Privacy Preserving Acoustic Model Training for Speech Recognition

Yuuki Tachioka
* Denso IT Laboratory, Tokyo, Japan
E-mail: ytachioka@d-itlab.co.jp

***Abstract*—In-domain speech data significantly improve the speech recognition performance of acoustic models. However, the data may contain confidential information and exposure of transcriptions may lead to a breach in speakers' privacy. In addition, speaker identification can be problematic when speakers want to hide their membership of a certain group. Thus, the in-domain data must be deleted after its period of use. However, once the data are deleted, models cannot be updated for future architectures. Privacy preservation is necessary when retaining speech data; it is important that the transcriptions cannot be reconstructed and the speaker cannot be identified. This paper proposes a privacy preserving acoustic model training (PPAMT) method that satisfies these requirements and formulates the sensitivities of three features (n-grams, phoneme labels, and acoustic features) for PPAMT. A sensitivity analysis showed that phoneme labels and acoustic features were less susceptible to PPAMT than n-grams, which is optimal because accurate phoneme labels and acoustic features are needed for acoustic model training. Speech recognition experiments showed that the word error rate degradation by PPAMT was less than 0.6% as a result of this property.**

I. Introduction

When training acoustic models for automatic speech recognition (ASR), in-domain data are effective, even if the amount is small [1], [2]. However, transcriptions of in-domain data tend to contain highly confidential information, and speaker identification can also be problematic when speakers want to hide their membership of a certain group. Thus, in-domain data must be discarded after its period of use, but once the data are deleted, the model cannot be updated for model architectures proposed in the future. Continuous use of in-domain data with privacy preservation is desirable. One of the goals of privacy preserving data mining (PPDM) [3], [4] is to reduce the risk of the identification and disclosure [5]. PPDM is necessary to prevent the disclosure of speech data and personal identities.

Few studies have been conducted on PPDM in the field of speech processing. One approach is a secure calculation [6], [7] but it requires more computation than a non-secure calculation. In addition, its operation protocol must be also modified when changing models, and it cannot be used for data preservation.

Another approach is data shuffling, which deletes personal information, but sequential discriminative training [8], [9] or an end-to-end approach [10] cannot be used; because the time sequence of acoustic features is completely lost. The time sequence of acoustic features must be preserved in order to use the aforementioned approaches.

Recently, restoring training data from trained models has become feasible [11]. Training data cannot be restored from conventional Gaussian mixture models (GMM) because they only retain the mean and variance of the acoustic features of the training data. Meanwhile, deep neural network (DNN) models trained using the in-domain data are vulnerable to attacks and they need to be trained using the privacy-preserved dataset.

As mentioned above, for the PPDM of speech data, original in-domain data must be deleted so that personal information cannot be restored from the anonymized dataset. At the same time, the time sequence of the privacy-preserved dataset must be retained. The use of real data is also important for acoustic model training. Although a deep autoencoder can be used to generate training data [12], generalization models cannot sufficiently represent speech dynamics [13]. This paper proposes a privacy preserving acoustic model training (PPAMT) method to satisfy these requirements. A PPDM survey paper [14] classified various PPDM techniques (Table 1 in [14]). According to this classification, our framework is related to perturbation, randomization, and anonymization techniques.

Privacy can be categorized into two types: input privacy and output privacy. Input privacy adds noise to data such as in privacy preserving data publishing (PPDP). PPAMT mainly preserves input privacy. Utterances are divided into word phrases which are then randomly concatenated to further randomize transcriptions. We formulated the probability of influence from PPAMT for three types of features: n-grams, phoneme labels, and acoustic features. In addition, to anonymize speakers, we used k-anonymization [15], [16] based on speaker clustering, which prevents attackers from specifying the target speaker from more than k candidates.

The output privacy of PPAMT appears in the prior distribution of the tri-phone states, which is used to convert DNN outputs into posteriors. The difference of prior distributions between before and after PPAMT is evaluated similarly to differential privacy (DP) [17] by perturbation [18] or random sampling [19].

The remainder of this paper is as follows. Section II proposes PPAMT, which uses perturbation and randomization, and analyzes the sensitivity of the above three features. Section III describes how to anonymize speakers by using speaker clustering. Experiments described in Sec. IV show the effectiveness of the proposed PPAMT for a large-vocabulary continuous speech recognition (LVCSR) task.

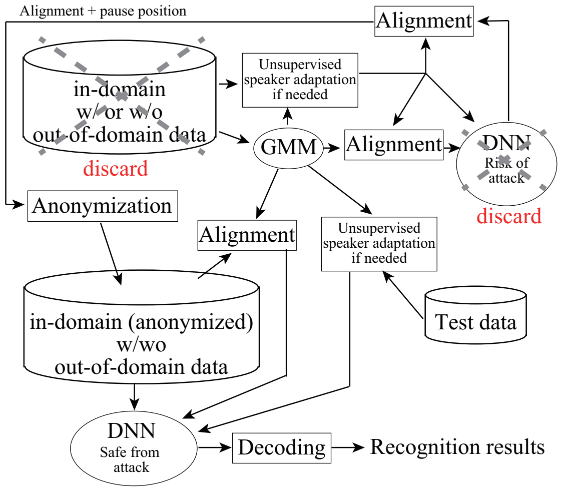

Fig. 1. System overview of PPAMT

II. Privacy preserving acoustic model training (PPAMT)

A. Overview

We propose a PPAMT framework shown in Fig. 1, which satisfies the requirements mentioned in the introduction. The requirements are to delete personal information while keeping the time sequence of features. The basic operations of PPAMT are perturbation and randomization [14]. First, sentences are divided into word phrases based on short pauses or boundaries between the phrases. Then, the phrases are randomly concatenated to construct new sentences.

There are $N(s)$ utterances for the s-th speaker ($1 \leq s \leq S$) where the total number of training speakers is S. The original $N(s)$ utterances are divided by $D(s)$ times into $N'(s)$ word phrases composed of a few words, i.e., $N'(s) = N(s) + D(s)$. After division, randomly selected $W(s)$ word phrases are concatenated to make $\lfloor N'(s)/W(s) \rfloor$ sentences, where $\lfloor \cdot \rfloor$ is a flooring function. Because $W(s)$ phrases are selected from $N'(s)$ phrases, the total number of combinations N_c is

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

where s is omitted for readability. At least one of the original sentence can be restored with probability $p_R = N'/N_c$, which is almost zero when $N' \gg W$. The subsections below describe the sensitivity of three types of features to PPAMT.

B. n-grams

Uni-grams do not change. In this framework, uni-gram privacy can be attained by dropping the corresponding word phrases. Bi-grams change at the beginning and the right-hand side at each division. The probability of being influenced after the total $\sum_s D(s)$ divisions is

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

where N_w is the total number of words in the training dataset. This probability indicates the sensitivity of bi-grams to PPAMT.

Tri-grams change at the beginning and the right-hand side at each division. The probability is

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

C. Phoneme labels

Mono-phone labels do not change. Tri-phone labels change at four parts: the beginning, both sides of the division, and the end. The probability under the assumption that each label has the same duration is

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

where the total number of tri-phone labels is N_{π_3}.

D. Acoustic features

Acoustic features are concatenated in contiguous $\pm\phi$ frames, i.e., for each frame, features over $(2\phi + 1)$ frames are used. When there are a total of N_F frames in the training data, the features used have $N_F(2\phi + 1)$ frames.

Acoustic features change at four parts, i.e., the left-hand part of the beginning, both sides of the division, and the right-hand side of the end. For each part, features over $\sum_{\varphi=1}^{\phi} \varphi = \frac{(\phi+1)\phi}{2}$ frames change. Thus, the division changes the features in $2(\phi + 1)\phi \sum_s D(s)$ frames. The probability is

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

E. Relation of three types of features

In general, the sensitivities in Eqs. (3), (4), and (5) can be expressed by the relation: $p_{L_3} \gg p_{\pi_3} > p_F$ because of the order $N_F > N_{\pi_3} \gg N_w$. This relationship indicates that phoneme labels and acoustic features are less susceptible to PPAMT than n-grams. This is optimal for training acoustic models because this means that n-grams are more randomized than phoneme labels and acoustic features. N-grams must be randomized sufficiently in order to prevent others from restoring the original transcriptions, whereas phoneme labels and acoustic features must be accurate in order to train accurate models on the dataset.

F. Output privacy of prior distribution of DNN

The prior distribution of tri-phone states t is also changed by PPAMT. The discrepancy between the original prior distribution P and the prior distribution after PPAMT, P', can be measured similarly to DP [17], as

$$\epsilon(t) = |\log(P(t)) - \log(P'(t))|. \qquad (6)$$

III. Speaker anonymization

In addition to the requirements in Sec. II, speaker anonymization can be achieved by speaker clustering.

A. Speaker clustering based on i-vector

k-anonymization, which is one of the PPDP techniques, is used to anonymize speakers. This technique can anonymize training speakers and the number of the training speakers. After training speakers are divided into multiple clusters, the utterances of k-different speakers belonging to the same cluster are mixed. If all utterances are composed of multiple speakers, the training data are robust against attacks based on speaker identification techniques. First, speaker clusters are created based on i-vectors [20]. i-vectors are derived from a factor analysis that decomposes speech into a speaker/channel invariant and a variant, i.e., $\boldsymbol{V}^n = \boldsymbol{v} + \boldsymbol{T}\boldsymbol{z}^n$ where $\boldsymbol{V}^n$ is a GMM super vector adapted for an utterance n and depends on a speaker and a channel; $\boldsymbol{v}$ is a GMM super vector, which is independent of the speaker and channel and is obtained from a universal background model; $\boldsymbol{T}$ is a low-rank rectangular matrix composed of basis vectors that span all variable spaces, and $\boldsymbol{z}^n$ is an i-vector for the utterance n. All N utterances are clustered by k-means algorithm based on a cosine similarity of $\boldsymbol{z}^n$ to anonymize speakers.

B. Random concatenation

After C clusters are created from speaker clustering, the c-th speaker cluster contains $\sum_{s\in\mathcal{S}(c)} N(s)$ utterances, i.e., $\sum_{s\in\mathcal{S}(c)} N'(s)$ word phrases, where $\mathcal{S}(c)$ is the speaker group belonging to the c-th cluster. These phrases are randomly concatenated. This adjusts the number of speakers in the training data, S, to the desired number of clusters C. If the number of speakers in every cluster is greater than or equal to k, k-anonymization is achieved, i.e., k is the minimum number of speakers in the same cluster. If each cluster has the same number of speakers, k is $\lfloor S/C \rfloor$. i-vectors in randomly concatenated utterances are the average of multiple speakers' i-vectors, which are robust against speaker identification. Compared with the PPAMT in Sec. II, it is possible to mix different language contexts and increase the language perplexity in the ratio S/C.

C. Speaker adaptation does not preserve privacy

Speaker adaptation is insufficient for preserving privacy. For example, feature-space maximum likelihood linear regression (fMLLR) [21] is a typical technique that transforms an acoustic feature vector $\boldsymbol{x}$ to obtain the feature vector $\boldsymbol{y}$ adapted to the s-th speaker by applying transformation matrix $\boldsymbol{A}_s$ and bias $\boldsymbol{b}_s$ as $\boldsymbol{y} = \boldsymbol{A}_s\boldsymbol{x} + \boldsymbol{b}_s$. If $\boldsymbol{A}_s$ and $\boldsymbol{b}_s$ are discarded after training to preserve privacy, they can still be estimated; because GMM models for speaker adaptation must be reserved to estimate transform matrices for unknown test users. These parameters can be estimated from target speaker utterances as $\hat{\boldsymbol{A}}_s$ and $\hat{\boldsymbol{b}}_s$. An original feature $\hat{\boldsymbol{x}}$ can be estimated from the inverse as $\hat{\boldsymbol{x}} = \hat{\boldsymbol{A}}_s^{-1}[\boldsymbol{y} - \hat{\boldsymbol{b}}_s]$., enabling speaker identification. This identification is accurate especially when the condition number of $\boldsymbol{A}_s$ is small.

IV. Experiments

A. Experimental setups

We validated PPAMT using the Corpus of Spontaneous Japanese (CSJ) [22], one of the most widely used LVCSR tasks for building Japanese ASR systems. The vocabulary size is about 70k. We used DNN training tools from the Kaldi toolkit [23] with an attached recipe to construct a baseline system. Acoustic features were transformed from a 13-dimensional mel-frequency cepstral coefficient with linear discriminant analysis to obtain 40-dimensional acoustic features concatenated in contiguous $\pm\phi(=17)$ frames. fMLLR-based unsupervised speaker adaptation was applied. The DNN was composed of seven layers with $9,388$ output nodes (triphone states) and each layer had $1,905$ nodes.

The CSJ dataset contained two domains. The in-domain data were academic lectures (CSJ A) and the out-of-domain data were general lectures and interviews (CSJ R&S). We evaluated the open CSJ A test set composed of ten lectures by ten different speakers in terms of word error rate (WER) [%]. For decoding, a tri-gram language model was constructed from the in-domain data, which was common to all systems. The in-domain training data (CSJ A set) originally contained $\sum_s N(s) = 159,297$ sentences in $N_F = 85,999,942$ [frames] (239 hours). The number of speakers was $S = 986$. In total, there were $N_w = 3,871,539$ words ($41,862$ different words) with $N_{\pi_3} = 12,004,648$ tri-phone labels, in which there were 41 types of monophones and $41^3 = 68,921$ types of triphones. After $\sum_s D(s) = 952,346$ divisions, $\sum_s N'(s) = 1,111,643$ word phrases were obtained. In this experiment, sentences were divided according to the occurrence of short pauses or Japanese post-positional particles. Before division, each sentence had $N_w / \sum_s N(s) = 24.3$ words on average. After division, each phrase contained $N_w / \sum_s N'(s) = 3.48$ words on average, and $W = 10$ word phrases were randomly concatenated to construct $111,509$ sentences. In addition, for the experiment in Sec. IV-E, we used the out-of-domain training data (CSJ R&S set) that contained 2,222 speakers and 101,208,464 [frames] (281 hours).

B. PPAMT and feature sensitivity

Fig. 2 shows examples of concatenated phrases. Slash marks indicate the boundaries between phrases, which each contained a few words. The average duration of each phrase was $N_F/N = 77.4$ [frames] (0.774 [sec]). In this case, the sensitivities were $p_{L_3} = 0.984 \gg p_{\pi_3} = 0.317 > p_F = 0.194$, which confirmed the relationship in Sec. II-E.

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

Fig. 2. Examples of randomly concatenated phrases.

TABLE I
WER[%] OF PROPOSED PRIVACY PRESERVATION WHEN 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 (PPAMT)	14.88	13.76
speaker anonymization (10 clusters)	15.09	14.17

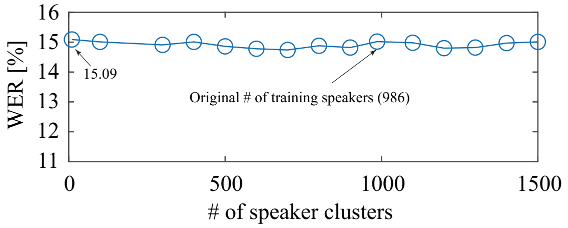

Fig. 3. WER[%] of CSJ test set when speaker clustering was used for speaker anonymization.

Table I shows the WER when only in-domain dataset was available. After a cross-entropy (CE) DNN acoustic model was obtained, sequential minimum Bayes risk (sMBR) discriminative training [9] was conducted. The first row is the upper limit WER when all in-domain speakers were used without privacy preservation. The second row is the WER with phrase division only. Although each word phrase contained only a few words and their duration was short, the acoustic model was properly trained. Random concatenation of the phrases, i.e., PPAMT, improved the WER because tri-phoneme variations were increased by the concatenation. sMBR was effective for PPAMT because the time sequences of acoustic features were reserved, which was an advantage of the proposed PPAMT. Speaker anonymization with ten clusters attained $k(=98)$-anonymization, although the WER were degraded by 0.2% for CE and 0.4% for sMBR.

C. *Number of speaker clusters*

Fig. 3 shows the relationship between the number of speaker clusters C and WER. The performance largely did not depend on the number of clusters for both $C < S$ and $C > S$. Thus speaker anonymization with arbitrary k is feasible.

D. *Subsampling of in-domain speakers without privacy preservation*

Another way to continuously use in-domain data is subsampling, in which partial in-domain data are used only when the speaker has agreed to the use of personal data. Fig. 4 shows the relationship between the number of training speakers, which were non-preserved, and WER. When the number of training speakers was less than 100, the performance significantly degraded. When the number of non-preserved speakers was 200 (approximately 1/5 of the total number), a 3% degradation of WER was observed.

The use of non-preserved speakers was also beneficial for PPAMT. Fig. 5 shows the relationship between the number of non-preserved speakers in PPAMT and WER. In the case of 200 non-preserved speakers, the WER degraded only 0.7%, although subsampling degraded the WER by 3%. Even when the number of non-preserved speakers was reduced, the performance degradation was less than that of subsampling. This indicates that PPAMT outperformed subsampling even when partial in-domain data were available without privacy preservation.

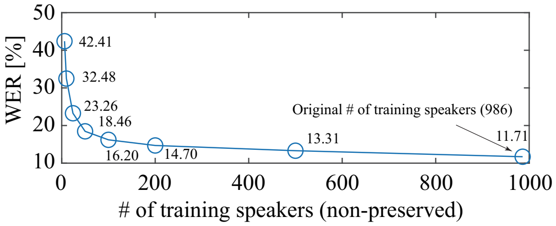

Fig. 4. WER[%] on the CSJ testset when training data were subsampled.

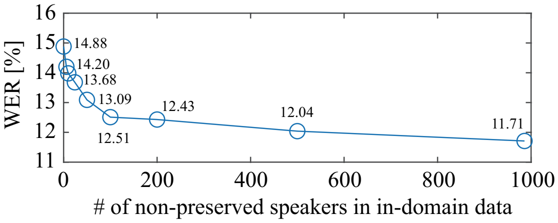

Fig. 5. WER[%] on the CSJ testset when partial speakers were preserved.

TABLE II
WER[%] OF PROPOSED PRIVACY PRESERVATION WHEN ADDITIONAL OUT-OF-DOMAIN DATASET WAS AVAILABLE. () SHOWS THE IMPROVEMENT FROM TABLE I.

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

E. *Effectiveness of additional out-of-domain dataset*

Table II shows the WER when additional out-of-domain dataset was available. Additional out-of-domain data were particularly effective for PPAMT because domain-independent knowledge could be learned. On the other hand, when only out-of-domain data were available, WER $(= 14.14\%)$ was much higher, indicating that privacy-preserved in-domain data significantly improved performance. In this case, speaker anonymization did not degrade the performance as in Sec. IV-C.

Fig. 6 shows the relationship between the number of non-preserved speakers in the in-domain dataset and WER. The WER degradation was 0.59% even without non-preserved speakers. When 200 speakers were used without privacy preservation, WER degradation was 0.26%.

F. *Output privacy for prior distribution*

Fig. 7 shows the cumulative probability in the prior distribution as in Eq. (6). The difference for 90% of the states was less than $\epsilon = 0.5$. The difference exceeded $\epsilon = 2$ for only 2.7% of the states. This shows that high privacy preservation

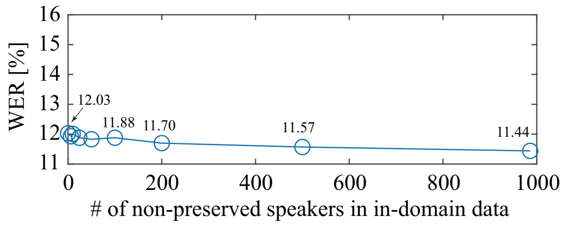

Fig. 6. WER[%] of CSJ test set when partial speakers were preserved with out-of-domain data.

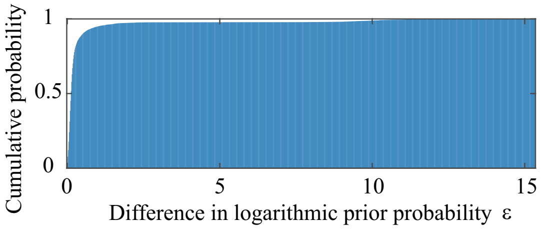

Fig. 7. Difference in logarithmic probabilities of prior distribution for tri-phone states, ϵ in Eq. (6).

of prior distributions can be attained using PPAMT in terms of output privacy because most prior distributions did not change.

V. Conclusion and future work

In this paper, we proposed a PPAMT method using perturbation and randomization as key operations. We formulated the sensitivities of three features (n-gram, phoneme labels, and acoustic feature) to PPAMT, i.e., the probabilities of being influenced by PPAMT. This indicates that the acoustic features and phoneme labels were less susceptible to PPAMT than the language features, which is optimal property for acoustic model training with privacy preservation. In addition, speaker anonymization was attained by speaker clustering. WER degradation by PPAMT was less than 0.6%, while the transcriptions could be restored with a negligible probability. Speaker anonymization did not degrade the performance with the inclusion of out-of-domain data. In future, we will conduct further theoretical analysis on the proposed PPAMT, particularly in terms of output privacy.

References

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

[2] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, “A big data approach to acoustic model training corpus selection,” in *Proceedings of INTERSPEECH*, 2014, pp. 2083–2087.

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

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

[5] D. Lambert, “Measure of disclosure risk and harm,” *Journal of Official Statistics*, vol. 9, no. 2, pp. 313–331, 1993.

[6] P. Smaragdis and M. Shashanka, “A framework for secure speech recognition,” *IEEE Transactions on Audio, Speech, Language Processing*, vol. 15, no. 4, pp. 1404–1413, 2007.

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

[8] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University Engineering Dept, 2003.

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

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

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

[12] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, “Privacy preserving synthetic data release using deep learning,” in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018.

[13] R. Masumura, Y. Ijima, S. Kobashikawa, T. Oba, and Y. Aono, “Can we simulate generative process of acoustic modeling data? Towards data restoration for acoustic modeling,” in *Proc. APSIPA*, 2019, pp. 655–661.

[14] A. Shah and R. Gulati, “Privacy preserving data mining: Techniques, classification and implications -a survey,” *International Journal of Computer Applications*, vol. 137, no. 12, 2016.

[15] P. Samrati and L. Sweeny, “Generalizing data to provide anonymity when disclosing information,” in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (PODOS)*, 1998, p. 188.

[16] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[17] C. Dwork, “Differential privacy,” in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, ser. Lecture Notes in Computer Science, 2006.

[18] A. D. Sarwate and K. Chaudhuri, “Signal processing and machine learning with differential privacy,” *IEEE Signal Processing Magazine*, pp. 86–94, 2013.

[19] J. Joy, D. Gray, C. McGoldrick, and M. Gerla, “K privacy: Towards improving privacy strength while preserving utility,” *Ad Hoc Networks*, pp. 16–30, 2018.

[20] 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*, vol. 19, no. 4, pp. 788–798, 2011.

[21] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[22] S. Furui, K. Maekawa, and H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” in *Proceedings of ASR*, 2000, pp. 244–248.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, “The Kaldi speech recognition toolkit,” in *Proceedings of ASRU*, 2011, pp. 1–4.