ENHANCING PRIVACY THROUGH DOMAIN ADAPTIVE NOISE INJECTION FOR SPEECH EMOTION RECOGNITION

Tiantian Feng[⋆†]*, Hanieh Hashemi*[†]*, Murali Annavaram*[†]*, Shrikanth S. Narayanan*[⋆†]

[⋆]Signal Analysis and Interpretation Lab (SAIL)
[†]Department of Electrical and Computer Engineering, University of Southern California
tiantiaf@usc.edu,hashemis@usc.edu,annavara@usc.edu,shri@sipi.usc.edu

ABSTRACT

Speech **E**motion **R**ecognition (**SER**) techniques have gained considerable interest in many applications including smart virtual assistants and health state tracking. SER systems often acquire and transmit speech data collected at the client-side to remote cloud platforms for inference and decision making. However, speech data carries rich information not only about emotions conveyed in vocal expressions, but also other sensitive demographic traits, such as gender, age, and language background. It is desirable to select only features that are necessary for the emotion classification while protecting sensitive features. However, there are some features that are necessary for emotion classification. These features may also reveal other demographic traits. In this work, we propose a method to improve inference privacy for sensitive features by injecting noise into the input speech data, but without degrading the SER system performance. The approach combines a noise representation learning architecture, called Cloak [1], with adversarial training to keep relevant information inside the data for emotion classification while removing information that would enable inferring sensitive demographic attributes. Experimental results show that our method can effectively prevent inference of sensitive demographic information, and that the improved privacy comes at a cost of only a minor utility loss for the emotion classification.

Index Terms— machine learning, statistical privacy, noise enjection, adversarial training, speech emotion, fairness

1. INTRODUCTION

Speech emotion recognition (SER) aims to identify emotional states conveyed in vocal expressions. Speech emotion recognition systems are currently deployed in a wide range of applications such as in smart virtual assistants [2], medical diagnoses [3, 4], and education [5]. The same speech signal carries rich information about individual traits (e.g., age, gender) and states (e.g., health), many of which are deemed sensitive from an application point of view. The present paper considers how privacy with respect to such sensitive traits discernible from speech can be achieved without compromising the SER accuracy. A real-time SER system typically has three parts: data acquisition, data transfer, and classification [6]. Prior approaches to protect privacy have eliminated the collection of raw audio data and instead extracted acoustic features from the raw speech signals on the client-side. Thus the client only shares the acoustic features to obfuscate the actual content of the conversation being disclosed to any third party, including remote cloud servers [7]. The acoustic features needed for each classification task may differ, and more critically some features that are needed for SER may also disclose some of the sensitive traits [8]. Attribute inference attacks would aim to reveal individuals' sensitive attributes (e.g., age and gender) that they did not intend or expect to share [9, 10]. These undesired/unauthorized usage of data may occur when the service provider is not trustworthy or when an intruder attacks the cloud system [11, 12, 13].

Traditional approaches to improve inference privacy are based on cryptography solutions such as Homomorphic Encryption and Multi-Party Computing [14, 15, 16]. However, these solutions have a massive computational burden, thus creating prolonged delays. Another approach to prevent attribute inference attacks is through noise perturbation. These methods improve input privacy by transforming original data input $\mathbf{x}$ to $\mathbf{x}'$ on the client-side by adding noise vectors to the input features x. A recently proposed framework to prevent attribute inference is called Cloak [1]. Cloak proposes noise injection on the input to keep relevant features for the primary application while removing the irrelevant features for this primary task. However, hiding the irrelevant features for the primary task does not necessarily prevent the adversary to infer other sensitive attributes. That is because some of the relevant features that are necessary for a primary task, which are not protected by noise, can in fact leak information about the sensitive attributes.

The goal of our work is to minimize the risk of sensitive attribute inference while maintaining the utility of the input features for the primary task in remote machine learning settings. While protecting any arbitrary attribute inference is challenging to analyze and evaluate, in this work we focus on domain-specific attributes that a user is particularly worried about. We propose to combine the Cloak framework and adversarial training to learn a noise function ϕ that balances the utility for the primary task and inference privacy for domain-specific secondary tasks. We evaluate our proposed noise injection method over SER (primary task) and gender prediction (secondary task) using IEMOCAP [17], Crema-D [18], and MSP-Improv [19] datasets. We show that our approach can effectively remove irrelevant features for the primary task while adding noise on the selected features to prevent secondary task inference.

2. SER DATA SETS

2.1. IEMOCAP

The IEMOCAP database [17] was collected using multi-modal sensors that capture motion, audio, and video of acted human interactions. The corpus contains 10,039 utterances from ten subjects targeting expressing categorical emotions. In addition, the utterances are divided into improvised conditions and scripted conditions based on whether the utterance is from a fixed script. We choose to remove data from script conditions as suggested in previous work [20].

2.2. CREMA-D

The CREMA-D [18] corpus is a multi-modal database of emotional speech collected from 91 actors, 48 of whom are male and 43 are female. The set contains 7,442 speech recordings that simulate emotional expressions including happy, sad, anger, fear, and neutral.

2.3. MSP-Improv

The MSP-Improv [19] corpus was created to study naturalistic emotions captured from improvised scenarios. The corpus includes audio and visual data of utterances spoken in natural condition (2,785 utterances), target condition (652 target utterances in improvised scenario), improvised condition (4,381 utterances from the remainder of the improvised scenario), and read condition (620 utterances). We decide to use the data only from the improvised scenarios.

2.4. Data setup

We combine these three data sets into a single data set and then divided it into three small subsets: 1. Data ($\mathbf{D_p}$) to train the primary task model, which has utterances from 40% of speakers in each subcorpus; 2. Data ($\mathbf{D_{adv}}$) to train the secondary task model (adversarial), which has utterances from 40% of speakers in each sub-corpus; 3. Data ($\mathbf{D_e}$) for evaluation, which contains the rest 20% of speakers in each subcorpus. In this way, we can train the primary task model and secondary task model separately. We created 5 different sets of data described above, with no overlap of the test data between each data set. We perform the same training process and testing procedure on each data set to obtain the reported average prediction results.

3. FEATURES AND LABELS

SER inference can be performed using a variety of ML models where each model requires a specific input format. For this work, we apply mel-spectrogram as the input for SER. We apply a sequence of overlapping Hamming windows to the raw speech signal, with a window size of 50 msec and time shift of 10 msec. The discrete Fourier transform (DFT) of length 800 is then calculated for each frame and the mel-spectrogram data is then generated from the DFT. We set the dimension for the mel-spectrogram as 128. Then, the mel-spectrogram of each utterance is divided into 2-second segments, with an overlap of 1 second (in training) or 0.5 seconds (in testing) between segments. Due to the data imbalance issue in the IEMOCAP corpus, previous works choose the most four common emotions (neutral, sad, happiness, and anger) for designing the SER experiment [20]. In this work, we pick these four emotion classes also because all three corpora contain these labels. Table 1 shows the label distribution of mel-spectrogram segments in these corpora.

4. METHOD

4.1. Problem Setup

We follow a setup in which we have labeled data set $\mathbf{D}$ including speech samples $\mathbf{x}_1, ..., \mathbf{x}_n$, where $\mathbf{x}_i \in \mathbb{R}^m$, primary task labels $y_1, ..., y_n$, and secondary task labels $z_1, ..., z_n$. In our present study, the primary task is SER and the secondary task is gender prediction. We define f_θ as the classifier with parameters θ that predicts the primary task labels y_i. We name f_g as the adversary classifier with parameters g to predict secondary task z_i. We also define two feature groups of conductive features and non-conductive features. The conductive features, $\mathbf{c} \in \mathbf{x}$, are more relevant features associated with the primary task, and the non-conductive feature set, $\mathbf{u} \in \mathbf{x}$, represents features that are less important for the primary task. One of our goals is to train noise perturbation function ϕ to map the input data $\mathbf{x}_i$ as $\phi(\mathbf{x}_i)$, such that the conductive features are kept for the primary task while the adversary classifier f_g is not able to correctly predict secondary label z_i using noisy representation $\phi(\mathbf{x}_i)$.

	Neutral	Happy	Sad	Angry	All
IEMOCAP	4950	3731	3379	1432	13492
CREAM-D	2726	1593	998	1637	6954
MSP-Improv	7600	4606	4312	2655	19173

Table 1. Statistics of emotion labels from extracted mel-spetrogram segments in three different data sets.

4.2. Primary task model and secondary task model

Figure 1 shows the architecture of the primary and the secondary task model used in this work. We construct both models with the same structure for simplicity. First, the mel-spectrogram data is fed into the 2d convolutional layers to generate a 128-channel representation. Batch normalization and ReLU activation functions are applied after each convolutional layer. For brevity, the ReLU activation function is not shown in the figure. The learned representation is then sent to the Recurrent Neural Network (RNN). Finally, we flatten the RNN model's output and pass it to the fully connected layer for classification. We use data set $\mathbf{D_p}$ and $\mathbf{D_{adv}}$ to train f_θ and f_g, respectively.

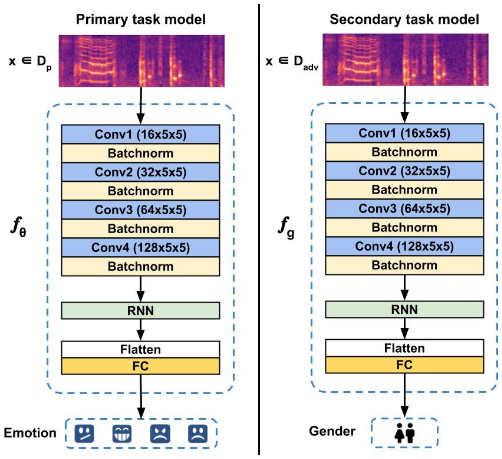

Fig. 1. Architecture of the primary task model in predicting emotion and secondary task model in predicting gender.

4.3. Cloak - Feature Selection Designed for a Single Task

To select conductive features from input data $\mathbf{x}$, the Cloak framework proposes to learn the function ϕ through the measure of mutual information. Specifically, the noise function ϕ is defined below:

$$\phi(\mathbf{x}) = \mathbf{x} + \mathbf{r}, \text{ where } \mathbf{r} \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

The parameters σ in diagonal covariance matrix Σ and μ are trainable tensors that are learned during the training process. Information leakage is defined as the Mutual Information (MI) between client's raw data and its noisy representation [21, 1, 22, 23]. Formally, we define the MI function as I [24]. Cloak attempts to minimize the mutual information $I(\phi(\mathbf{x}); \mathbf{u})$ between $\phi(\mathbf{x})$ and non-conductive feature sets $\mathbf{u}$ while maximizing the mutual information $I(\phi(\mathbf{x}); \mathbf{c})$ between $\phi(\mathbf{x})$ and conductive feature sets c. To optimize

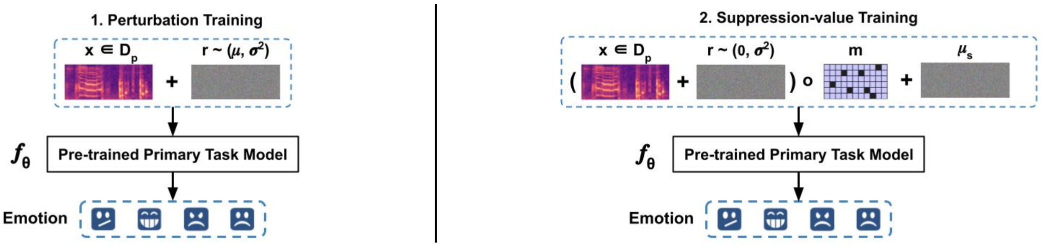

Fig. 2. Training procedures of Cloak which include perturbation training and suppression-value training.

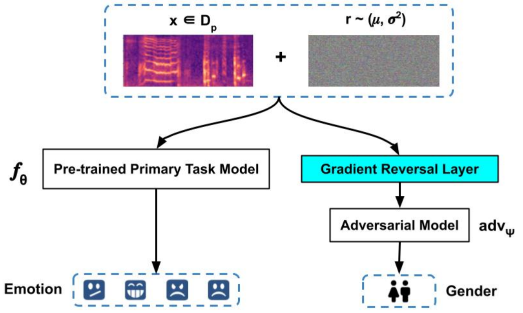

Fig. 3. Architecture of the Cloak + Adversarial training setup

this problem, we first need to find the upper bound for the term $I(\phi(\mathbf{x});\mathbf{u})$. Since $\mathbf{u}$ is a subset of $\mathbf{x}$, then the following holds:

$$I(\phi(\mathbf{x});\mathbf{u}) \leq I(\phi(\mathbf{x});\mathbf{x}) = H(\phi(\mathbf{x})) - H(\phi(\mathbf{x})|\mathbf{x}) \quad (2)$$

$$\min \ -I(\phi(\mathbf{x});\mathbf{c}) \quad (3)$$

Consequently, we convert the original problem to optimizing the upper bound of $I(\phi(\mathbf{x});\mathbf{u})$. This optimization problem has been proved to be the same as minimizing the following loss function [1]:

$$L_c = -\log(\frac{1}{m}\sum_{j=0}^{m}\sigma_j^2) \quad (4)$$

Secondly, we need to find the lower bound for the term $I(\phi(\mathbf{x});\mathbf{c})$. This can be seen as minimizing the empirical cross-entropy loss over all training samples [1]. As a result, we want to minimize the following loss function, where L is the cross-entropy loss function:

$$\min_{\sigma,\mu} \lambda L_c + L(f_\theta(\phi(\mathbf{x})), y) \quad (5)$$

The lemma and accompanying proof for the upper bound of $I(\phi(\mathbf{x});\mathbf{u})$ and lower bound of $I(\phi(\mathbf{x});c)$ can be found in [1]. λ in equation 5 is a hyper-parameter that sets the focus on minimizing the first term $I(\phi(\mathbf{x});\mathbf{u})$. The training process in learning σ and μ is named as perturbation training in the original paper shown in figure 2. In order to take gradients over σ and μ in $\phi(\mathbf{x})$, the noise $\mathbf{r}$ can be rewritten as $\mathbf{r} = \sigma \odot \mathbf{e} + \mu$, where $\mathbf{e} \sim N(0,1)$. In addition, we redefine σ shown in equation 6 to constrain the range of σ, since the variance cannot be negative. During the training process, we freeze the parameters in f_θ and take the gradients with respect to σ, μ.

$$\sigma = \frac{1 + tanh(\rho)}{2} \quad (6)$$

The second part of the training is called suppression-value training which aims to find representations to replace the non-conductive features. The learned noise function $\phi(\mathbf{x})$ denotes the importance of features from input data $\mathbf{x}$. Intuitively, higher σ_j represents lower importance of the associated feature in predicting the primary task, since such feature can tolerate higher noises. Then, the Cloak method selects the conductive features by applying a cutoff threshold T. We can define a mask m, where $m_j = 0$, if $\sigma_j > T$, otherwise $m_j = 1$. We can rewrite noise function $\phi(\mathbf{x})$ as $\phi(\mathbf{x}) = (\mathbf{x}+\mathbf{r}) \odot \mathbf{m} + \mu_\mathbf{s}$, where $\mathbf{r} \sim (0, \sigma)$, and $\mu_\mathbf{s}$ is set to replace non-conductive features. We set initial $\mu_\mathbf{s}$ as μ, and we wish to minimize the following loss function with respect to $\mu_\mathbf{s}$:

$$\min_{\mu_s} L(f_\theta(h_{\mu_s}(\mathbf{x})), y) \quad (7)$$

4.4. Adversarial Training - Domain Specific Attribute Removal

Please note in the cloak, secondary task does not play a role in defining objective function. To further decrease the accuracy of the secondary classifier, we propose to minimize the mutual information $I(\phi(\mathbf{x});z)$ between noise representation $\phi(\mathbf{x})$ and secondary task label. Since it is infeasible to estimate the mutual information between two arbitrary distributions, this problem is typically turned into the following adversarial training objectives as suggested in [25]:

$$\min_{\psi} L(adv_\psi(\phi(\mathbf{x})), z); \quad \max_{\sigma,\mu} L(adv_\psi(\phi(\mathbf{x})), z) \quad (8)$$

Here adv_ψ represents an adversary in inferring the secondary task. The objective of training $\phi(\mathbf{x})$ is to create the noise representation that is minimally informative of the secondary task. In practice, this optimization problem is usually implemented using the gradient-reversal layer (GRL) [26]. Here, the gradient reversal layer g_α is inserted between $\phi(\mathbf{x})$ and adv_ψ. During the forward pass, the GRL acts as the identity function, while it scales the gradients passed through by $-\alpha$ in the back-propagation stage. Consequently, we aim to minimize the loss function below:

$$\min_{\psi,\sigma,\mu} L(adv_\psi(g_\alpha(\phi(\mathbf{x}))), z) \quad (9)$$

In the end, we aim to combine the Cloak framework with adversarial training to select the most important features for the primary task, while minimizing the attribute inference for the secondary task and potentially other tasks. We follow the training procedure that performs the perturbation training and suppression-value training in sequence. The model architecture is shown in Figure 3. Formally, we aim to minimize the following loss function:

$$\min_{\psi,\sigma,\mu} \lambda L_c + L(f_\theta(\phi(\mathbf{x})), y) + L(adv_\psi(g_\alpha(\phi(\mathbf{x}))), z) \quad (10)$$

5. RESULTS

5.1. Primary task model and secondary task model

First, we show the prediction results of the primary task model f_θ for SER and the secondary task model $f_\mathbf{g}$ (adversarial) of predicting gender. We report the results for the main test set and each subcorpus. We implement both models and training processes using PyTorch. We choose ReLU as the activation function and the dropout rate as 0.2. We set the learning rate to 10^{-4}, and apply Adam optimizer to train both models. We also augment the speech samples

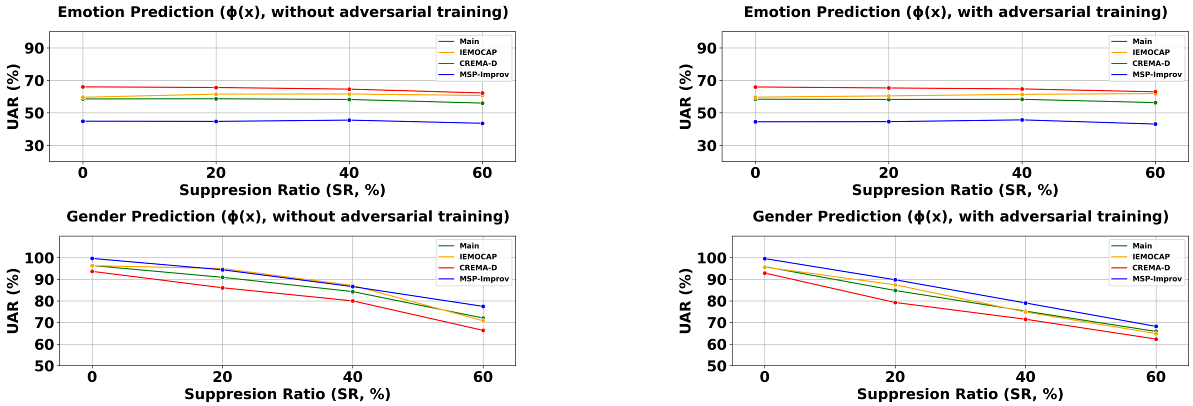

Fig. 4. Prediction results for both primary task model and secondary task model using noise perturbation $\phi(x)$.

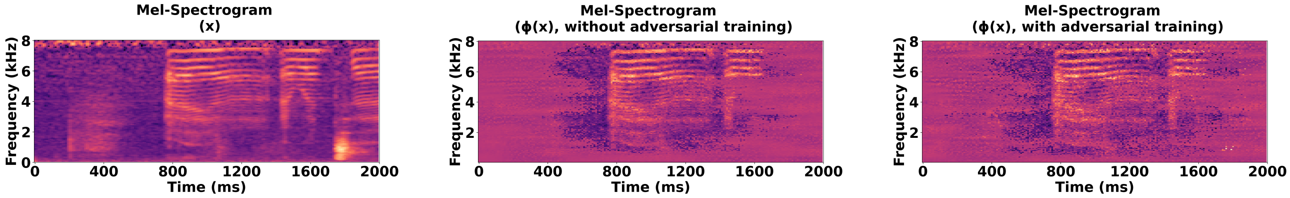

Fig. 5. An example of noise representation (SR=60%) trained with and without adversarial setup.

	Emotion Model ($\mathbf{f}_\theta$)		Gender Model ($\mathbf{f_g}$)	
	Acc	**UAR**	**Acc**	**UAR**
Main	57.7%	58.9%	97.2%	96.8%
IEMOCAP	57.6%	59.7%	96.3%	96.6%
CREAM-D	69.8%	66.9%	93.5%	93.8%
MSP-Improv	45.8%	44.9%	99.6%	99.7%

Table 2. Prediction results for both primary task model and secondary task model using original input $\mathbf{x} \in \mathbf{D_e}$. We report results for whole $\mathbf{D_e}$ and also each subcorpus in $\mathbf{D_e}$.

by adding white noises within each speaker based on the emotion label distribution. We aggregate the predictions among all segments in an utterance as the final prediction. We report both accuracy and unweighted average recall (UAR) results in Table 2. We can observe that the primary task model f_θ can predict emotion with a UAR of 58.9% on the test data set $\mathbf{D_e}$. The primary task model yields highest performance on CREMA-D data (UAR: 66.9%) and lowest on MSP-improv data set (44.9%). We also find that the secondary task model can infer gender with high accuracy (97.2%) and high UAR (96.8%), and this is consistent for each subcorpus.

5.2. Noise perturbation

In this experiment, we trained the noise representation with and without adversarial training. We first perform the perturbation training to learn μ and σ. We initialize the noise model $\phi(\mathbf{x})$ with $\mu = \mathbf{0}$ and $\rho = -\mathbf{10}$ in $\sigma = \frac{1+tanh(\rho)}{2}$. We set λ in the loss function as 0.1 in equation 10. After we learned the noise map σ in perturbation training, we empirically choose the threshold value T as the 20%, 40%, 60% percentile in σ for the suppression-value training process. We use suppression ratio (SR) to represent the amount of non-conductive features to remove, and higher SR (lower threshold value T) means more features are suppressed (or removed). We choose $\alpha = 0.1$ in GRL when adversarial training setup is included.

The results of the primary task model in predicting emotions and the secondary task model (adversarial) in predicting gender using noise representation $\phi(\mathbf{x})$ is shown in Figure 4. SR = 0% represents the noise model $\phi(\mathbf{x})$ trained with only perturbation training. We can observe that lower SR is associated with less performance decrease in predicting emotions. We also find that emotion prediction performance (UAR score) decreases only from 57.5% to 56.0% and to 56.3% with adversarial training and without adversarial training, respectively. This shows that the selected feature are informative of the SER task. In addition, we find that noise representation trained without adversarial setup can decrease the performance of the secondary task (gender) from > 95% to 73.0% when SR = 60%. The secondary task prediction UAR further decreases to 65.8% by adding the adversarial training in learning noise model $\phi(\mathbf{x})$. We also plot an example of learned noise representation in Figure 5. We can observe that $\phi(\mathbf{x})$ trained using adversarial setup have more features suppressed in the range of 500-2k Hz.

6. CONCLUSION AND FUTURE WORK

In this work, we propose a combination of Cloak and adversarial training to learn a noise inject function $\phi(\mathbf{x})$ to balance the utility for the SER and inference privacy for sensitive demographic information. The code used in this work is under [1]. Our results show that by adding the adversarial training in the Cloak framework, the injected noise can effectively prevent demographic attributes, such as gender, from being inferred. The prediction results also show that the injected noise on original input data does not decrease the emotion recognition performance. In the future, we plan to extend our current work to multitask learning scenarios to hide sensitive information inside data for multiple demographic labels.

7. ACKNOWLEDGMENT

The work was supported by the USC-Amazon Center on Trusted AI.

[1] https://github.com/usc-sail/speech-emotion-privacy-trust

8. REFERENCES

[1] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh, "Not all features are equal: Discovering essential features for preserving prediction privacy," in *Proceedings of the Web Conference 2021*, 2021, pp. 669–680.

[2] Ming-Che Lee, Shu-Yin Chiang, Sheng-Cheng Yeh, and Ting-Feng Wen, "Study on emotion recognition and companion chatbot using deep neural network," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 19629–19657, 2020.

[3] Srinivasan Ramakrishnan and Ibrahiem MM El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[4] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 189–196, September 2017.

[5] Wu Li, Yanhui Zhang, and Yingzi Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Third International Conference on Natural Computation (ICNC 2007)*. IEEE, 2007, vol. 5, pp. 809–813.

[6] Shashidhar G Koolagudi and K Sreenivasa Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[7] Tiantian Feng, Amrutha Nadarajan, Colin Vaz, Brandon Booth, and Shrikanth Narayanan, "Tiles audio recorder: an unobtrusive wearable solution to track audio activity," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, 2018, pp. 33–38.

[8] Mimansa Jaiswal and Emily Mower Provost, "Privacy enhanced multimodal neural representations for emotion recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7985–7993, Apr. 2020.

[9] Neil Zhenqiang Gong and Bin Liu, "Attribute inference attacks in online social networks," *ACM Transactions on Privacy and Security (TOPS)*, vol. 21, no. 1, pp. 1–30, 2018.

[10] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh, "Privacy in deep learning: A survey," *arXiv preprint arXiv:2004.12254*, 2020.

[11] Josep Domingo-Ferrer, Oriol Farras, Jordi Ribes-González, and David Sánchez, "Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges," *Computer Communications*, vol. 140, pp. 38–60, 2019.

[12] P Ravi Kumar, P Herbert Raj, and P Jelciana, "Exploring data security issues and solutions in cloud computing," *Procedia Computer Science*, vol. 125, pp. 691–697, 2018.

[13] Hamed Tabrizchi and Marjan Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," *The journal of supercomputing*, vol. 76, no. 12, pp. 9493–9532, 2020.

[14] Nitin Agrawal, Ali Shahin Shamsabadi, Matt J Kusner, and Adrià Gascón, "Quotient: two-party secure neural network training and prediction," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1231–1247.

[15] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *International conference on machine learning*. PMLR, 2016, pp. 201–210.

[16] Miguel Dias, Alberto Abad, and Isabel Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2057–2061.

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[18] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[19] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[20] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[21] Paul Cuff and Lanqing Yu, "Differential privacy as a mutual information constraint," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 43–54.

[22] Chuan Guo, Awni Hannun, Brian Knott, Laurens van der Maaten, Mark Tygert, and Ruiyu Zhu, "Secure multiparty computations in floating-point arithmetic," *arXiv preprint arXiv:2001.03192*, 2020.

[23] Hanieh Hashemi, Yongqin Wang, Chuan Guo, and Murali Annavaram, "Byzantine-robust and privacy-preserving framework for fedml," *arXiv preprint arXiv:2105.02295*, 2021.

[24] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[25] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon, "Learning controllable fair representations," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2164–2173.

[26] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.