ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446570

CONCEALING MEDICAL CONDITION BY NODE TOGGLING IN ASR FOR DEMENTIA PATIENTS

Wei-Tung Hsu, Chin-Po Chen, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

It is important to make automatic speech recognition (ASR) be inclusive to all users, including those with disorders. Besides model performances, privacy concerns, such as leakage of medical condition, are severe and harmful for this already vulnerable population. Hence, developing privacy-preserving machine learning (PPML) algorithms is important. Recent node cancellation strategies, while repeatedly showing their privacy protection efficacy, involve complex multi-branched structures with manually-tuned thresholds. In this work, we focus on learning ASR for dementia patients without revealing their medical condition. Specifically, we present a dementia attribute cancellation strategy (DACS) that trains a single toggling network in an end-to-end manner to toggle off particular node dimensions at ASR decoding, concealing a subject's dementia status. We show that using DACS can achieve 33% dementia protection efficacy (DPE), and further configuring for higher protection efficacy achieves 44% DPE, with only a slight decrease of 0.1% WER in ASR performance.

Index Terms— Privacy-preserving machine learning (PPML), automatic speech recognition, node-cancellation, dementia

1. INTRODUCTION

Emerging focus on Inclusive-AI aims at making AI services accessible to all users regardless of demographic categories, health status and other conditions [1]. An example is that people with disability (POD) are vulnerable populations to access reliable automatic speech recognition (ASR) [2, 3]. The speech of POD is prone to erroneous transcriptions (due to low-resourced condition) and unwanted violations of user privacy, e.g., exposure of medical condition [4]. Recently, privacy-preserving machine learning (PPML) algorithms are being actively developed to address privacy concerns. Specifically, learning-based PPML is commonly used to protect feature embeddings in voice technologies, such as ASR and speech emotion recognition (SER), from privacy leakage.

Learning-based PPML algorithms can be broadly categorized into attribute-elimination and node-cancellation approaches. The attribute-elimination approach often involves using a gradient reversal layer (GRL) or its variants to eliminate unwanted attribute's information in the feature space. For example, Jalal et al. applied a gradient reversal strategy to remove speaker identity from ASR embedding [5]. Similar strategies using a GRL to remove unwanted attributes can be found in privacy preservation studies [6, 7]. On the other hand, node-cancellation approach, inspired by disentanglement learning, concentrates attribute-specific information on particular dimensions and then masks those dimensions containing sensitive information. For example, Huang et al. proposed two variants of align-then-mask strategies for privacy-aware speech emotion recognition [8, 9], and showed superior results compared to those of attribute-elimination methods.

	People	Utterance	Age	Gender (Male / Female)
Train	AD: 54 HC: 54	1868	AD: 66.8±6.6 HC: 66.4±6.5	AD: 24 / 30 HC: 24 / 30
Test	AD: 24 HC: 24	800	AD: 66.1±7.4 HC: 66.1±7.1	AD: 11 / 13 HC: 11 / 13

Table 1. Demographics of healthy controls (HC) and people with Alzheimer's disease (AD) in the dataset we use

However, these recent methods contain major limitations. Firstly, their use of multi-task learning structures requires independent networks, to align each of the attribute-specific information with the node dimensions. This makes the method undesirably complex when dealing with multiple attributes. Secondly, the threshold in determining whether to mask a node has to be set manually. In this work, to address these limitations, we propose to learn a single toggling network that can directly toggle off nodes on-the-fly at inference to achieve privacy protection while maintaining task performances.

In specifics, this study contributes to an end-to-end node-cancellation learning strategy that protects user medical condition with a single toggling network. This toggling network takes frame-wise embedding and toggles off those nodes containing dementia status at ASR decoding. We evaluate this approach on task of ASR for dementia patients, where the main task is ASR and the sensitive attribute is their disease status. Our approach exhibits a 44% higher protection efficacy compared to that of the non-protected baseline while maintaining a competitive ASR performance. Additionally, we show that our network can be configured to adapt to different demands for protection efficacy and downstream task performance.

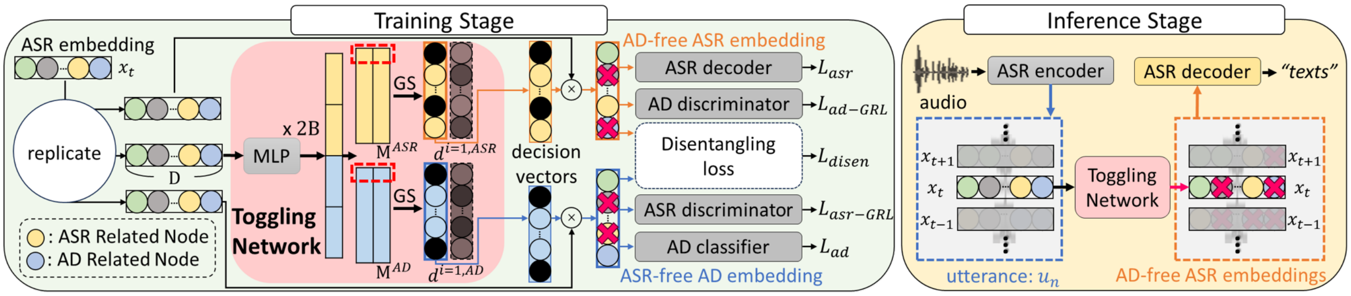

Fig. 1. The proposed DACS adds a toggling network in ASR to hide dementia status in frame-wise embeddings. The network is trained with two pairs of adversarial losses and disentangling loss, calculated from the AD-free and ASR-free branches.

2. METHODOLOGY

2.1. Dataset description

ADReSS Challenge dataset [10] containing transcribed speech recordings of participants during the Boston Diagnostic Aphasia Examination is used in this study. The average duration of each conversation session is 75.3 seconds, and there are 78 healthy controls (HC) and 78 people with Alzheimer's disease (AD), with their gender and age-matched. This dataset has been used in developing a wide range of speech algorithms for dementia diagnoses [11, 12, 13, 14]. In addition, we segmented each session into utterances and further removed utterances that were less than 0.1 seconds. We list the demographics of the dataset used in this study in Table 1.

2.2. The Dementia Attribute Cancellation Strategy

To protect dementia status in ASR, we utilize a toggling network (f) that takes an ASR embedding (x) and outputs a decision vector (m) to create a masked ASR embedding ($\tilde{x}$) using the equation: $\tilde{x} = mx = xf(x)$. $\tilde{x}$ is fed into an ASR decoder for word sequence, containing no dementia status. The training of the toggling network and ASR is described below. The choice of hyperparameters follows the original settings, with an exception of adjusting learning rate for the toggling network to 10^{-3} to ensure proper training of ASR branch (preventing WER from being around 50% to 80%).

2.2.1. The toggling network

The toggling network is a Multilayer Perceptron (MLP) with Gumbel-Softmax (GS) as output. The use of a GS function can generate differentiable-categorical outputs [15] with values approximating zeros and ones, and they can be jointly optimized during training. Additionally, to enable the toggling network to turn off nodes containing dementia status, we utilize a multi-task learning structure including two branches: an ASR-free branch and an AD-free branch. Each branch generates a masked ASR embedding to its downstream classifier.

The MLP expands the embedding into a vector with dimension $2{*}B{*}D$, where B represents the number of branches ($B = 2$), and this vector is rearranged into two matrices $M^{ASR} \in R^{2*D}$, and $M^{AD} \in R^{2*D}$. Then, a Gumbel-Softmax function

$$d_k^i = \frac{exp(\frac{s_k^i + g_k^i}{\tau})}{\sum_{j=1}^{2} exp(\frac{s_k^j + g_k^j}{\tau})} \tag{1}$$

takes each row of the matrix s_k^i and outputs corresponding decisions d_k^i, where $1 <= k <= D$ and $i \in \{1, 2\}$ represent the rows and columns of matrices; g_k^i is a random value sampled from Gumbel Distribution. The output is in the form of matrices, whose first and second column values are mutually exclusive (i.e., if $d_k^{i=1}$ = 1, then $d_k^{i=2}$ will be 0, and vice versa). Then, we arbitrarily select $d^{i=1}$ as the node-toggling decision vector, whose values approximating 1 or 0.

Further, we define losses to help cancelling the dementia state in masked ASR embedding $\tilde{x}$ while retaining ASR performances. For AD-free branch, we define ASR loss L_{asr} and reversed AD loss L_{ad-GRL} as an adversarial pair of losses. On the other hand, in ASR-free branch, we use another adversarial loss pair $L_{asr-GRL}$ and L_{ad} to create ASR-free embedding. This helps us to disentangle dementia attributes from AD-free embedding by enlarging the distance between that and ASR-free embedding. The disentanglement is done with disentangling loss, L_{disen}, implemented using AM-softmax [16], which penalizes the toggling model for insignificant difference between AD-free and ASR-free embeddings.

$$L_{disen} = \frac{-1}{N} \sum_{i=1}^{N} log \frac{e^{s(W_{y_i}^T f_i - m)}}{e^{s(W_{y_i}^T f_i - m)} + \sum_{j=1, j \neq y_i}^{C} e^{s W_j^T f_i}} \tag{2}$$

N is the number of samples; y_i represents the classes (ASR-free or AD-free embeddings), and C is the number of classes (C=2). $W_{y_i}^T f_i$ denotes the projection from input embedding to a single value, where W_{y_i} is the weight of a projection layer of class y_i and f_i represents the input embedding for sample i. The scaling factor s and margin m are set to 30.0 and 0.4.

The exact form of L_{asr} and L_{ad} are CTC loss and recall loss respectively. Here, instead of using conventional cross entropy-based losses for the AD downstream classifier, we used recall loss whose calculation depends on the recall score of the classification result. This approach allows us to emphasize protecting participants with AD labels. Similar to [17], we define L_{recall} as $L_{recall} = 1 - \frac{1}{N}\sum_{c=1}^{C}\sum_{k:y_k=c} w_c P_k^{y_k}$, where $k : y_k = c$ denotes the samples whose ground truth label y_i is class c. The weight for class c, w_c, is a tunable hyper-parameter controlling the type of participants to be protected. We set w_c for HC to 0.1 and for AD to 0.9 to enhance the protection on AD participants. $P_k^{y_k}$, the output probability of class y_k, is then calculated to represent the recall of that class. The other parameters are C, the number of classes, and N, the number of samples. L_{asr} and L_{ad} are computed from the downstream classifiers, and the reversed losses $L_{asr-GRL}$ and L_{ad-GRL} are from the downstream classifiers going through a gradient-reversal layer (GRL). The total loss, L_{toggle}, is as follows: $L_{toggle} = L_{asr} + L_{asr-GRL} + L_{ad-GRL} + L_{ad} + L_{disen}$. Our implementation of the toggling network is on github[1].

2.2.2. *Additional components*

The additional component is an AD classifier, serving as a downstream classifier of the ASR-free branch. It is a 2-layer MLP projecting D-dimensional embeddings to a 2-dimensional vector to calculate L_{ad} and L_{ad-GRL}. To calculate these losses using frame-level ASR embedding when AD label is given at a session-level, we perform aggregation. Let $u_n^P = \{x_1, x_2, ...x_T\}^P$ be embeddings of an utterance, where x_t is the ASR embedding at timestep t with length T; n denotes the n'th utterance of participant P. We calculate $\overline{u_n^P} = \frac{1}{T}\sum_{t=1}^{T} x_t$, to represent that utterance. Then, the participants are up-sampled to be paired with the utterance-level vectors $(\overline{u_n}^P, L^P)_n$. This component is pre-trained with the ASR system and then used for training the toggling network.

2.2.3. *ASR architecture*

Our ASR system is a data2vec-based end-to-end framework [18], taking 16kHz waveform of segmented audios as input. The output dimension, D, of the encoder is 1024. The decoder serves as the downstream classifier of the AD-free branch. We take the pretrained data2vec model, and fine-tune on the ADReSS dataset prior to toggling network training.

3. EXPERIMENTS

3.1. Experimental setup

In this work, we propose a privacy preserving ASR that aims to delete an user's dementia condition while maintaining ASR performance. WER is used to evaluate ASR; 1 - Acc(%), 1 - F1(%), and protection efficacy (DPE) are used to evaluate the dementia protection. Acc and F1 are the accuracy and f1-score of an AD classifier, which is seen as the attacker's AD model. DPE is a metric inspired by vaccine efficacy evaluation [19], which measures the attributable proportion of the protection by our method, defined as: DPE = (PRU - PRP)/PRU × 100%. PRU represents the risk (diagnosis exposure) of unprotected users, whereas PRP represents the risk of protected users. The risk of unprotected users is calculated by the true positives of the non-protect baseline model (Fine-tune) in dementia detection task (TP_n) divided by the number of AD subjects (N_{AD}). Similarly, PRP is calculated by the true positives of a protected model (TP_p) divided by N_{AD}.

3.2. Comparison methods

Multiple methods were implemented to compare with DACS:
Data2Vec [18]: Data2Vec with the pre-training setting 'data2vec-audio-large-960h' as an unprotected ASR system.
Fine-tune: fine-tuning the pre-trained Data2Vec model on the ADReSS dataset.
GRL: using the pair L_{asr} and L_{ad-GRL} to fine-tune pre-trained Data2Vec model on the ADReSS dataset, also known as the attribute-elimination approach for privacy preservation.
Single Toggling: using L_{asr} and L_{ad-GRL} (AD loss implemented with cross entropy) to optimize a toggling network sec.2.2.1, and using it for dementia nodes cancellation.
STOA [9]: State-of-the-art node-cancelling strategy in [9].

3.3. Experimental Result

3.3.1. *Comparison of DACS with other baseline models*

First, the WER of the fine-tuned Data2Vec model is 25.7%, which is similar to that of the state-of-the-art ASR model reported in this dataset [12] (please refer to the second row in Table 2). The fine-tuned model significantly improves ASR performance but also induces medical condition leakage, with dementia undetected rate of 20.83%. This model is set as the non-protected baseline for DPE computation (please refer to sec.3.1). Next, the GRL method improves DPE by 22% but degrades the ASR performance to 27.2% WER. The single toggling method retains the ASR performance (WER of 25.9%) but still lacks any protection (0% DPE). Furthermore, STOA model has an overall protection (1-Acc.) 2.09 % higher

	WER(%)	1-Acc. (%)	1-F1 (%)	DPE (%)
Data2Vec [18]	47.5	–	–	–
Fine-tune	25.7	20.83	21.74	0.00
GRL	27.2	27.08	31.71	22.23
Single Toggling	25.9	25.00	25.00	0.00
STOA [9]	25.9	22.92	22.45	-5.56
DACS	25.8	41.67	45.45	33.33

Table 2. ASR performances in WER(%) and privacy preservation performances in 1-Acc. (%), 1-F1(%) and dementia protection efficacy (DPE) (%) of different models

[1] https://github.com/Victoria-Wei/dementia-attribute-cancellation-strategy.

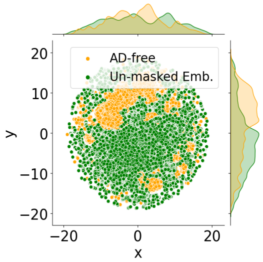

Fig. 2. t-SNE plots of original and toggled ASR embeddings

than that of Fine-tune, though the DPE is negative. The result implies that the learning strategy of the STOA model is to protect both HC and AD participants. Hence, when the protection efficacy is measured only on AD participants (revealing HC's condition is less relevant), the DPE decreases to negative. Lastly, when compared with Fine-tune (baseline), our proposed DACS improves DPE by 33% (the best-performing method) while retaining similar ASR performance.

3.3.2. Ablation study

We further perform ablation study to examine the effect of different components of our framework. First, when we use cross entropy instead of recall loss, the overall protection decreases to 29.17% (refer to column: 1-Acc, and the row GS+Disen-recall in Table 3), and the DPE decreases to 11%. When we remove disentangling loss (the row GS-Disen+recall), the overall protection decreases to 22.92%, and the DPE drops directly to 0%. This result shows that the toggling network doesn't enhance protection for dementia subjects, but protects a few more HC participants. Lastly, if the disentangling loss is not used and cross entropy is used instead of recall loss (the row GS-Disen-recall), the overall performance is just the same as that of the baseline model. In our ablation study, we observe that both recall and disentangling loss are important to improve the protection efficacy. Lastly, a t-SNE plot shows that the toggled AD-free embeddings shift to upper plane of the y-axis, while the original ASR embeddings distribute in the lower plane (Fig 2).

3.3.3. Aggressive and passive modes of DACS

With a trained DACS model, the node-toggling decision vector is determined by the score matrix M processed through the Gumbel-softmax function (sec.2.2.1). In this section, we illustrate how DACS can be configured into two distinct node-toggling modes: aggressive and passive. First, we retain the nodes already toggled on/off by the Gumbel-softmax output. Next, we calculate Δs, representing the difference between $s_k^{i=1}$ and $s_k^{i=2}$ ($s_k^{i=1} - s_k^{i=2}$), to sort the toggling decisions. A higher Δs indicates a higher likelihood of toggling a node on, while a lower Δs means otherwise. In the passive toggling mode, we toggle on the nodes corresponding to the top P% highest Δs among the remaining nodes, whilst in our aggressive mode, we further toggle off the nodes corresponding to P% lowest Δs.

	WER(%)	1-Acc. (%)	1-F1 (%)	DPE (%)
DACS	25.8	41.67	45.45	33.33
GS+Disen-recall	25.7	29.17	30.43	11.11
GS-Disen+recall	25.8	22.92	23.40	0.00
GS-Disen-recall	25.9	20.83	21.74	0.00

Table 3. ASR performances in WER(%) and privacy preservation performances in 1-Acc. (%), 1-F1(%) and dementia protection efficacy (DPE) (%) for ablation study

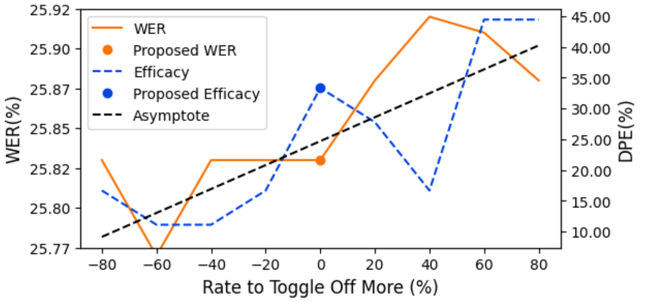

Fig. 3. WERs and DPE in aggressive (when x-axis values are positive) and passive modes (x-axis values are negative).

Fig 3 shows the result of the performance in our tasks by setting P∈{-80, -60, -40, -20, 0, 20, 40, 60, 80}, where the negative values of P indicate passive mode. There is a trend that the protection efficacy improves when more nodes are toggled off. Specifically, the protection efficacy significantly increases to 44% when over 60% additional nodes are off. In this setting, ASR performance only drops slightly to WER=25.9%. In passive mode, while the ASR performance improves with WER≈25.77% (60% additional nodes toggled on), the protection efficacy decreases to ≈10%. These results demonstrate that a trained DACS model can be configured to different demands of the downstream tasks, e.g., in this study it seems more suitable to configure it in aggressive mode.

4. CONCLUSION

This study presents a novel learning-based PPML strategy. In this research we propose a DACS that allows ASR service provider to protect dementia user's private medical condition while maintaining ASR performance. Compared to prior approaches of PPML, the use of a single toggling network is more desirable at ASR decoding. Our results show that DACS achieves 33% of DPE while maintaining almost the same ASR performance. We further demonstrate how P% can be configured to shift the optimal model to focus more on protection or task performance. In this study, we introduce a novel learning-based PPML strategy. An immediate direction is to study the strategy across multiple datasets, including larger scale of speech data and diverse medical conditions.

5. REFERENCES

[1] Tero Avellan, Sumita Sharma, and Markku Turunen, "Ai for all: defining the what, why, and how of inclusive ai," in *Proceedings of the 23rd International Conference on Academic Mindtrek*, 2020, pp. 142–144.

[2] Dhanya Srinivasan, B Bharathi, Thenmozhi Durairaj, et al., "Ssncse_nlp@ lt-edi-acl2022: Speech recognition for vulnerable individuals in tamil using pre-trained xlsr models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 317–320.

[3] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris, "Toward fairness in ai for people with disabilities sbg@ a research roadmap," *ACM SIGACCESS Accessibility and Computing*, , no. 125, pp. 1–1, 2020.

[4] Jennifer Williams, Karla Pizzi, Shuvayanti Das, and Paul-Gauthier Noé, "New challenges for content privacy in speech and audio," *ISCA SPSC Symposium 2022*.

[5] Md Asif Jalal, Pablo Peso Parada, Jisi Zhang, Karthikeyan Saravanan, Mete Ozay, Myoungji Han, Jung In Lee, and Seokyeong Jung, "On-device speaker anonymization of acoustic embeddings for asr based onflexible location gradient reversal layer," *Proc. Interspeech 2023*, 2023.

[6] Diep Luong, Minh Tran, Shayan Gharib, Konstantinos Drossos, and Tuomas Virtanen, "Representation learning for audio privacy preservation using source separation and robust adversarial learning," *arXiv preprint arXiv:2308.04960*, 2023.

[7] Peng-Fei Zhang, Guangdong Bai, Hongzhi Yin, and Zi Huang, "Proactive privacy-preserving learning for cross-modal retrieval," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–23, 2023.

[8] Yu-Lin Huang, Bo-Hao Su, Y-W Peter Hong, and Chi-Chun Lee, "An attribute-aligned strategy for learning speech representation," *Proc. Interspeech 2022*, 2021.

[9] Yu-Lin Huang, Bo-Hao Su, Y-W Peter Hong, and Chi-Chun Lee, "An attention-based method for guiding attribute-aligned speech representation learning," *Proc. Interspeech 2022*, pp. 5030–5034, 2022.

[10] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *Proceedings of INTERSPEECH 2020*, 2020.

[11] Loukas Ilias, Dimitris Askounis, and John Psarras, "Detecting dementia from speech and transcripts using transformers," *Computer Speech & Language*, p. 101485, 2023.

[12] Tianzi Wang, Jiajun Deng, Mengzhe Geng, Zi Ye, Shoukang Hu, Yi Wang, Mingyu Cui, Zengrui Jin, Xunying Liu, and Helen Meng, "Conformer based elderly speech recognition system for alzheimer's disease detection," *Proceedings of Interspeech 2022*, 2022.

[13] Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov, "Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models," *ACL 2022*, 2022.

[14] Shahla Farzana, Ashwin Deshpande, and Natalie Parde, "How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 37–48.

[15] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[16] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[17] Junjiao Tian, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-pang Chiu, and Zsolt Kira, "Recall loss for imbalanced image classification and semantic segmentation," *openreview.net*.

[18] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[19] Walter A Orenstein, Roger H Bernier, Timothy J Dondero, Alan R Hinman, James S Marks, Kenneth J Bart, and Barry Sirotkin, "Field evaluation of vaccine efficacy.," *Bulletin of the World Health Organization*, vol. 63, no. 6, pp. 1055, 1985.