Computer Speech and Language 95 (2026) 101854

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Privacy-preserving feature extractor using adversarial pruning for TBI assessment from speech ☆

Apiwat Ditthapron [a], Emmanuel O. Agu [a,*], Adam C. Lammert [b]

[a] Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, USA
[b] Computer Science Department, College of the Holy Cross, Worcester, MA, USA

ARTICLE INFO

Keywords:
Health assessment
Privacy
DNN pruning
Adversarial training

ABSTRACT

Speech is an effective indicator of medical conditions such as Traumatic Brain Injury (TBI), but frequently includes private information, preventing novel passive, real-world assessments using the patient's smartphone. Privacy research for speech processing has primarily focused on hiding the speaker's identity, which is utilized in authentication systems and cannot be renewed. Our study extends privacy to include the content of speech, specifically sensitive words during conversation. Prior work has proposed extracting privacy-preserving features via adversarial training, which trains a neural network to defend against attacks on private data that an adversarial network is simultaneously attempting to access. However, adversarial training has an unsolved problem of training instability due to the inherent limitations of minimax optimization. Instead, our study introduces Privacy-Preserving using Adversarial Pruning (PPA-Pruning). Nodes are systematically removed from the network while prioritizing those contributing most to the recognition of personal data from a well-trained feature extractor designed for TBI detection and adversarial tasks. PPA-Pruning was evaluated for various privacy budgets via a differential privacy setup. Notably, PPA-Pruning outperforms baseline methods, including adversarial training and Laplace noise, achieving up to an 11% improvement in TBI detection accuracy at the same privacy level.

1. Introduction

Motivation: Speech contains rich information, including the speaker's identity, linguistic content (Nautsch et al., 2019; Ramanarayanan et al., 2022) and underlying medical conditions. Recently, speech processing for health assessments from speech has gained attention due to its potential to support non-invasive, cost-effective, and passive detection and monitoring of neurological disorders, including Traumatic Brain Injury (TBI) and depression (Al Mamun et al., 2017; Ramanarayanan et al., 2022; Renn et al., 2018). TBI and depression affect speech production, including impairing the coordination between brain regions responsible for speech planning and emotion. This results in difficulties with articulation, fluency, and voice modulation. Current speech assessments are typically performed in-clinic using speech tasks that require subject engagement or a questionnaire (Norman et al., 2013; Low et al., 2020). These in-clinic assessments can be tedious and time-consuming, and the requirement for physical presence at a clinic

☆ This material is based on research sponsored by DARPA, USA under agreement number FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. Results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI, USA grant DMS-1337943 to WPI.

* Corresponding author.
E-mail addresses: aditthapron@wpi.edu (A. Ditthapron), emmanuel@wpi.edu (E.O. Agu), alammert@wpi.edu (A.C. Lammert).

https://doi.org/10.1016/j.csl.2025.101854
Received 27 March 2024; Received in revised form 4 June 2025; Accepted 6 June 2025
Available online 23 June 2025

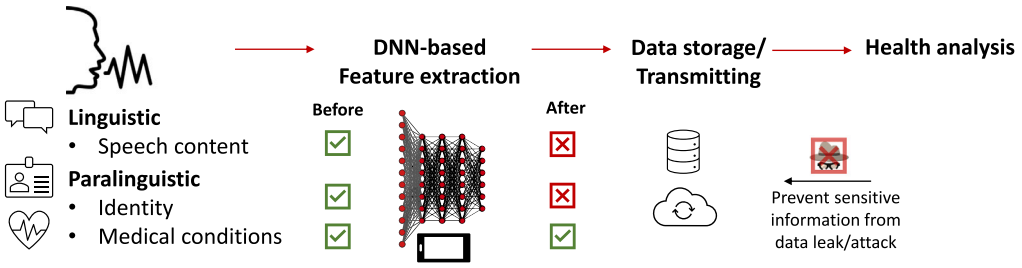

Fig. 1. Pipeline of privacy-preserving feature extraction using PPA-Pruning for health assessment from spontaneous speech.

and associated costs can be a barrier for some individuals. Enabled by advancements in speech technology and machine learning, automatic assessment and passive detection and monitoring of various ailments from speech (Banerjee et al., 2019; Ditthapron et al., 2022; Ramanarayanan et al., 2022; Renn et al., 2018; Talkar et al., 2020). Such technology-based automatic assessments can detect early signs of neurological disorders using mobile devices, reducing the healthcare burden associated with over 1.4 million TBI cases in the USA (Bavikatte et al., 2021; Faul et al., 2010). Much of prior work has focused on improving assessment accuracy with only a few addressing privacy concerns around recording and analyzing the speech or voice, which contain speaker identity (such as gender and age) and sensitive biometric data (Vildjiounaite et al., 2006; Kröger et al., 2020).

The problem of privacy concerns: As defined by Article 4 of European Parliament, Council of the European Union (2016), personal data are "any information relating to an identified or identifiable natural person". The regulation mandates the incorporation of a "privacy by design" engineering approach, requiring proactive measurement to address the risks of violating personal data throughout the system. For speech recordings, an individual's voice is considered personal data and, specifically biometric data that is utilized for authentication in personal devices and assistive technology.

Challenges: Unlike the protection of non-biometric data, which employs pseudonymization to replace sensitive information and prevent direct identification, voice characteristics pose a unique challenge. While pseudonymization techniques such as voice-style transfer (Qian et al., 2019) and Generative Adversarial Networks (GANs) for gender and fundamental formant transfer (Prajapati et al., 2022) exist for Automatic Speech Recognition (ASR), they are not suitable for health assessment. This limitation arises from the inherent complexity of individual medical conditions embedded within the voice, making it challenging to apply standard pseudonymization methods to ensure privacy in health-related contexts.

Previous work: Previous work addresses privacy concerns in health assessment by employing adversarial training, which aims to defend against attacks by another network, to train the feature extractor. Prior work incorporated the adversarial model to recognize speaker identity in the features extracted by the feature extractor network (Lavania et al., 2023; Srivastava et al., 2019). This involves a minimax optimization of two competing tasks: health assessment and the identification of personal data. The primary objective is to extract features for the target assessment task that are resistant to use in the adversarial task, specifically speaker identification and word recognition in this work. While adversarial training has demonstrated state-of-the-art results in many applications, it is not without challenges. The minimax or zero-sum game optimization inherent in adversarial training can lead to non-convergence and overfitting issues, where the model excels in one task at the expense of the other, rendering the latter uncompetitive (Wang et al., 2019). For instance, an exceptionally accurate assessment model may fail to preserve privacy, or vice versa.

Our proposed approach: To avoid issues with minimax optimization, we introduce an innovative approach to derive a well-performing and privacy-preserving feature extractor. Instead of relying on minimax optimization, a pruning scheme is employed within the adversarial training framework. Pruning nodes or connections in neural networks reduces network complexity, accelerating the network (Vadera and Ameen, 2022) by eliminating information from the network (Srivastava et al., 2014). An example in the latter case is dropout, which mitigates overfitting by preventing some nodes from adapting to new samples. Dropout removes information from a DNN, but does so randomly. For privacy preservation, Gong et al. (2020) introduced a pruning optimization to identify sparse patterns in a CNN that preserve data privacy during private training. This alternative seeks to overcome the training instability in traditional adversarial training, offering a potential solution to the challenges of non-convergence and overfitting in privacy-sensitive applications. Our proposed Privacy-Preserving using Adversarial Pruning (PPA-Pruning) is designed to safeguard the personal data of patients, specifically focusing on voice biometrics and linguistic content. This method is implemented during speech recording and feature extraction on smartphones, aiming to eliminate speech content and speaker identity from the extracted features, as illustrated in Fig. 1. Instead of retaining raw audio, privacy-preserving features are extracted from speech to protect sensitive information against potential malicious activities such as data leaks or attacks on stored data during data transmission for health analysis on the server. With the overarching objective of privacy-preserving feature extraction, our PPA-Pruning offers more comprehensive protection compared to anonymization and pseudonymization methods. The proposed method operates on a feature extractor previously pre-trained to extract features for both health assessment and adversarial tasks, systematically removing nodes that contribute the most to the adversarial task one by one. After each pruning iteration, the networks for both tasks are updated to seamlessly adapt to the pruned feature extractor. Pruning also reduces the computational complexity of the feature extractor on the smartphone and improves power utilization, a critical factor in mobile applications.

The pruning method was evaluated for the TBI detection task, selected as an example of a health assessment task from speech. The adversarial tasks included speaker identification and word recognition. Classification metrics for each task are reported, along with the privacy budget derived from differential privacy to measure the level of protected personal data. Adversarial training using minimax optimization, was performed.

Our contributions can be summarized as follows:

1. We proposed PPA-Pruning using a pruning method that is more stable than the traditional minimax method for extracting privacy-preserving speech features.
2. We evaluated our proposed PPA-Pruning in terms of accuracy and privacy, using a privacy budget to measure upper bound of privacy leaks, for TBI assessment on conversational speech.
3. We present results that demonstrate a higher level of privacy protection in extracted features compared to baselines, and comply with GDPR regulations; the pruning method achieves a higher detection accuracy of up to 11% than baselines at the same privacy-preserving level. Baselines included state-of-the-art speech feature sets for health assessment, formant coordinations, and bag-of-audio-word (BOAW).
4. We show that pruning using a gradient of nodes in the network is the only method that is "unlinkable" between speaker identities.

The rest of this paper is organized as follows: background and related work of privacy-preserving features for health assessment from speech are in Section 2. Our adversarial pruning method is described in Section 3, followed by experimental setup in Section 4. Evaluation results of the proposed method are reported in Section 5 with a discussion. Finally, we conclude our study in Section 6.

2. Background and related work

2.1. Regulation and guidelines on personal data

Protecting personal data is imperative during data collection and processing, and extends beyond anonymization. Analyzing labeled speech that obstructs speaker identity is crucial to avoid compromising speech-based authentication systems. Well-known regulations governing data collection and processing include the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). In the USA (2023), states such as Illinois, Texas, and Washington have initiated measures to safeguard *biometric identifiers*, encompassing voiceprints, iris scans, fingerprints, face scans, and face geometry. These states mandate an additional layer of data security. For instance, Illinois's Biometric Information Privacy Act (Illinois General Assembly, 0000) aims to ensure that entities "store, transmit, and protect all biometric identifiers and biometric information" with a higher level of protection than other confidential and sensitive information. Notably, Illinois empowers private individuals with the right of action, enabling them to seek redress ($1000–$5000 per violation) for any breaches of this protection.

GDPR recommends that systems processing personal identity information take measures to prevent collected data from being traced back to a consumer's identity. This is particularly relevant in the context of automatic health assessment from speech, which involves the speaker's identity. The GDPR states that the collection and usage of biometric data should adhere to the following principles (Nautsch et al., 2019; European Parliament, Council of the European Union, 2016):

1. Unlinkable: Speech representations from the same speaker should not be individually identifiable.
2. Renewable: References of speech must be renewable without the need for recollection.
3. Irreversibility: Speech signals should not be reconstructable from stored data.

Beyond speaker identity, concerns also arise regarding linguistic or intelligible spoken words, especially in the analysis of unscripted or conversational speech that requires no speaker attention but may contain words revealing the speaker's identity (Ditthapron et al., 2022; Talkar et al., 2020). To this end, we introduce a novel approach for privacy-preserving feature extraction using adversarial pruning. To the best of our knowledge, this is the first work that endeavors to preserve both speaker identity and speech content for health assessment applications.

2.2. Differential privacy (DP)

DP is a mathematical framework for protecting the privacy of individuals in statistical databases or data analysis. It gained attention in the context of data protection regulations, including the GDPR and other similar frameworks. This study uses privacy budget, a fundamental measurement of privacy in DP, to evaluate the privacy protection of the proposed pruning method. DP is formalized using the notion of a *privacy budget* and a *privacy loss function*. The definition involves the introduction of randomness through noise to obscure individual contributions to the data. DP can be defined as a mechanism M that is said to satisfy ε-differential privacy if, for all adjacent datasets D and D', and for all possible outcomes S in the range of M, the following holds:

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \times \Pr[M(D') \in S] \tag{1}$$

where ε is a privacy budget, a non-negative parameter that quantifies the privacy guarantee. Smaller values of ε correspond to stronger privacy guarantees. $M(D)$ represents the output of the mechanism M on dataset D. S is a subset of possible outcomes. D and D' are neighboring datasets, meaning that they differ by the inclusion or exclusion of a single individual's data. The formula

states that the probability of obtaining a specific outcome S on the dataset D is roughly the same as obtaining that outcome on dataset D', up to a multiplicative factor e^{ϵ}. A ϵ value of zero indicates perfect privacy and larger values imply weaker privacy guarantees.

Privacy Budget Parameter: The parameter ϵ is the *privacy loss parameter*. Smaller values of ϵ provide stronger privacy guarantees but may reduce the accuracy of the output.

Noise Injection: DP protects data privacy by adding random noise to individual data points before they are aggregated. This method is used as a baseline for privacy protection on handcrafted speech features, which excludes the proposed method and adversarial baselines. The added noise makes it difficult to determine the contribution of a specific individual's data during training. However, this study adopts this concept for inference of the health assessment, where speech input is D in the DP formula. Although DP provides substantial protection of personal data, it requires careful consideration of parameters and potential trade-offs between privacy and system performance, controlled by the privacy parameter ϵ.

De-identification and preserving anonymity in speech features: DP aids in the de-identification of individual recordings. Even if an adversary has auxiliary information about some individuals, in differentially private datasets (speech sample in this study), the added noise makes it challenging to re-identify specific individuals, thereby protecting against re-identification attacks. However, we considered a speech sample from one speaker as a dataset and each temporal snippet of speech as a data entry—in a way that aggregating data with added noise is preserved under DP. DP provides a mathematically rigorous framework for privacy protection. It ensures that the privacy guarantees hold even in the presence of sophisticated adversaries attempting various attacks to de-anonymize the data. DP can be used to meet the *unlinkable* and *irreversibility* requirements of the GDPR.

2.3. Voice anonymization methods

Voice anonymization has become an important area of research that aims to protect sensitive information about speakers while retaining the meaning of speech and emotions. To balance these competing objectives, a number of recent studies have explored adversarial frameworks and Differential Privacy (DP) mechanisms. In this section, we provide an organized review of these prior methods.

2.3.1. Adversarial-based methods

Prior work has explored adversarial training to conceal speaker-specific information in the learned representations. Wang et al. (2024), Srivastava et al. (2019) and Chen et al. (2023) proposed adversarial losses that can be used to make embeddings less sensitive to the speaker's identity. An adversarial model is trained to guess the speaker's identity based on encoded features, and the speech feature extractor is penalized based on how well the adversarial model does. Min–max optimization, which is built into these methods, can cause problems such as training instability and convergence because the adversarial network may overfit or underfit at different stages of training.

Qian et al. (2019) and Prajapati et al. (2022) propose extensions to the adversarial paradigm by incorporating generative frameworks. In 2019, Qian et al. created a zero-shot voice style transfer model using an autoencoder structure. In 2022, Prajapati et al. used CycleGAN architectures along with time-scale modification to make voices less recognizable. While these methods are good at retaining the meaning of the language, they often struggle with achieving the right balance between performance and anonymization, especially when sensitive features are linked to the speaker's language patterns.

In addition, Gu et al. (2024) and Hua et al. (2024) focus on preserving not only privacy but also other aspects of speech, such as emotional content. Gu et al. (2024) use a strategy called *disentanglement* to separate features that are related to the content from those that are not related to the content. Hua et al. (2024) created fake speaker speech that retained emotional cues. These works provide a comprehensive understanding of voice anonymization. However, while it is important to protect speaker privacy, it is just as important to keep the speech signals still usable for other tasks, such as emotion detection and automatic speech recognition.

2.3.2. Differential privacy in voice anonymization

Another research direction focuses on applying DP to the task of voice anonymization. Shamsabadi et al. (2023) present a framework that integrates DP directly into the adversarial training process. Their method adds calibrated noise to the gradients and intermediate feature representations, ensuring that the information from any speaker only makes a small difference in the final output. Shamsabadi et al. (2023) show that their method achieves state-of-the-art anonymization while still keeping acceptable levels of speech intelligibility for later tasks. Their proposed method uses noise injection that is carefully tuned to the extracted features, following ϵ-DP. This implies that the presence or absence of a single speaker's data in the training set will not significantly alter the learned representations.

In contrast, our proposed PPA-Pruning takes a different direction to achieve privacy preservation. PPA Pruning systematically determines and removes feature extractor nodes that change the adversarial task the most, which corresponds to discovering speaker-sensitive attributes. This is done without adding noise to the model's gradients or intermediate representations. This pruning-based strategy inherently reduces the risk of privacy leakage by removing the specific network components that capture personal data.

A notable advantage of the PPA-Pruning approach is its improved training stability. The noise injection method used by Shamsabadi et al. (2023) requires knowing the noise level (and, by extension, the privacy budget) to be carefully tuned. PPA-Pruning, on the other hand, does not utilize min–max adversarial optimization at all. Our method maintains the network's performance high on the target health assessment task while ensuring that the pruned network is unable to reconstruct sensitive speaker information. We achieve this by repeatedly pruning the network and then adapting it through model re-training. PPA-Pruning also used DP as

a technical measure to find privacy leaks. However, Shamsabadi et al. (2023) employs DP by enhancing the randomness of the network's calculations.

In addition to suppressing speaker identity, our approach also targets the verbal content in the speech signal. Although prior work (Ahmed et al., 2020) considers specific words, such as names, locations, or other unique identifiers, which increases the risk of privacy leakage, we simply suppress all verbal content in the speech. This is because speech is inherently contextual; even words that appear neutral in isolation may reveal sensitive information when combined with surrounding context. Given the difficulty in reliably distinguishing which words are more indicative of privacy risks, to ensure comprehensive protection, our method treats every word as potentially sensitive.

2.4. Neural network pruning

Most pruning aims to reduce the number of parameters in a large DNN without compromising accuracy, fundamentally by removing connections between layers, nodes, or filters in a large pre-trained network (Vadera and Ameen, 2022; Kulkarni et al., 2022). The main criteria for pruning a neural network can be categorized into three groups: magnitude-based, similarity clustering, and sensitivity analysis (Vadera and Ameen, 2022). All three methods share the same goal: reducing network complexity with minimal change to the final prediction, which involves identifying nodes that contain the same information or do not contribute significantly to the prediction. However, as our proposed method aims to remove nodes that specifically contribute to a certain measurement, such as private data, only sensitivity analysis criteria for pruning meet our objectives.

In addition to the objective of reducing network parameters, pruning is frequently employed during training as dropout, which temporarily removes random connections of nodes to avoid overfitting (Srivastava et al., 2014). Although dropout was not initially proposed as a privacy-preserving method, Jain et al. (2015) demonstrated that dropout preserves privacy under the differential privacy terms, in addition to improving learning stability in deep belief networks. In evaluating our proposed adversarial pruning method, dropout is considered and evaluated for random pruning to preserve speaker and speech content privacy.

3. Proposed method

3.1. Overview and requirements

We propose Privacy-Preserving using Adversarial Pruning (PPA-Pruning) to preserve speaker identity and linguistic content during health assessments from speech recordings. Speech features extracted using a deep neural network (DNN) are denoted as h, obtained through the feature extractor $\mathcal{F}$. These features are protected against personal data attacks, represented by the adversarial network $\mathcal{G}_{ADV}$, while simultaneously performing well on the TBI detection task, denoted as $\mathcal{G}_{TBI}$. This aligns with the common setup of adversarial networks trained using minimax optimization. However, the inner maximization in this setup is globally non-concave, posing training challenges between the $\mathcal{G}_{ADV}$ and $\mathcal{G}_{TBI}$ networks. In an effort to derive a stable privacy-preserving feature extractor, we propose PPA-Pruning as an adversarial method, which simply removes nodes in $\mathcal{F}$ that contribute the most to $\mathcal{G}_{ADV}$ as part of the outer minimization process. Unlike minimax optimization, the PPA-Pruning assumes that networks $\mathcal{F}$, $\mathcal{G}_{ADV}$, and $\mathcal{G}_{TBI}$ have already converged through multi-task learning for the TBI detection and adversarial tasks before the pruning stage.

Prior to pruning to preserve privacy, multi-task learning is used to train feature extractors $\mathcal{F}$, which comprise l layers each containing n_l nodes, for both the TBI detection and adversarial tasks (speaker identification and ASR). The entire pruning process is depicted in Fig. 2. Considering different ranges and sensitivities of loss functions in each task, gradients in each task are normalized using GradNorm (Chen et al., 2018), which weights the loss (w_t) in each training batch, as outlined in Eq. (2). The pruning starts after $\mathcal{F}$ converges for both tasks.

$$L_{grad}(\theta_t) = \sum_t \left| \left\| \nabla_\theta w_t f_t(\theta_t) \right\| - \overline{\left\| \nabla_\theta w f(\theta) \right\|} \right| \tag{2}$$

3.2. Node pruning

This study proposes two criteria for selecting and pruning nodes in $\mathcal{F}$ that contribute the most to the training objectives in adversarial tasks. The weight of the node is not considered a pruning criterion, as it signifies the importance of the node for both TBI and adversarial tasks. Instead, this study explores SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017), two algorithms commonly employed to explain feature importance in DNN. Both algorithms capture the contribution of each node in DNN to the predictions, including speaker identity and spoken words that we aim to exclude from extracted features.

1. **SHAP** estimates the contribution of each feature to speaker identity and word recognition predictions. The Shapley value, a concept from cooperative game theory, is used to assign a fair value to each feature based on its marginal contribution to all possible coalitions of features. In the context of feature importance, the Shapley value of the node n_{ij} where i, j represents the node j of the layer i in the feature extractor $\mathcal{F}$, is calculated as the average marginal contribution of that feature across all possible permutations of features. The formula for the Shapley value is

$$\mathrm{Sh}_i(f) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} [f(S \cup \{i\}) - f(S)], \tag{3}$$

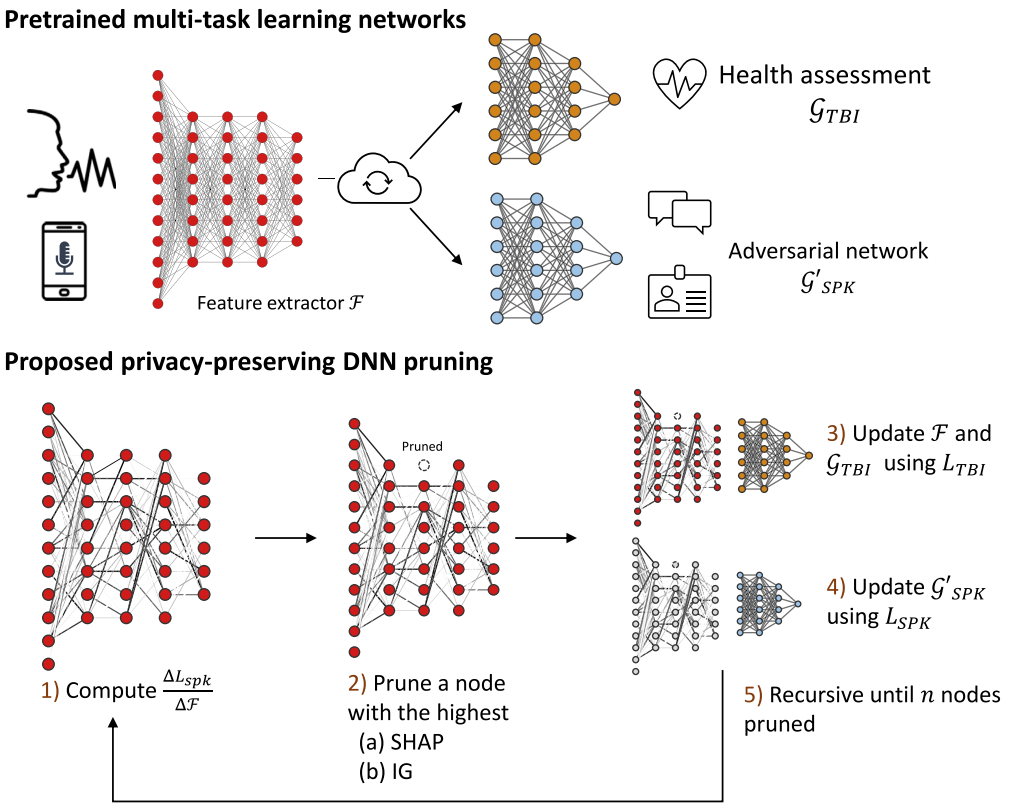

Fig. 2. The proposed PPA-Pruning preserves speaker privacy in extracted speech features. Shapley Additive Explanations (SHAP) and Integrated Gradients (IG) are adopted as measures to remove a node with the highest contribution to the adversarial network, which violates the speaker's privacy. In each iteration, a node is pruned, followed by network adaptation in discriminators to match changes due to pruning.

where $f(S)$ is the model's prediction when using only the features in the set S, M is the set of all features, $|S|$ is the number of features in set S, $|M|$ is the total number of features, and $S \subseteq M \setminus \{i\}$ denotes all possible subsets of features excluding feature i. This gives a fair distribution of the model prediction among the features, considering all possible combinations. In practice, since evaluating all subsets is computationally expensive, approximation methods like Monte Carlo simulations or kernel SHAP are often used to estimate the Shapley values.

2. **IG** assigns an importance score to a node n_{ij} by integrating the model's predictions over a path from a baseline (usually a reference input with zero influence) to the actual input. The formula for IG for a feature i is given by

$$\mathrm{IG}_i(f) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{z} + \alpha \times (\mathbf{x} - \mathbf{z}))}{\partial x_i} \, d\alpha, \tag{4}$$

where f is the model's prediction function, $\mathbf{x}$ is the actual input, $\mathbf{z}$ is the baseline input, x_i and x_i' are the ith feature values of $\mathbf{x}$ and $\mathbf{z}$, respectively.

In each pruning iteration, all DNN connections to node n_{ij} that have the highest $\mathrm{Sh}_i(f)$ or $\mathrm{IG}_i(f)$ are removed. This removal impacts subsequent DNN layers, particularly the normalization layer and bias, necessitating network retraining for the DNN to make accurate predictions.

3.3. Training procedures and network adaptation

PPA-Pruning performs four main steps as follows:

1. **Forward Pass:** Run the updated feature extractor $\mathcal{F}$ (with fewer nodes) on a batch of training samples, thereby generating updated feature representations.
2. **Update** $\mathcal{G}_{TBI}$: Compute the TBI detection loss (L_{TBI}) using the updated representations and backpropagate through $\mathcal{G}_{TBI}$ and $\mathcal{F}$. This step ensures that $\mathcal{G}_{TBI}$ adjusts its parameters to accommodate the pruned structure while preserving detection accuracy.

3. **Update $\mathcal{G}_{SPK}$ and $\mathcal{G}_{ASR}$ (Adversarial Tasks)**:
Following the TBI update, we compute losses for the adversarial tasks of speaker verification and automatic speech recognition. Let $\mathbf{h}$ denote the updated features from the pruned extractor $\mathcal{F}(\mathbf{x})$. For speaker verification, the loss L_{SPK} is calculated via cross-entropy over C_{SPK} possible speaker classes and over N samples in the training batch:

$$L_{SPK}(\theta_{\mathcal{F}}, \theta_{SPK}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C_{SPK}} y_{SPK}^{(i,c)} \log\big(\mathcal{G}_{SPK}(\mathbf{h}^{(i)})_c\big), \tag{5}$$

where $y_{SPK}^{(i,c)} \in \{0,1\}$ indicates whether a sample i belongs to the speaker class c, and $\mathcal{G}_{SPK}(\mathbf{h}^{(i)})_c$ is the predicted probability for that class.
Similarly, for the word recognition task, L_{ASR} is computed by comparing the predicted token distribution against the ground-truth tokens for each input. Assuming a vocabulary of size V, let $\{w_t\}_{t=1}^{T}$ be the predicted word (or subword) sequence and $\{z_t\}_{t=1}^{T}$ the ground truth:

$$L_{ASR}(\theta_{\mathcal{F}}, \theta_{ASR}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{v=1}^{V} z^{(i,t,v)} \log\big(\mathcal{G}_{ASR}(\mathbf{h}^{(i,t)})_v\big), \tag{6}$$

where $z^{(i,t,v)} \in \{0,1\}$ indicates if the ground-truth token at time t matches the vocabulary index v, and $\mathcal{G}_{ASR}(\mathbf{h}^{(i,t)})_v$ is the predicted probability for the token v at time t.
We then form a combined adversarial loss by summing or weighting these two terms,

$$L_{ADV} = \alpha L_{SPK} + \beta L_{ASR}, \tag{7}$$

where α and β control the relative importance of each adversarial objective. The parameters of $\mathcal{F}$, $\mathcal{G}_{SPK}$, and $\mathcal{G}_{ASR}$ are updated via stochastic gradient descent (Adam optimizer):

$$(\theta_{\mathcal{F}}, \theta_{SPK}, \theta_{ASR}) \leftarrow (\theta_{\mathcal{F}}, \theta_{SPK}, \theta_{ASR}) - \eta \nabla_{\theta} L_{ADV}.$$

This ensures that both the speaker-verification and speech-recognition networks adapt to the pruned architecture, striving to recover lost information, while the pruned extractor $\mathcal{F}$ continues to suppress sensitive details in each subsequent iteration.
4. **Iterative Pruning:** This process is repeated for each pruning iteration until either the desired pruning rate or a specified convergence criterion is reached. Throughout these iterations, the model continuously adapts to its reduced capacity, balancing TBI detection performance with suppressed sensitive information.

Following each pruning step, both $\mathcal{F}$ and $\mathcal{G}_{TBI}$ are updated using gradients computed from the L_{TBI} objective, and $\mathcal{G}_{ADV}$ is updated using the L_{ADV} objective. It is crucial to note that this is not a minimax adversarial process, as F is not trained to outperform $\mathcal{G}_{ADV}$. Instead, $\mathcal{G}_{ADV}$ is retrained (adapted) at each pruning iteration to learn from the current state of the evolving, privacy-preserved features. This continuous retraining of $\mathcal{G}_{ADV}$, starting from its initial training on rich unpruned features, represents a highly adaptive adversarial agent. Its reported performance (EER for speaker verification and WER for ASR) at each pruning iteration inherently reflects its best attempt to extract private data from the features as they become progressively more anonymized. This dynamic evaluation provides a robust indicator of privacy leakage, demonstrating the ongoing challenge an attacker would face, effectively simulating a worst-case scenario where the attacker's model is continuously optimized to extract any remaining sensitive information.

3.4. DNN models

Our DNN model extends AM-MobileNet1D (Nunes et al., 2020), a compact Convolutional Neural Network (CNN) utilizing additive margin softmax for learning speaker representations. This extension aims to learn common features for TBI detection, speaker identification, and speech recognition tasks. The DNN architecture, including the number of CNN blocks, CNN filter size, and dropout rate in $\mathcal{F}$, $\mathcal{G}_{TBI}$, and $\mathcal{G}_{ADV}$, was optimized using AsyncHyperBand neural architecture searching implemented in the Ray Tune library (Liaw et al., 2018). Additionally, other optimization parameters, such as learning rate and decay rate, were fine-tuned. The best-performing architectures for TBI detection and privacy preservation have 3 residual blocks in the feature extractor, 2 blocks of the Fully-Connected layer (FC) in the adversarial networks, and 2 blocks of FC in the TBI detection network. Each residual block in the feature extractor is the inverted residual block from MobileNetV2 (Sandler et al., 2018) that implements depth-wise convolutions to reduce computational complexity. After the depth-wise convolutions, point-wise convolution is performed using a 1×1 convolution filter to combine the output from each depth-wise convolution. Batch normalization and the ReLU activation function were included at the end of each residual block. The configuration of the channel (depth) is reported in Table 1. All nodes in the last layer of the feature extractor are flattened to one dimension and an FC layer of 896 channels is applied. A vector of shape (1896) is used to extract speech features on which we aim to preserve privacy in this study. Each block in $\mathcal{G}_{TBI}$, $\mathcal{G}_{SPK}$, and $\mathcal{G}_{ASR}$ contains a Dropout layer, followed by a single layer of FC, batch normalization, and the ReLU activation function. In the last block, the softmax activation function was used instead of ReLU.

Table 1
Model configuration.

Parameters	Tuned value
Model $\mathcal{F}$	
Input channel	32
Numbers of residual channels	16,24,32
Output channel	896
Learning rate	2.2e−3
Model $\mathcal{G}_{TBI}$	
Numbers of channel	56,16
Dropout rate	0.24,0.20
Learning rate	3.43e−5
Model $\mathcal{G}_{SPK}$	
Numbers of channel	64,60
Dropout rate	0.46,0.28
Learning rate	3.01e−5
Model $\mathcal{G}_{ASR}$	
Numbers of channel	64,32
Dropout rate	0.35,0.24
Learning rate	4.2e−5

3.5. Identification and suppression of sensitive words

While our primary goal is to obscure sensitive linguistic content, our method does not rely on a predefined external lexicon or a Named Entity Recognition (NER) system. Instead, we adopt an adversarial approach that treats all spoken words as potentially sensitive. Specifically, the adversarial network $\mathcal{G}_{ASR}$ is trained to recognize any intelligible content in the extracted features. During pruning, any node or connection that strongly contributes to correct word recognition is considered a candidate for removal. Consequently, words that are easily identified by $\mathcal{G}_{ASR}$ are more aggressively suppressed, since their gradient contributions are higher. This strategy avoids the need for explicit lists of sensitive terms (e.g., personal names, locations) or more sophisticated NER pipelines. Instead, we empirically ensure that features facilitating the recognition of any token, potentially revealing personal or private details in context, are pruned. This provides a broader privacy guarantee, especially when conversational or spontaneous speech is used and sensitive information may not be captured easily by a static dictionary of keywords.

4. Evaluation method

PPA-Pruning, which employs SHAP and IG as criteria to prune the networks, was evaluated for the TBI assessment task and adversarial tasks, including speaker identification and word recognition. Baselines for adversarial learning, involving minimax optimization and DP methods applied to different speech features, were considered.

4.1. Speech corpora

4.1.1. Coelho TBI corpus

PPA-Pruning was trained and evaluated for the TBI detection task using the Coelho TBI corpus (Coelho et al., 2002). The dataset comprises conversational speech collected during discourses following TBI, which included story retelling, story generation, and conversations from 55 native English speakers with non-penetrating head injuries, as well as 52 native English speakers with no brain injury (control subjects). Only conversational speech produced by the subjects was considered in this study. Each discourse lasted around 10–15 min and was examined by a speech-language pathologist. The conversation was initiated by the examiner with the question "Why are you here at the hospital/rehabilitation center today?" and was then continued by either the subject or the examiner. The dataset exhibits a diverse demographic range in terms of gender (Male:39, Female:16), age (28.53 ± 12.31), education level (13.01 ± 2.38 years), working class (Unskilled:19, Skilled:18, Professional:18), TBI severity (Coma duration: 16.95 ± 22.10 days) and recording on-set time after the accident (10.35 ± 17.78 months). The causes of brain injury for subjects in this corpus include motor vehicle accidents (45 subjects), falls (6 subjects), struck by cars (3 subjects), and others (1 subject). The Coelho corpus was utilized to train $\mathcal{G}_{TBI}$ with the binary cross-entropy loss function for classifying TBI and healthy subjects in both pre-training and network adaptation. The dataset was also employed during the pruning (calculating attribution scores) of $\mathcal{F}$ and the network adaptation of $\mathcal{G}_{ADV}$. The model and optimization details are explained in Section 3.4.

While the Coelho corpus is highly relevant for assessing clinical speech, its limited size poses challenges for training complex models and could impact the generalizability of our models. We partially mitigate this constraint through a multi-stage approach: first, we pre-train our adversarial networks using the much larger Librispeech dataset, which contains read speech from over a thousand speakers. This pre-training step enables the learning of robust speaker and linguistic representations, exploiting the extensive and varied speaker characteristics present in Librispeech.

4.1.2. Librispeech

The Librispeech corpus (Panayotov et al., 2015) comprises a collection of audiobooks with 1166 speakers (564 females, 602 males) and has been widely utilized for pre-training and evaluating various speech processing tasks, including speaker identification and ASR. In this study, the Librispeech corpus was employed to pre-train the adversarial model $\mathcal{G}_{ADV}$ due to the limited number of speakers and vocabulary in the Coelho corpus. For the ASR task, the vocabulary size (V) was derived from the Librispeech corpus, consisting of approximately 10,000 subword tokens (SentencePiece).

Nevertheless, we acknowledge a potential domain mismatch: Librispeech contains read, audiobook-style utterances, whereas the Coelho corpus features spontaneous speech that inherently differs in fluency, disfluency patterns, and lexical richness. To reduce the impact of domain mismatch, all networks are fine-tuned during each pruning step on the Coelho corpus.

4.1.3. Audio pre-processing and data splitting

All audio files were initially downsampled to 16 kHz. In the Coelho corpus, noise suppression was applied to reduce background noise, utilizing the minimum mean square error estimated from the spectral amplitude from the MATLAB Voicebox toolbox. The Coelho corpus provides a timestamp, which was used to eliminate non-speech segments and split pathologist and subject speeches. Finally, the speech from each subject was divided into chunks of 200 ms, with a 10 ms overlap. In both datasets, speech samples were split using subject-wise five-fold cross-validation, with test subjects excluded from the training set.

4.2. Evaluation metrics

4.2.1. TBI detection

Balanced Accuracy (BA), computed as $\text{BA} = \frac{\text{Sensitivity+Specificity}}{2}$, was used to assess the classification performance of privacy-preserving feature extractors $\mathcal{F}$ and the binary classification network $\mathcal{G}_{TBI}$ for the TBI detection task. The accuracies of positive and negative classes were given equal weight. The accuracy was subsequently averaged over five-fold cross-validation, with the standard error calculated using $\text{Standard Error} = \frac{\sum_{i=1}^{5}(m_i-\bar{x})^2}{\sqrt{5}}$, where m_i represents the performance of the validation fold i out of 5 folds.

4.2.2. Speaker verification

The model $\mathcal{G}_{SPK}$ was initially trained for speaker classification using the Librispeech corpus for pretraining and the Coelho corpus (training set) for network adaptation. However, we assessed $\mathcal{G}_{SPK}$ for the speaker verification task using the Equal Error Rate (EER) on the test set of the Coelho corpus. This choice was made because testing subjects were excluded from the training set.

For each test sample, we treated the output to the last layer of $\mathcal{G}_{SPK}$ after each pruning iteration as a d-vector representing the speaker's voice. This d-vector was then used to compute the cosine similarity to the ten reference d-vectors extracted from the same speaker. The averaged cosine similarity was considered as the speaker probability in EER. EER is the point where the False Acceptance Rate (FAR) is equal to the False Rejection Rate (FRR). In other words, it represents the threshold at which the probability of incorrectly accepting a different speaker is the same as the probability of incorrectly rejecting the true speaker. In order to illustrate the performance comparison between speaker verification and TBI detection, we also calculated the BA at the EER point.

4.2.3. Automatic speech recognition

Word Error Rate (WER), computed as $\frac{S+D+I}{N}$ was used to evaluate speech recognition network ($\mathcal{G}_{ASR}$), the adversarial model. WER takes the number of substitutions (S), number of insertions (I), number of deletions (D) and the total number of words (N) in the ground truth into account. A lower WER indicates a smaller difference between the generated output and the ground truth and better ASR system performance.

4.2.4. Privacy budget

Privacy budget (ϵ) is a parameter in DP that controls the level of privacy loss. A mechanism M or network is ϵ-differentially private if Eq. (1) holds for all measurable sets S in the range of M. The equation can be rewritten to estimate ϵ as

$$e^{\varepsilon} \geq \frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]}$$

where D' is a dataset that differs from the dataset D by one entry. The privacy budget is computed for the proposed method and baselines, where S is the label in speaker identity and speech recognition.

4.3. Baseline

4.3.1. Adversarial using minimax optimization

Srivastava et al. (2019) previously proposed an adversarial loss function for ASR privacy preservation. The training objective is to enhance the performance of ASR models $\mathcal{F}$ and $\mathcal{G}_{ASR}$ relative to adversarial networks and the speaker identification model $\mathcal{G}_{SPK}$ using $\min_{\mathcal{F},\mathcal{G}_{ASR}} \max_{\mathcal{G}_{SPK}} \mathcal{L}_{ASR} - \alpha\mathcal{L}_{SPK}$ The parameter α is employed to control the degree of privacy preservation, where a high value degrades ASR performance. However, the speaker identification model $\mathcal{G}_{SPK}$ might struggle to converge if the α value is excessively small. This study substituted the ASR task in the adversarial baseline with the TBI detection task.

4.3.2. Laplace noise method

In order to incorporate differential privacy into our baseline Bag-of-Audio-Words (BOAW), Low-Level Descriptor (LLD), and Formant-based feature sets, we add Laplace noise to each dimension of the extracted feature vectors. Specifically, for each original feature vector $\mathbf{x} \in \mathbb{R}^d$, we compute the noisy version $\mathbf{x}'$ as:

$$\mathbf{x}' = \mathbf{x} + \mathbf{n}, \quad \text{where } \mathbf{n} \sim \mathcal{L}\left(0, \frac{\Delta}{\epsilon}\right).$$

Here, $\mathcal{L}(0, \frac{\Delta}{\epsilon})$ denotes the Laplace distribution with mean 0 and scale parameter $b = \frac{\Delta}{\epsilon}$, where Δ is the global sensitivity of the feature vector and ϵ is the privacy budget. After noise is added, we then train our TBI detection model on the noise-infused features. Please note that these baselines are not evaluated for speaker verification and ASR. The three speech feature sets are as follows:

Acoustic feature: A collection of Low-Level Descriptor (LLD) from COMPARE 2016 acoustic features (Schuller et al., 2013) was extracted from speech using the OpenSMILE library (Eyben et al., 2010). A total of 130 LLD features, including energy, spectral, and voicing features, were extracted over 20 ms audio length with a 10 ms time step. To obtain a trade-off between the BA of TBI detection and privacy, the DP method was applied with a range of privacy budgets from 0.01 to 500. Principal Component Analysis (PCA) of 12 components, experimentally tuned, was used to reduce the feature dimension after the DP method.

Bag-of-Audio-Word (BOAW): The BOAW technique clusters the extracted COMPARE features into groups also called audio words. Next, histograms of these audio words are constructed to count the occurrences of common audio words in each speech sample. In this study, BOAW is employed as an alternative to PCA for dimensionality reduction. After extracting COMPARE 2016 features, two BOAW codebooks, each containing 1000 audio words, were generated using OpenXBOW (Schmitt and Schuller, 2017). These codebooks were then used to vectorize the COMPARE 2016 features. Due to the clustering method employed, DP could not be applied. The privacy budget was estimated from BOAW, inherently preserving privacy as it only involves the histogram of the reference point.

Formant coordinations: The Praat software was employed for extracting the first three formant tracks from speech. Following the procedures outlined in Williamson et al. (2019), we constructed a matrix of correlation and covariance coefficients for the first three formants across four time-delayed scales: 10 ms, 30 ms, 70 ms, and 150 ms. In the final step, total power and entropy constant values were estimated and incorporated into the feature set. PCA was then applied to reduce the dimensionality of the data, resulting in 8 components. Both DP and PCA were applied in the same manner as in the acoustic feature extraction process.

4.4. Experimental setup

The evaluation was performed using the tuned model for the TBI detection task, treating speaker identification and automatic speech recognition as adversarial tasks. PPA-Pruning operates iteratively, pruning one node in the network at a time. We evaluated the method across a range of pruning rates, from 0% to 30%, representing the ratio to the total nodes in the feature extractor $\mathcal{F}$, with increments of 5. In the first experiment, we evaluated the proposed pruning method for its impact on decreasing TBI detection accuracy and gaining privacy budget in each iteration of pruning, followed by comparisons to the baselines. Dropout was also considered a pruning method that randomly removed nodes. It is important to note that the measured privacy budget depends on the array of features rather than the specific adversarial tasks. Subsequently, we report the relationship between the privacy budget in DP and speaker identification as well as speech recognition.

All training and evaluations were performed on an NVIDIA A100 GPU using the PyTorch library version 1.12 (Paszke et al., 2019). The Adam optimizer was used with different learning rates for each network, as reported in Table 1. Pre-training was performed for up to 100 epochs with an early stop when loss values stopped improving for 10 epochs. Model adaptation was performed for one epoch at the end of the pruning iteration.

5. Results and discussion

5.1. Model tuning

The size of the DNN architecture was limited by the size of the Coelho corpus, which is considerably small for the TBI and speaker identification tasks and extremely small for the word recognition task. Pre-training using an external corpus (Librispeech) was employed for the feature extractor and adversarial block with fine-tuning on the Coelho corpus, but overfitting occurred rapidly. We selected the network with the lowest TBI and adversarial errors, equally normalized using GradNorm. The two lowest trials came from the same architecture, comprising 3 feature blocks, 2 TBI detection blocks, and 2 adversarial blocks, as illustrated in Fig. 3.

5.2. Privacy-preserving features using pruning method

The TBI detection accuracy decreases as the pruning rate increases from 0% to 30%. As illustrated in Fig. 4, these trends vary based on the pruning schemes employed. In the result without model adaptation, pruning more than 5% of all nodes significantly damages the TBI detection model; BA is not better than a random guess. This highlighted the need for model adaptation. In the results of pruning with model adaptation, pruning based on SHAP yields higher TBI accuracy but does not preserve privacy as well as pruning based on IG. However, pruning based on IG demonstrates a large standard error in the privacy budget. Overall, the proposed PPA-Pruning performs better than the dropout baseline, which was previously considered a privacy-preserving method (Scheliga et al., 2023), on both metrics. In the comparison between DP and TBI accuracy, each method demonstrates a noticeable plateau on

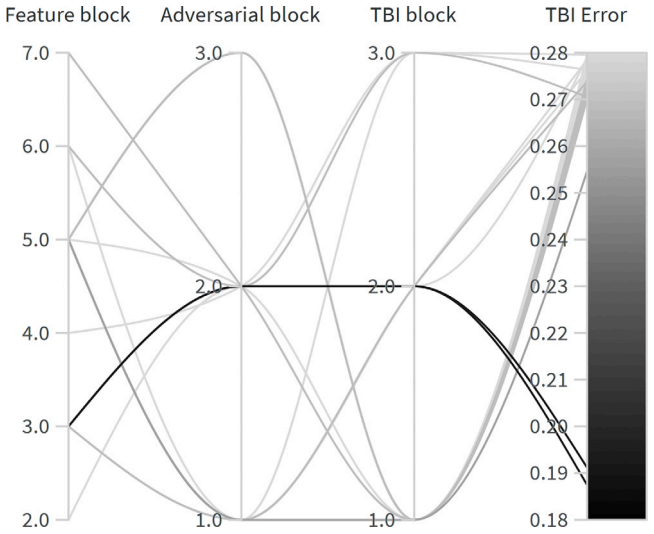

Fig. 3. Selected trials of tuning for model architecture before the pruning.

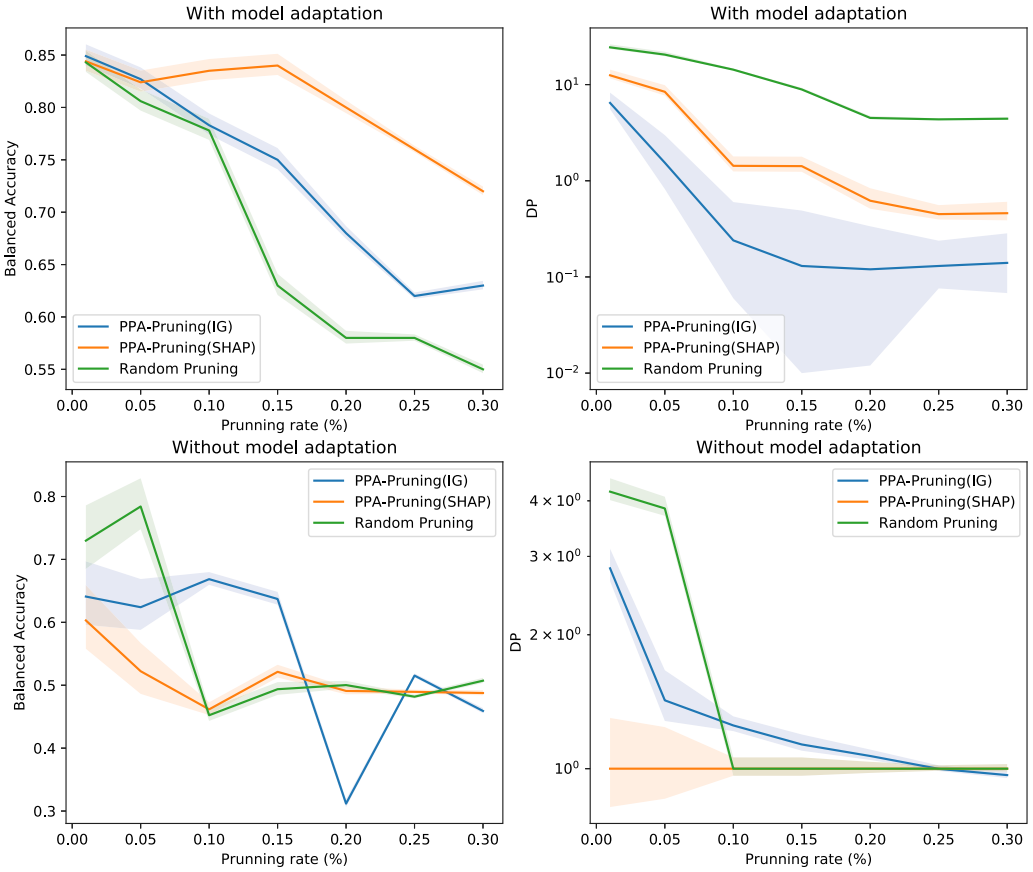

Fig. 4. TBI balanced accuracy and privacy budget of DP at different pruning rates of three pruning methods. TOP: the proposed pruning method with model adaptation between each pruning iteration; BOTTOM: the proposed pruning method without model adaptation.

the DP plot, but such a plateau is not evident on the TBI accuracy plot. This suggests that the pruning methods had a more stabilized behavior in privacy budget than the TBI accuracy.

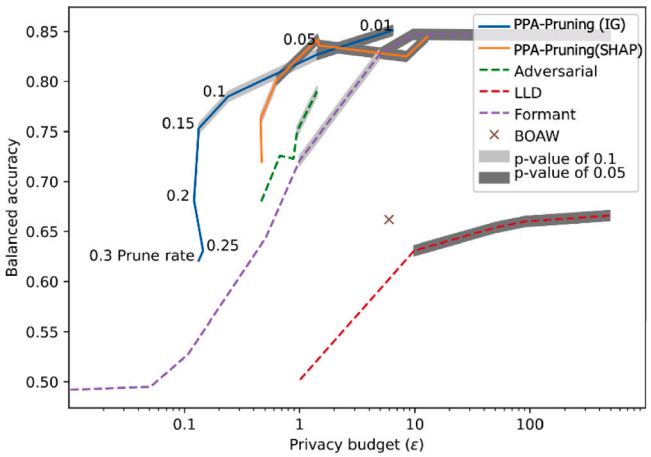

Fig. 5. Trade-off between TBI balanced accuracy and privacy budget of the proposed PPA-Pruning and the baselines.

The trade-off between TBI BA and privacy budget is depicted in Fig. 5 for the proposed adversarial pruning method and all the baselines. When considering SHAP and IG, their TBI BAs are competitive when the privacy budget is higher than 0.6. However, IG outperforms SHAP by achieving a lower privacy budget when the privacy budget is below 0.6. These results were derived from pruning rates ranging between 0.01 and 0.3. To further investigate the trade-off in TBI BA, we conducted a Wilcoxon signed-rank test across all cross-validations to assess the significance of the difference in TBI BA when nodes are pruned compared to the results without any pruning (no privacy-preservation). A privacy budget of 0.7 and 0.1 can be employed to maintain a p-value of 0.05 and 0.1, respectively. In comparison to all baselines, both IG and SHAP pruning achieve a higher BA and a lower privacy budget, signifying more privacy preservation in the extracted features. Despite the training instability in adversarial training using minimax, the adversarial training baseline outperforms DP methods applied to Low-Level Descriptor (LLD), formant coordinations, and Bag-of-Audio-Word (BOAW).

Although this approach provides differential privacy guarantees compared to the local-DP baseline, where Laplace noise is added to the Bag-of-Audio-Words (BOAW), Low-Level Descriptor (LLD), and Formant-based features, it notably degrades diagnostic accuracy. At a comparable privacy budget ϵ, our proposed pruning-based method outperforms the Laplace baseline by up to 11% in TBI detection accuracy. This result underscores the fact that selectively pruning nodes to remove speaker- and content-specific information preserves health-related features more effectively than indiscriminate noise addition.

5.3. Effects of privacy budget in speech processing task

Introducing noise to the features input to the DNN layer, which involves modifying the input through a product of weights, can contribute to preserving data privacy. The privacy budget, determining the extent of private information preservation, is influenced by both the input and classification functions. The privacy budgets required to achieve the minimum BA for TBI detection, speaker identification, and speech recognition are presented in Fig. 6. The WER in the speech recognition task is reported separately in Fig. 7 demonstrating that at the same privacy budget, different pruning methods yield different WER values. Note that in our context, higher WER is preferred because it indicates that more linguistic content has been suppressed, which is beneficial for privacy preservation. The model successfully achieves a perfect score for speaker verification without privacy concerns by utilizing the threshold at EER. A notable gain in preserving information about speaker identity occurs between a privacy budget of 0.1 and 5, accompanied by a significant decrease in accuracy at a privacy budget of 0.5. The performance of speech recognition decays the most between a privacy budget of 0.5 and 5. In summary, at a privacy budget of 5, minor privacy preservation can be achieved without any performance loss in TBI assessment. The range between a privacy budget of 0.5 and 5 exhibits a trade-off that can be leveraged when designing a privacy-preserving health assessment system based on spontaneous speech.

5.4. Relationship between pruning and privacy budget

While our experimental results (Fig. 4) show a trade-off between TBI detection accuracy and the privacy budget ϵ, the exact relationship between the number of pruned nodes ϵ is not governed by a simple closed-form expression. following the definition of DP, in this work, we rely on an empirical measurement of ϵ after each pruning iteration. Specifically, we compute the ratio of probabilities for the adversarial tasks (speaker identity and word recognition) on adjacent datasets, thereby quantifying how effectively pruning reduces the model's capacity to encode personal data.

Despite the intuitive expectation that removing more nodes should result in stronger privacy protections (i.e., a smaller ϵ), the non-linear nature of deep neural networks makes it difficult to derive a direct formula linking pruning rate to privacy budget. Empirically, we observe that pruning gradually lowers ϵ, but the rate of decrease can vary depending on which nodes are removed and how they contribute to speaker or linguistic information.

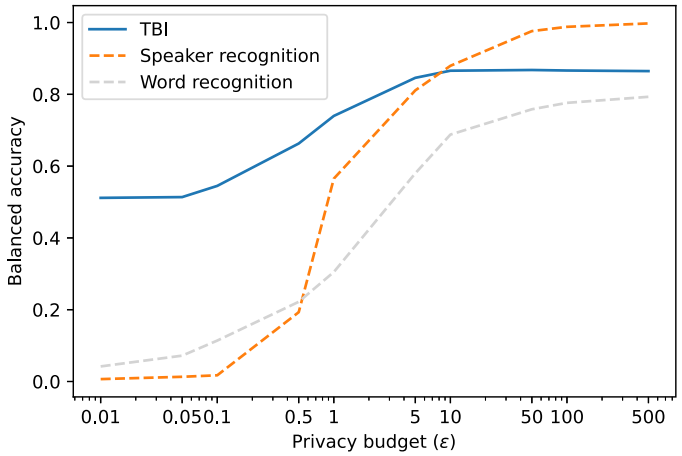

Fig. 6. Effects of privacy budget in preserving privacy in TBI detection, speaker recognition and word recognition balanced accuracy. Balanced accuracy for speaker recognition is computed from 1-EER.

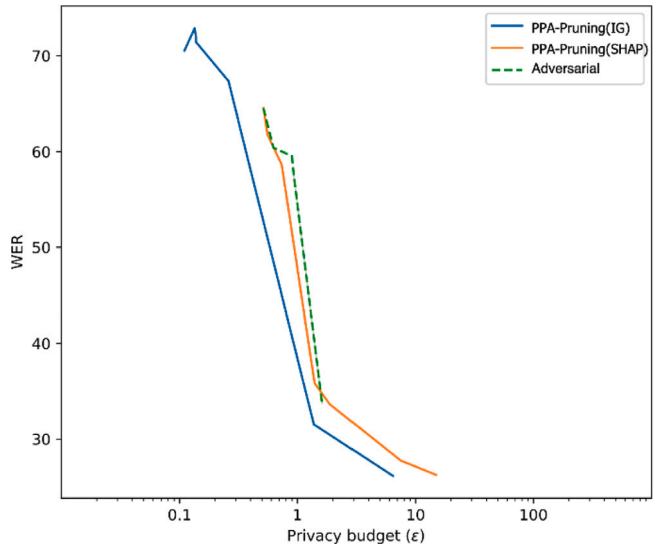

Fig. 7. Word error rate using the PPA-Pruning.

5.5. Limitations and future work

First, the proposed PPA-Pruning method only works on an end-to-end DNN that allows the computation of contribution scores at the feature level, which restricts its application to DNN-based feature extractors. It also considers only SHAP and IG as pruning criteria. While other variants of gradient analysis or contributions of nodes to network prediction have been previously proposed, they are often specific to tasks or network architectures. Future work may involve revisiting the criteria for network pruning, in addition to SHAP and IG.

More rigorous analyses could estimate the upper and lower bounds on ϵ by leveraging information-theoretic approaches or analyzing how the global sensitivity Δ shifts as the feature-space dimension decreases. Furthermore, a dimensionality-based sensitivity model may help formalize how pruning affects local or global DP guarantees. Our current study lays the groundwork for such work by proposing and demonstrating an empirical approach for measuring these changes. We plan to explore more formal theoretical approaches in future work.

6. Conclusion

The proposed PPA-Pruning aims to preserve personal data related to both speaker identity and speech content. This pruning method removes nodes and all their connections in the DNN, which significantly contributes to the recognition of personal data

during adversarial tasks. Notably, the pruning method exhibits more stability during training and outperforms adversarial training that utilizes minimax optimization. Additionally, it surpasses handcrafted features with noise introduced by DP. When comparing the pruning criteria of SHAP and IG, IG preserves more privacy than SHAP and baselines while achieving the same TBI detection accuracy. These results underscore the effectiveness of the proposed method for privacy preservation, which can be utilized to mitigate the risk of data leakage when collecting and analyzing speech data.

CRediT authorship contribution statement

Apiwat Ditthapron: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Emmanuel O. Agu:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Adam C. Lammert:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Apiwat Ditthapron reports financial support was provided by Defense Advanced Research Projects Agency. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ahmed, S., Chowdhury, A.R., Fawaz, K., Ramanathan, P., 2020. Preech: A system for {Privacy-Preserving} speech transcription. In: 29th USENIX Security Symposium (USENIX Security 20). pp. 2703–2720.

Al Mamun, K.A., Alhussein, M., Sailunaz, K., Islam, M.S., 2017. Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications. Futur. Gen. Comput. Sys. 66, 36–47.

Banerjee, D., et al., 2019. A deep trans. learning approach for improved post-traumatic stress disorder diagnosis. Knowl. & Info. Sys. 60 (3), 1693–1724.

Bavikatte, G., Subramanian, G., Ashford, S., Allison, R., Hicklin, D., 2021. Early identification, intervention and management of post-stroke spasticity: expert consensus recommendations. J. Central Nerv. Syst. Dis. 13.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A., 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Proc ICML. PMLR, pp. 794–803.

Chen, M., Lu, L., Wang, J., Yu, J., Chen, Y., Wang, Z., Ba, Z., Lin, F., Ren, K., 2023. VoiceCloak: Adversarial example enabled voice de-identification with balanced privacy and utility. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7 (2).

Coelho, C.A., Youse, K.M., Le, K.N., 2002. Conversational discourse in closed-head-injured and non-brain-injured adults. Aphasiology 16 (4–6), 659–672.

Ditthapron, A., Lammert, A.C., Agu, E.O., 2022. Continuous TBI monitoring from spontaneous speech using parametrized Sinc filters and a cascading GRU. IEEE J. Biomed. Heal. Inform. 26 (7), 3517–3528.

European Parliament, Council of the European Union, 2016. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Off. J. Eur. Union Apr.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: ACM International Conference on Multimedia. pp. 1459–1462.

Faul, M., Wald, M.M., Xu, L., Coronado, V.G., 2010. Traumatic brain injury in the United States; emergency department visits, hospitalizations, and deaths, 2002–2006. Centers for Disease Control and Prevention, National Center for Injury.

Gong, Y., Zhan, Z., Li, Z., Niu, W., Ma, X., Wang, W., Ren, B., Ding, C., Lin, X., Xu, X., et al., 2020. A privacy-preserving-oriented DNN pruning and mobile acceleration framework. In: Proceedings of the 2020 on Great Lakes Symposium on VLSI. pp. 119–124.

Gu, W., Liu, Z., Chen, L., Wang, R., Guo, C., Guo, W., Lee, K.A., Ling, Z.-H., 2024. A voice anonymization method based on content and non-content disentanglement for emotion preservation. In: 4th Symposium on Security and Privacy in Speech Communication. pp. 116–120.

Hua, H., Shang, Z., Li, X., Shi, P., Yang, C., Wang, L., Zhang, P., 2024. Emotional speech anonymization: Preserving emotion characteristics in pseudo-speaker speech generation. In: 4th Symposium on Security and Privacy in Speech Communication. pp. 55–60.

Illinois General Assembly, 0000. Biometric Information Privacy Act, Illinois General Assembly Public Acts 740 ILCS 14.

Jain, P., Kulkarni, V., Thakurta, A., Williams, O., 2015. To drop or not to drop: Robustness, consistency and differential privacy properties of dropout. arXiv preprint arXiv:1503.02031.

Kröger, J.L., Lutz, O.H.-M., Raschke, P., 2020. Privacy implications of voice and speech analysis–information disclosure by inference. In: IFIP International Summer School on Privacy and Identity Management. Springer, pp. 242–258.

Kulkarni, U., Hallad, S.S., Patil, A., Bhujannavar, T., Kulkarni, S., Meena, S., 2022. A survey on filter pruning techniques for optimization of deep neural networks. In: Proc I-SMAC. IEEE, pp. 610–617.

Lavania, C., Das, S., Huang, X., Han, K.J., 2023. Utility-preserving privacy-enabled speech embeddings for emotion detection. In: Interspeech 2023. ISCA, pp. 3612–3616.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I., 2018. Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118.

Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investig. Otolaryngol. 5 (1), 96–116.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Nautsch, A., et al., 2019. Preserving privacy in speaker and speech characterisation. Comput. Speech & Lang. 58, 441–480.

Norman, R.S., et al., 2013. Traumatic brain injury in veterans of the wars in Iraq and Afghanistan: Communication disorders stratified by severity of brain injury. Brain Inj. 27 (13–14), 1623–1630.
Nunes, J.A.C., Macêdo, D., Zanchettin, C., 2020. Am-mobilenet1d: A portable model for speaker recog.. In: Proc IJCNN. IEEE, pp. 1–8.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an ASR corpus based on public domain audio books. In: Proc ICASSP. IEEE, pp. 5206–5210.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Proc NeurIPS 32.
Prajapati, G.P., Singh, D.K., Amin, P.P., Patil, H.A., 2022. Voice privacy using CycleGAN and time-scale modification. Comput. Speech & Lang. 74, 101353.
Qian, K., Zhang, Y., Chang, S., Yang, X., Hasegawa-Johnson, M., 2019. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In: International Conference on Machine Learning. PMLR, pp. 5210–5219.
Ramanarayanan, V., Lammert, A.C., Rowe, H.P., Quatieri, T.F., Green, J.R., 2022. Speech as a biomarker: opportunities, interpretability, and challenges. Perspect. the ASHA Spec. Interes. Groups 7 (1), 276–283.
Renn, B.N., Pratap, A., Atkins, D.C., Mooney, S.D., Areán, P.A., 2018. Smartphone-based passive assessment of mobility in depression: Challenges and opportunities. Ment. Heal. Phys. Act. 14, 136–139.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proc. CVPR. pp. 4510–4520.
Scheliga, D., Mäder, P., Seeland, M., 2023. Dropout is not all you need to prevent gradient leakage. In: Proc. AAAI, vol. 37, (8), pp. 9733–9741.
Schmitt, M., Schuller, B., 2017. OpenXBOW: introducing the passau open-source crossmodal bag-of-words toolkit. J. Mach. Learn. Res. 18 (1), 3370–3374.
Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al., 2013. The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Interspeech. pp. 148–152.
Shamsabadi, A.S., Srivastava, B.M.L., Bellet, A., Vauquier, N., Vincent, E., Maouche, M., Tommasi, M., Papernot, N., 2023. Differentially private speaker anonymization. Proc. Priv. Enhancing Technol. 1, 98–114.
Srivastava, B.M.L., Bellet, A., Tommasi, M., Vincent, E., 2019. Privacy-preserving adversarial representation learning in ASR: Reality or illusion? In: Interspeech. pp. 3700–3704.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: Proc ICML. PMLR, pp. 3319–3328.
Talkar, T., Yuditskaya, S., Williamson, J.R., Lammert, A., Rao, H., Hannon, D., O'Brien, A., Vergara-Diaz, G., DeLaura, R., Sturim, D., et al., 2020. Detection of subclinical mild traumatic brain injury (mTBI) through speech and gait. In: Proc. Interspeech 2020. pp. 135–139.
Vadera, S., Ameen, S., 2022. Methods for pruning deep neural networks. IEEE Access 10, 63280–63300.
Vildjiounaite, E., Mäkelä, S.-M., Lindholm, M., Riihimäki, R., Kyllönen, V., Mäntyjärvi, J., Ailisto, H., 2006. Unobtrusive multimodal biometrics for ensuring privacy and information security with personal devices. In: International Conference on Pervasive Computing. Springer, pp. 187–201.
Wang, R., Chen, L., Lee, K.A., Ling, Z.-H., 2024. Asynchronous voice anonymization using adversarial perturbation on speaker embedding. In: Interspeech 2024. pp. 4443–4447.
Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., Gu, Q., 2019. On the convergence and robustness of adversarial training. In: Proc ICML. PMLR, pp. 6586–6595.
Williamson, J.R., Young, D., Nierenberg, A.A., Niemi, J., Helfer, B.S., Quatieri, T.F., 2019. Tracking depression severity from audio and video based on speech articulatory coordination. Comput. Speech & Lang. 55, 40–56.