{S}(c)$ は c 番目のクラスタに所属する話者の集合である．これらの文節はランダムに結合される．これにより，学習データ中の話者数 S を意図する数 C に調整することができる．各クラスタ数の話者数が 1 以上になるようにすれば，k 匿名化が達成できる．k は同一クラスタに所属する話者数の最小数である．各クラスタに話者が同一数含まれれば，k は $\lfloor S/C \rfloor$ である．ランダムに結合された発話の i-vector は平均的には話者クラスタのセントロイドとなるので，話者特定に対して頑健である．2 節での PPAMT と比して，異なる言語文脈を混ぜられ，S/C 倍に言語複雑性を大きくすることができる．

3.3 話者適応ではプライバシーは保護されない

話者適応ではプライバシー保護には十分でない．例えば，典型的な適応手法である特徴空間最尤線形回帰 (feature-space maximum likelihood linear regression; fMLLR) [20] を例にとる．fMLLR では，音響特徴量ベクトル $\boldsymbol{x}$ を変換し，s 番目の話者に適応した特徴量 $\boldsymbol{y}$ を得る．この時に，変換行列 $\boldsymbol{A}_s$ とバイアス $\boldsymbol{b}_s$ を $\boldsymbol{y}=\boldsymbol{A}_s\boldsymbol{x}+\boldsymbol{b}_s$ のように適用する．これではプライバシーは保護できない．もし $\boldsymbol{A}_s$ と $\boldsymbol{b}_s$ が学習後に破棄されていたとしても，これらを推定することは可能である．なぜならば，適応に使うための GMM は，未知のテスト話者に対して変換行列を推定するために保存しておく必要があるからである．対象話者の発話が得られれば，これらのパラメータ $\hat{\boldsymbol{A}}_s$ と $\hat{\boldsymbol{b}}_s$ が得られるので，元の特徴量 $\hat{\boldsymbol{x}}$ がそれらの逆行列から推測されてしまう．$\hat{\boldsymbol{x}}=\hat{\boldsymbol{A}}_s^{-1}[\boldsymbol{y}-\hat{\boldsymbol{b}}_s]$. このようになれば，話者特定が可能である．これは $\boldsymbol{A}_s$ の条件数が小さいときには特に高精度である．

4. 実験

4.1 実験条件

日本語話し言葉コーパス (Corpus of Spontaneous

ex 1: 仕事の / その情報の / えー / 四番目と / 高いと /
誤り率は / おります / 人に / 調べ物を / 結果を / 現在までに
ex 2: すぐ検索して / えー / 本発表は / 納めまして /
だから / 途中で / えー / 最も一致が / おー / の / います /
基づくフィードバックだと / 認知活動の

図 2　ランダムに結合された文節の例
Fig. 2　Examples of randomly concatenated phrases.

Japanese; CSJ) [21] を用いて PPAMT の有効性を検証した．CSJ は最も広く用いられている日本語の LVCSR タスクである．語彙数は約 70k である．ここでは Kaldi toolkit [22] の “nnet1” 実装と付属の CSJ レシピにより，ベースラインシステムを構築した．音響特徴量は，13 次元のメル周波数ケプストラム係数 (mel-frequency cepstral coefficient; MFCC) を線形判別解析により変換して得られた 40 次元の音響特徴量を連続 $\pm\phi(=17)$ フレーム結合したものとした．fMLLR による教師なし話者適応を適用した．DNN は 7 層 (各層 1,905 ノード) からなり，9,388 の出力ノード (tri-phone 状態) を持つ．

CSJ の中には 2 つのドメインがある．ここでドメイン内データとして扱うのは，学術講演 (CSJ A) であり，ドメイン外データは一般講演とインタビュー (CSJ R&S) である．10 名の異なる話者からなるオープンな CSJ A テストセット 10 講演を単語誤り率 (word error rate; WER) [%] の観点で評価した．デコード時，tri-gram 言語モデルは，ドメイン内データから学習したものを，全システムで共通に用いた．ドメイン内の学習データ (CSJ A set) は，もともと $\sum_s N(s) = 159,297$ 文章 ($N_F = 85,999,942$ フレーム (239 hours) からなる) を含む．話者数は $S = 986$ である．全部で，$N_w = 3,871,539$ 単語 (41,862 異なり単語) があり，$N_{\pi_3} = 12,004,648$ の tri-phone ラベルが付けられている．$\sum_s D(s) = 952,346$ 分割ののち，$\sum_s N'(s) = 1,111,643$ 文節が得られた．この実験では，文章は短いポーズや助詞ごとに文節に分割した．分割前は，各文は平均 $N_w/\sum_s N(s) = 24.3$ 単語を含む．分割後は，各文節は平均 $N_w/\sum_s N'(s) = 3.48$ 単語を含むため，$W = 10$ 単語をランダムに結合することで，111,509 文を生成した．4.5 節での実験では，ドメイン外の学習データ (CSJ R&S set) として，話者数 2,222，フレーム数 101,208,464(281 hours) のデータを使った．

4.2　PPAMT

図 2 はランダムに結合された文節の例である．これを見るとほぼ言語内容は失われていることが分かる．スラッシュマークは，文節の境界を示し，各分節は数単語からなる．各分節は平均的に $N_F/N = 77.4$ フレーム (0.774 [sec]) の継続長となった．この場合，影響の受けやすさは，$p_{L_3} = 0.984 \gg p_{\pi_3} = 0.317 > p_F = 0.194$ のようになり，2.5 節で示した関係性が満たされることが分かった．

表 1　ドメイン内のデータしか利用できない場合に，提案のプライバシー保護を施した場合の WER[%]
Table 1　WER[%] of the proposed privacy preservation where only in-domain dataset was available.

	CE	sMBR
all in-domain data available	11.71	11.05
phrase division	15.43	14.44
random concatenation	14.88	13.76
speaker anonymization (10 clusters)	15.09	14.17

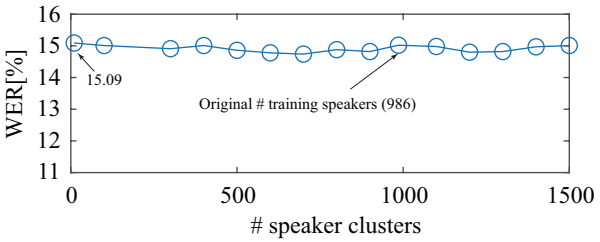

図 3　話者匿名化のために話者クラスタリングを利用した場合の CSJ testset での WER[%]
Fig. 3　WER[%] on the CSJ testset when speaker clustering was used for speaker anonymization.

表 1 は，ドメイン内データセットのみが利用可能な場合の WER である．クロスエントロピー (cross-entropy; CE) DNN 音響モデルが得られたのち，系列ベイズリスク最小化 (sequential minimum Bayes risk; sMBR) 識別学習 [9] を行った．第 1 行目はプライバシー保護なしに全部のドメイン内話者を使った場合の上限性能を示す．第 2 行目は文節への分割のみを行った場合である．数単語を含むだけの文節であって継続長が短くても，音響モデルの学習自体は行えている．文節のランダム結合により tri-phone の多様性が増すことで，性能が向上した．特徴量の時系列が保存されていることから，sMBR は PPAMT に対しても有効であり，これは提案の PPAMT の利点であるといえる．10 クラスタによる話者匿名化により，CE 学習時に 0.2%，sMBR 学習時に 0.4%の WER 低下が見られたが，これにより，98-匿名化が達成できている．

4.3　話者クラスタ数

図 3 には話者クラスタ C と WER の関係を示す．$C < S$ の場合，$C > S$ の場合いずれも，性能はほとんどクラスタ数に依存せず，任意の k に対する話者匿名化が達成できた．

4.4　ドメイン内データでプライバシー保護の必要ない話者が部分的に利用可能な場合

これとは別に，サブサンプリングにより，個人データの永続的利用に同意した話者のみを利用することで，部分的にドメイン内データを利用することも考えられる．図 4 には，プライバシー保護されない学習話者数と WER の関係を示す．プライバシー保護されない学習話者数が 100 名未満の場合，性能が顕著に低下した．プライバシー保護さ

ⓒ 2019 Information Processing Society of Japan

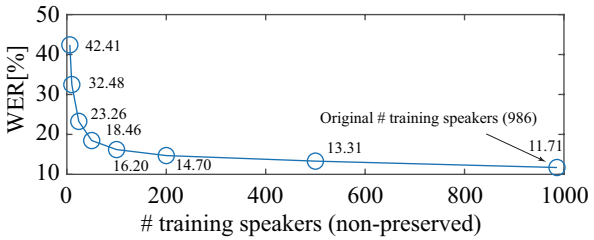

図 4　学習データを間引いた場合の CSJ testset での WER[%]

Fig. 4　WER[%] on the CSJ testset when training data were subsampled.

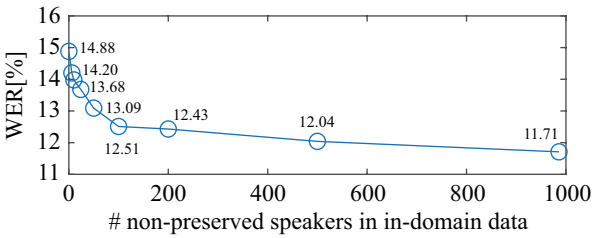

図 5　部分的に話者のプライバシー保護を行った場合の CSJ testset での WER[%]

Fig. 5　WER[%] on the CSJ testset when partial speakers were preserved.

れない話者が 200 名 (全体のおおよそ 1/5) の場合，3%の WER 低下が見られた．

図 5 では，PPAMT でプライバシー保護されない話者数を変えた場合の WER を示す．200 名プライバシー保護されない話者がいる場合，0.7%の WER の低下が見られた．これに対して，同条件でサブサンプリングの場合には 3%，WER が低下している．

プライバシー保護されない話者数が少なくなるにつれ，性能の劣化はサブサンプリングよりも小さくなる．これにより，PPAMT はサブサンプリングよりも，プライバシー保護なしで部分的にドメイン内データが使える場合よりも優れることがわかった．

4.5　他のドメイン外データセットが利用可能な場合

表 2 には，ドメイン外データが付加的に利用可能な場合の WER を示す．特に PPAMT において，付加的なドメイン外データは有効である．これはドメインによらない知識が利用できるためと考えられる．一方でドメイン外データののみしか利用できない場合には，WER は 14.14%であり，これらよりも著しく悪い．これにより，プライバシー保護されたドメイン内データは著しく性能を向上させた．この場合，4.3 節に示すように，話者匿名化は性能を低下させなかった．

図 6 は，ドメイン内データセットのプライバシー保護されない話者数と WER の関係を示す．プライバシー保護されない話者がいない場合でさえ，WER の低下は 0.59%である．200 名のプライバシー保護されていない話者がいる場合には，WER の低下は 0.26%である．

表 2　他のドメイン外データセットが利用可能な場合に，提案のプライバシー保護を施した場合の WER[%]．() には，表 1 からの改善量を示す

Table 2　WER[%] of the proposed privacy preservation where additional out-of-domain dataset was available. () shows the improvement from Table 1.

	CE
all in-domain data available	11.44 (0.27)
random concatenation	12.03 (2.85)
speaker anonymization (10 clusters)	11.98 (3.11)
cf. only out-of-domain data available	14.14

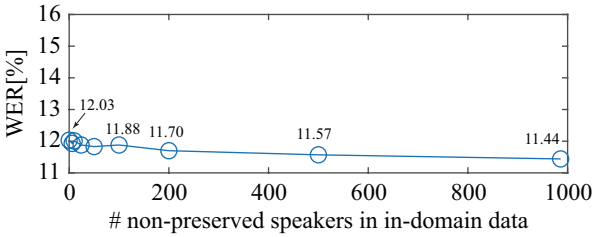

図 6　他のドメイン外データセットとともに，部分的に話者に提案のプライバシー保護を施した場合の WER[%]

Fig. 6　WER[%] on the CSJ testset when partial speakers were preserved with out-of-domain data.

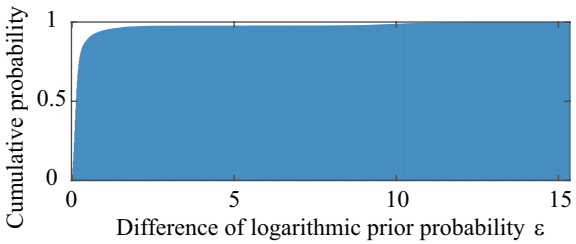

図 7　式 (6) の ϵ で表されるトライフォン状態の事前分布の対数確率の差異

Fig. 7　Difference of logarithmic probabilities of the prior distribution for tri-phone states, ϵ in Eq. (6).

4.6　事前分布の出力プライバシー

図 7 に，式 (6) に示す事前分布の累積確率を示す．全状態の 90%の差異が，$\epsilon = 0.5$ を下回っており，$\epsilon = 2$ を超えるのは，全体の 2.7%の状態である．これにより，ほとんどの事前分布が PPAMT により変化しておらず，PPAMT により事前分布のプライバシー保護がよく実現されていることが分かる．

5.　まとめと今後の課題

本報では，プライバシー保護音響モデル学習法 (privacy preserving acoustic model training; PPAMT) を提案した．“perturbation” と “randomization” がカギとなる操作である．3 種の特徴量 (n-gram，音素ラベル，音響特徴量) に対して，PPAMT により影響を受ける確率，すなわち PPAMT に対する敏感さ，を定式化した．これにより，音響特徴量や音素ラベルは言語特徴量よりも PPAMT の影響を受けにく

ⓒ 2019 Information Processing Society of Japan

いことが分かった．これは個人情報を保護して音響モデルの学習を行うのには良い性質である．これに加えて，話者クラスタリングにより話者匿名化を達成できた．PPAMT による WER の悪化は 0.6%未満であり，この時，書き起こしが再生される確率は無視できるほど小さい．話者匿名化は，ドメイン外データを使った際には，性能を低下させなかった．今後の課題は，提案の PPAMT に，特に出力プライバシーの観点から理論的な解析を加えることである．

参考文献

[1] Bocchieri, E., Riley, M. and Saraclar, M.: Methods for Task Adaptation of Acoustic Models with Limited Transcribed In-Domain Data, *Proceedings of INTERSPEECH*, pp. 326–329 (2004).
[2] Kapralova, O., Alex, J., Weinstein, E., Moreno, P. and Siohan, O.: A Big Data Approach to Acoustic Model Training Corpus Selection, *Proceedings of INTERSPEECH*, pp. 2083–2087 (2014).
[3] Agrawal, R. and Srkant, R.: Privacy-preserving Data Mining, *Proceedings of Special Interest Group on Management of Data (SIGMOD)*, pp. 439–450 (2000).
[4] Lindell, Y. and Pinkas, B.: Privacy Preserving Data Mining, *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (CRYPTO)*, pp. 36–54 (2000).
[5] Lambert, D.: Measure of Disclosure Risk and Harm, *Journal of Official Statistics*, Vol. 9, No. 2, pp. 313–331 (1993).
[6] Smaragdis, P. and Shashanka, M.: A Framework for Secure Speech Recognition, *IEEE Transactions on Audio, Speech, Language Processing*, Vol. 15, No. 4, pp. 1404–1413 (2007).
[7] Pathak, M. A., Raj, B., Rane, S. and Smaragdis, P.: Privacy-Preserving Speech Processing, *IEEE Signal Processing Magazine*, pp. 62–74 (2013).
[8] Povey, D.: *Discriminative Training for Large Vocabulary Speech Recognition*, PhD thesis, Cambridge University Engineering Dept (2003).
[9] Veselý, K., Ghoshal, A., Burget, L. and Povey, D.: Sequence-discriminative Training of Deep Neural Networks, *Proceedings of INTERSPEECH*, pp. 2345–2349 (2013).
[10] Graves, A. and Jaitly, N.: Towards End-to-end Speech Recognition with Recurrent Neural Networks, *Proceedings of the 31st International Conference on Machine Learning*, pp. 1764–1772 (online), available from ⟨http://proceedings.mlr.press/v32/graves14.pdf⟩ (2014).
[11] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *Proceedings of ACM Conference on Computer and Communications Security (CCS)* (2015).
[12] Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B. and Sweeney, L.: Privacy Preserving Synthetic Data Release Using Deep Learning, *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2018).
[13] Shah, A. and Gulati, R.: Privacy Preserving Data Mining: Techniques, Classification and Implications -A Survey, *International Journal of Computer Applications*, Vol. 137, No. 12 (2016).
[14] Samrati, P. and Sweeny, L.: Generalizing Data to Provide Anonymity When Disclosing Information, *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (PODOS)*, p. 188 (1998).
[15] Sweeney, L.: k-anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp. 557–570 (2002).
[16] Dwork, C.: Differential Privacy, *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, Lecture Notes in Computer Science (2006).
[17] Sarwate, A. D. and Chaudhuri, K.: Signal Processing and Machine Learning with Differential Privacy, *IEEE Signal Processing Magazine*, pp. 86–94 (2013).
[18] Joy, J., Gray, D., McGoldrick, C. and Gerla, M.: K Privacy: Towards Improving Privacy Strength While Preserving Utility, *Ad Hoc Networks*, pp. 16–30 (2018).
[19] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).
[20] Gales, M.: Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, *Computer Speech and Language*, Vol. 12, pp. 75–98 (1998).
[21] Furui, S., Maekawa, K. and Isahara, H.: A Japanese national project on spontaneous speech corpus and processing technology, *Proceedings of ASR*, pp. 244–248 (2000).
[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Petr, M., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K.: The Kaldi Speech Recognition Toolkit, *Proceedings of ASRU*, pp. 1–4 (2011).

ⓒ 2019 Information Processing Society of Japan