音響モデル構築用音声データのプライバシー保護*

○太刀岡 勇気 (デンソーアイティーラボラトリ)

1 はじめに

音響モデルを学習する際にはたとえ少量でも、ドメイン内データが有効である [1] が、ドメイン内データは、利用期間終了後は破棄されることが多い。ただし、一旦学習データが削除されてしまうと、モデルを再学習できないため、データを削除せず、個人情報を保護できる技術が必要である。これは privacy preserving data mining(PPDM) [2, 3] 問題の一つで、特定・暴露リスクを低減する [4]。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 に関する研究は多くない。計算方法の手順を秘匿化する方法 [5, 6] は、非秘匿な場合に比べ多くの計算量を必要とし、モデル変更の際には操作のプロトコルを変更する必要がある。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 [7] や end-to-end の手法 [8] を使うことはできない。

また、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おり [9]、ドメイン内データで学習した深層神経回路網 (DNN) は攻撃の危険にさらされる。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。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 [10] では、様々な PPDM 手法を分類している (文献 [10] の表 1)。これによれば、我々の方法は “perturbation”、“randomization”、“anonymization” に分類される。

PPAMT では、発話を文節に分割し、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 [11, 12] 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。

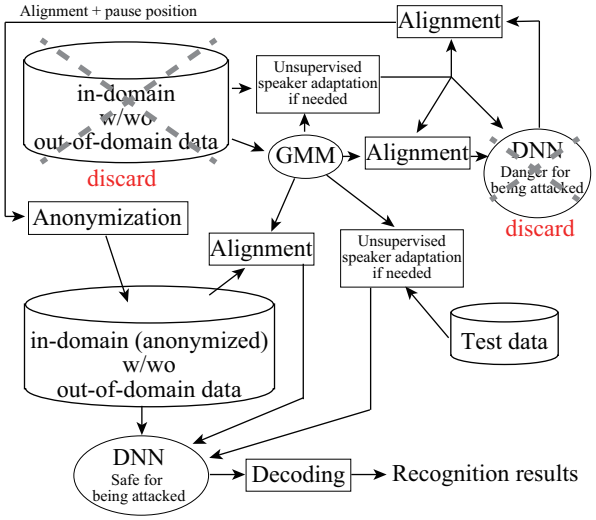

Fig. 1 System overview of PPAMT where GMM and DNN are the abbreviations of Gaussian mixture model and deep neural network, respectively.

2 PPAMT

図 1 に、「はじめに」に記した要求条件を満足する PPAMT の枠組みを示す。PPAMT の基本の操作は “perturbation” と “randomization” [10] である。文を短いポーズや文節境界で、文節に分割し、これらをランダムに結合して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。$W(s)$ 文節は $N'(s)$ 文節中から選ばれ、組み合わせ数 N_c は、$N_c = {}_{N'(s)}C_{W(s)} \times {}_{N'(s)-W(s)}C_{W(s)}$ のようになる。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。. 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。

2.1 n-gram

uni-gram は、不変である。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

$$p_{L_2} = \frac{2}{N_w} \sum_s D(s), \tag{1}$$

*Privacy Preservation of Speech Data for Constructing Acoustic Models. by TACHIOKA, Yuuki (Denso IT Laboratory)

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、

$$p_{L_3} = \frac{4}{N_w} \sum_s D(s), \tag{2}$$

である。

2.2 音素ラベル

mono-phone ラベルは変化しない。tri-phone ラベルは、4 箇所変化する。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、

$$p_{\pi_3} = \frac{4}{N_{\pi_3}} \sum_s D(s), \tag{3}$$

である。ここで、tri-phone ラベルの数は、N_{π_3} である。

2.3 音響特徴量

音響特徴量は連続 $\pm\phi$ フレーム結合される。すなわち、各フレームで、$(2\phi+1)$ 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、

$$p_F = \frac{2(\phi+1)\phi}{N_F(2\phi+1)} \sum_s D(s). \tag{4}$$

である。

2.4 3 種の特徴量の関連

一般的に、式 (2)、(3)、(4) で表される確率には、$p_{L_3} \gg p_{\pi_3} > p_F$ の関係がある[1]。これより、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。元の書き起こしの再生を防ぐには、n-gram は十分ランダム化されている必要がある一方、正確なモデル学習のためには、音素ラベルや音響特徴量は正確な必要があるからである。

3 話者秘匿化

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 [13] 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。

4 実験

4.1 実験条件

日本語話し言葉コーパス (CSJ) を用いて PPAMT の有効性を検証した。語彙数は約 70k である。ここでは Kaldi toolkit の “nnet1” 実装と付属の CSJ レシピにより、ベースラインを構築した。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。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 には 2 つのドメインがあり、学術講演 (CSJ A) をドメイン内データとして扱い、一般講演とインタ

[1] $N_F > N_{\pi_3} \gg N_w$ の順序が成り立つ。

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
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

Fig. 2 Examples of randomly concatenated phrases.

ビュー (CSJ R&S) をドメイン外データとした。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 (WER) [%] の観点で評価した。デコード時の tri-gram 言語モデルは、ドメイン内データから学習した。ドメイン内の学習データは、$\sum_s N(s) = 159,297$ 文章 ($N_F = 85,999,942$ フレーム (239 時間)) を含む。話者数 S は 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。4.4 節での実験では、ドメイン外の学習データとして、2,222 話者、101,208,464 フレーム (281 時間) 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。これより、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 秒) の継続長となった。この場合、$p_{L_3} = 0.984 \gg p_{\pi_3} = 0.317 > p_F = 0.194$ で、2.4 節で示した関係性が満たされることが分かった。

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 (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 (sMBR) 識別学習 [7] 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。

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

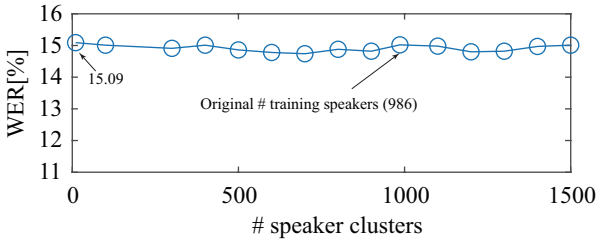

Fig. 3 WER[%] on the CSJ testset when speaker clustering was used for speaker anonymization.

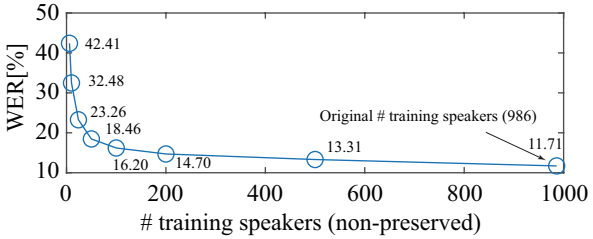

Fig. 4 WER[%] on the CSJ testset when training data were subsampled.

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。

4.3 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、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。プライバシー非保護話者数が 100 名未満の場合、性能が顕著に低下した。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 でプライバシー非保護話者数を変えた場合の WER を示す。200 名の場合、0.7%の WER の低下が見られた。これに対して、同条件のサブサンプリングの場合、3%WER が低下した。

プライバシー非保護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。

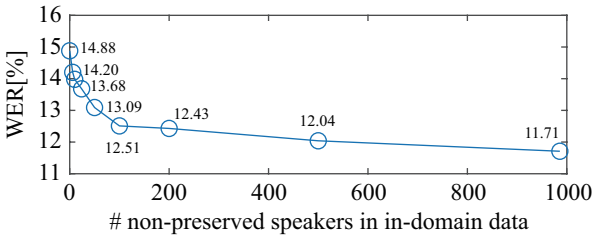

Fig. 5 WER[%] on the CSJ testset when partial speakers were preserved.

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

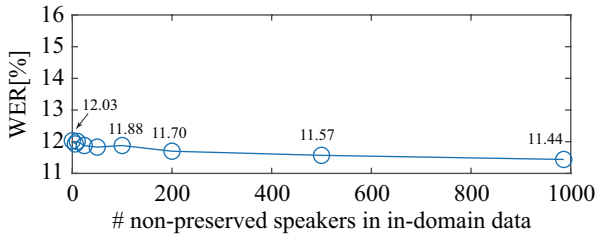

Fig. 6 WER[%] on the CSJ testset when partial speakers were preserved with out-of-domain data.

4.4 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。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、上と同様、話者匿名化は性能を低下させなかった。

図 6 のは、ドメイン内データセットのプライバシー非保護話者数と WER の関係を示す。プライバシー非保護話者がいない場合でさえ、WER の低下は 0.59%である。200 名のプライバシー非保護話者がいる場合には、WER の低下は 0.26%である。

5 まとめ

本報では、PPAMT を提案し、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。

参考文献

[1] E. Bocchieri, M. Riley, and M. Saraclar, “Methods for task adaptation of acoustic models with limited transcribed in-domain data,” Proceedings of INTERSPEECH, pp.326–329 (2004).

[2] R. Agrawal and R. Srkant, “Privacy-preserving data mining,” Proceedings of Special Interest Group on Management of Data (SIGMOD), pp.439–450 (2000).

[3] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (CRYPTO), pp.36–54 (2000).

[4] D. Lambert, “Measure of disclosure risk and harm,” Journal of Official Statistics, **9**, 313–331 (1993).

[5] P. Smaragdis and M. Shashanka, “A framework for secure speech recognition,” IEEE Transactions on Audio, Speech, Language Processing, **15**, 1404–1413 (2007).

[6] M.A. Pathak, B. Raj, S. Rane, and P. Smaragdis, “Privacy-preserving speech processing,” IEEE Signal Processing Magazine, 62–74 (2013).

[7] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” Proceedings of INTERSPEECH, pp.2345–2349 (2013).

[8] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” Proceedings of the 31st International Conference on Machine Learning, pp.1764–1772 (2014). http://proceedings.mlr.press/v32/graves14.pdf

[9] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” Proceedings of ACM Conference on Computer and Communications Security (CCS) (2015).

[10] A. Shah and R. Gulati, “Privacy preserving data mining: Techniques, classification and implications -a survey,” International Journal of Computer Applications, **137** (2016).

[11] P. Samrati and L. Sweeny, “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).

[12] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10**, 557–570 (2002).

[13] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, **19**, 788–798 (2011).