APSIPA Transactions on Signal and Information Processing, 2023, 12, e17

Overview Paper

A Review of Speech-centric Trustworthy Machine Learning: Privacy, Safety, and Fairness

Tiantian Feng*, Rajat Hebbar†, Nicholas Mehlman†, Xuan Shi†, Aditya Kommineni† and Shrikanth Narayanan

University of Southern California, Los Angeles, CA, USA

ABSTRACT

Speech-centric machine learning systems have revolutionized a number of leading industries ranging from transportation and healthcare to education and defense, fundamentally reshaping how people live, work, and interact with each other. However, recent studies have demonstrated that many speech-centric ML systems may need to be considered more trustworthy for broader deployment. Specifically, concerns over privacy breaches, discriminating performance, and vulnerability to adversarial attacks have all been discovered in ML research fields. In order to address the above challenges and risks, a significant number of efforts have been made to ensure these ML systems are trustworthy, especially private, safe, and fair. In this paper, we conduct the first comprehensive survey on speech-centric trustworthy ML topics related to privacy, safety, and fairness. In addition to serving as a summary report for the research community, we highlight several promising future research directions to inspire researchers who wish to explore further in this area.

*Corresponding author: Tiantian Feng, tiantiaf@usc.edu.

†These authors contributed equally to this work.

Received 13 December 2022; Revised 24 February 2023

ISSN 2048-7703; DOI 10.1561/116.00000084

1 Introduction

In the last few years, machine learning (ML), particularly deep learning, has empowered tremendous breakthroughs in a variety of research fields and applications, including natural language processing [47], image classification [83], video recommendation [45], healthcare analysis [119], and even mastering the chess game [152]. The deep learning model typically consists of multiple processing layers with a combination of both linear and non-linear computations. Although training a deep learning model with the multi-layer architecture demands the accumulation of massive datasets and access to large-scale computational infrastructures [18], the trained model usually achieves state-of-the-art (SOTA) performance compared to the traditional modeling approaches. The broad success of deep learning has enabled a more profound understanding of the human condition (state, trait, behavior, and interaction) and revolutionized technologies that support and enhance human experiences. Alongside the success that ML has had in these areas, significant progress has also been made in speech-centric ML.

Speech is a natural and prominent form of communication between humans that exists in almost every spectrum of human life, whether chatting with friends, discussing with colleagues, or having a remote call with the family. The advancement in speech-centric machine learning has enabled the ubiquitous usage of smart assistants such as Siri, Google Voice, and Alexa. In addition, speech-centric modeling has created numerous research topics in human behavior understanding, human-computer interface (HCI) [41], and social media analysis, involving several widely used speech modeling techniques like automatic speech recognition [110], speech emotion recognition [5], automatic speaker verification [88], and keyword spotting [179].

Despite the prospect of the broad deployment of ML systems in a wide range of speech-centric applications, two intertwined challenges remain unaddressed in most of these systems: understanding and illuminating the rich diversity across people and contexts while creating trustworthy ML technologies that work for everyone in all contexts. Trust is fundamental in human life, whether to trust friends, colleagues, family members, or AI-powered services. While ML practitioners, such as researchers, and decision-makers, conventionally focus on improving the system performance of ML models using performance metrics such as the F1 score, ensuring ML application that is trustworthy stays a challenging topic. In the past few years, we have witnessed a significant amount of research work targeting trustworthy AI and ML, and the objective of this paper is to provide a comprehensive review of related research activities, with an emphasis on speech-centric ML. This survey aims to outline salient design pillars and the latest research trends in speech-centric trustworthy ML.

Trustworthiness in ML has been defined differently across the literature. For example, Huang *et al.* [85] described the term trustworthiness based

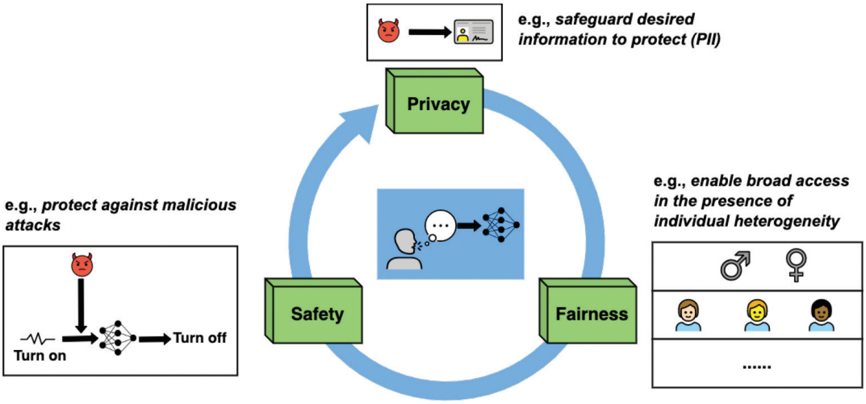
e.g., safeguard desired information to protect (PII)
Privacy
e.g., protect against malicious attacks
Turn on
Turn off
Safety
Fairness
e.g., enable broad access in the presence of individual heterogeneity
♂ ♀
......

Figure 1: Summary of key factors contribute to speech-centric trustworthy machine learning: privacy, safety, and fairness.

on the industry practices, involving the implementation of the certification process and explanation process. The certification process consists of testing and verification modules to detect potential fabrications or perturbations in the input data. The explanation refers to the ability to explain why ML reached a specific decision based on the input data. Furthermore, the ethics guidelines for trustworthy artificial intelligence published by the EU [154] recognized an AI system, to be considered trustworthy, must comply with laws and regulations, adhere to ethical principles, and function robustly. More recently, Liu *et al.* [101] summarized the trustworthy AI from safety, fairness, explainability, privacy, accountability, and environmentally friendly aspects. Likewise, our review recognizes robustness, reliability, safety, security, inclusiveness, and equity as core design elements of the trustworthiness ML system. Based on these criteria, our paper surveys the literature on speech-centric trustworthy ML from privacy, safety, and fairness perspectives, as illustrated in Figure 1:[1]

> **Privacy:** The speech-centric ML systems rely heavily on collecting speech data that is from, about, and for people in potentially sensitive environments and contexts, like homes, workplaces, hospitals, and schools. The collection of speech data frequently raises significant concerns about compromising user privacy, such as revealing sensitive information that people might want to keep private [99]. It is critical to make sure that the speech data that is either shared by an individual or collected by an ML system is protected from unjustifiable and unauthorized uses.

[1] The figure in this paper uses images from https://openmoji.org/.

Safety: Over the past few years, it has been discovered that ML systems are susceptible to adversarial attacks that aim to exploit vulnerabilities in the model's prediction function for malicious purposes [71]. For example, by introducing minor perturbations to the speech data, a malicious actor can cause the keyword spotting model to drastically misclassify the desired input speech commands. Therefore, a trustworthy ML system must generate reliable outputs for the same inputs even if the inputs have been intentionally altered by the malicious attacker [114].

Fairness: Recently, it has come to light that ML systems can perform unfairly. Why an ML system mistreats people is multifold [115]. One factor is the societal aspect, where the ML system generates biased outputs because of the societal biases in the training data or the assumptions/decisions throughout the ML development processes. Another reason that causes unfairness in AI is the imbalance of dataset characteristics, where limited data samples exist for some groups of people. As a result, the model needs to accommodate the needs of these groups to avoid biased outputs. It is also crucial to note that deploying unfair ML systems can amplify societal biases and data imbalance issues. To evaluate the trustworthiness of the speech-centric ML system, ML practitioners need to evaluate whether the ML model produces discriminatory outcomes towards individuals or groups.

The remainder of this article is organized as follows. Section 2 briefly summarizes popular speech-centric tasks, datasets, and SOTA modeling frameworks. Section 3 overviews the related survey papers in trustworthy speech-centric applications. Section 4 comprehensively discusses safety considerations in the speech-centric ML system. Section 5 discusses privacy risks and defenses in speech modeling. Section 6 reviews the emerging fairness issues with speech modeling tasks. Section 7 elaborates on potential developments and challenges in the future for speech-centric trustworthy machine learning. Finally, Section 8 concludes with a summary of the key observations in this article. Specifically, our contributions are summarized as follows:

1. To the best of our knowledge, this is the first review work to provide a comprehensive review of designing trustworthy ML focused on speech-centric modeling. We survey most, if not all, the published and pre-print works covering automatic speech recognition, speech emotion recognition, keyword spotting, and automatic speaker verification.

2. We create the taxonomies to systematically review design pillars related to the trustworthiness of speech-centric ML systems. We further compare a variety of literature on each key factor.

3. We discuss the outstanding challenges of designing speech-centric ML systems facing trustworthiness considerations related to privacy, safety, and fairness. Based on the literature reviewed, we also discuss the challenges yet to be solved and suggest several promising future directions.

2 Speech-centric Machine Learning

To begin with, we introduce the fundamentals of speech-centric ML systems to offer the audience a more in-depth understanding of the trustworthy aspect of such systems. Our aim is to provide a concise summary of speech-centric ML tasks, commonly used speech datasets, and speech modeling approaches.

2.1 Speech-centric ML Tasks

Automatic Speech Recognition (ASR): ASR system is one of the most prominent tasks in the speech domain, used to convert human speech into readable text. ASR techniques have found wide applications in modern human-computer interactive scenarios, such as Siri and Alexa. An ASR model typically involves pre-processing, feature extraction, classification, and the language model. The most widely adopted speech processing methods include framing, normalization, and pre-emphasis. After pre-processing, a feature extraction module extracts speech features as the input to the classifier. Mel-frequency cepstral coefficients (MFCCs) and Mel Spectrogram are commonly used speech features. The classification then predicts the spoken text based on the input features. Finally, a language model may be used to recognize the phoneme predicted by the classification model. It is worth noting that language models can significantly increase the efficiency of ASR systems, but it is not a necessary component in ASR systems. Many modern ASR systems can still function without using language models. Notably, word error rate (WER) is a standard metric to measure the performance of an ASR system, calculated as the number of errors in transcripts divided by the total number of spoken words. Readers interested in ASR systems may refer to Malik *et al.* [110] for a systematic introduction.

Speech Emotion Recognition (SER): The task of speech emotion recognition involves the classification of emotions (such as neutral, happy, sad, and angry) from speech signals, where the emotion labels are typically obtained from multiple human annotators. Equivalent to ASR, SER systems extract speech features passing through a classifier for emotion prediction. Due to the imbalanced label distributions in many existing SER datasets, the unweighted average recall (UAR) is often used as the evaluation metric for SER systems. For readers interested in a more comprehensive introduction to SER systems, we recommend referring to Akçay and Oğuz [5].

Automatic Speaker Verification (ASV): ASV aims to determine the authorization of the claimed identity based on voice fingerprint. ASV is composed of two stages: the enrolment and the testing stage. During the enrolment stage, embeddings or latent representations are extracted from speakers and stored. In testing, embeddings from a test speaker are extracted and compared to the claimed speaker's embeddings. Verification pairs from the same speaker are referred to as *genuine* pairs, while pairs from different speakers are "imposter" pairs [131]. Equal error rate (EER) is one of the most common metrics to evaluate the performance of ASV systems, which is the value at which the false acceptance rate (FAR) equals the false rejection rate (FRR). Further information on recent ASV developments can be found in Irum and Salman [88].

Keyword Spotting (KWS): Keyword spotting refers to the detection of predefined keywords or phrases in a speech recording, such as "Hello, Siri." This speech task is widely adopted in commercially available smart assistants such as Apple's Siri, Google Assistant, and Amazon's Alexa. Compared to ASR systems, KWS systems require substantially fewer computational resources and are well-suited for on-device learning.

2.2 Speech Datasets

The speech datasets are fundamental for training and testing speech-centric ML systems. This section provides a brief overview of the commonly used datasets for various downstream speech tasks. These datasets are also frequently used as benchmarks for evaluating speech-centric trustworthy ML. Then, we summarize the datasets along with their published years and total recording time in Table 1.

ASR: Among all the datasets listed in Table 1, Librispeech [128] is one of the most commonly used datasets for training and evaluating ASR algorithms. This dataset includes 1000 hours of audiobooks and transcriptions. Common Voice [11] is another widely used dataset supported by Mozilla. The dataset is entirely open source, and any people can contribute to the dataset by providing their speech recordings. Moreover, the CHiME-5 [15] and TED-LIUM [142] datasets consist of audio recordings from the home environment and TED talks.

SER: One of the most frequently referred datasets in the SER field is the IEMOCAP dataset [26]. This dataset consists of audio-visual data collected from 10 actors engaged in dyadic interactions. Each recorded utterance is provided with annotations into categorical emotion labels and transcripts. In addition to IEMOCAP, other commonly used datasets include the CMU-MOSEI dataset, which is based on YouTube videos [189], the MSP-Podcast

Table 1: Table showing the List of speech datasets that can be used for training the ASR model, the SER model, the AVS model, and the KWS model.

Speech Task	Dataset Name	Year	Hours
ASR	Librispeech [128]	2015	1000
	WSJ [66]	1994	162
	Common Voice [11]	2019	1900
	The CHiME-5 [15]	2018	50
	TED-LIUM [142]	2012	452
	The Spoken Wikipedia [94]	2016	1005
	Voxpopuli [178]	2021	1800
SER	IEMOCAP [26]	2008	12
	CMU-MOSEI [189]	2018	65
	MSP-Podcast [112]	2020	100
	MELD [133]	2018	-
	ESD [194]	2021	29
AVS	Voxceleb1 [123]	2017	352
	Voxceleb2 [39]	2018	2442
	VoxMovies [25]	2021	-
KWS	Speech Commands [179]	2018	-

dataset, which is based on podcasts [112], and the MELD dataset, which is based on movies [133].

AVS: The two most frequently used datasets for training the AVS system are Voxceleb1 [123] and Voxceleb2 [39]. The voxceleb1 dataset contains speech data from over 1000 celebrities on Youtube. Shortly after the success of the voxceleb1 dataset, the authors introduced the voxceleb2 dataset which includes over 6000 celebrities' speech data on Youtube.

KWS: Surprisingly, there are few purposely designed datasets for this downstream task. The Google Speech Commands dataset [179] is the most commonly used KWS dataset that includes 35 frequently used spoken words from the everyday vocabulary. The dataset includes a total of 105,829 audio recordings from 2618 speakers, wherein 2112 speakers are in the training set and the rest are in the test set.

2.3 Speech Modeling Approach

2.3.1 Conventional Modeling Approach

Conventional speech modeling systems heavily relied on the Gaussian mixture model (GMM), Hidden Markov Model (HMM), and the support vector machine

(SVM). For example, traditional speaker verification systems have used the Gaussian mixture model based universal background model (GMM-UBM) since early 2000 [139]. Later, the researchers proposed the i-vector framework [46], which reduced the high-dimensional GMM-UBM supervectors into low-dimensional vectors using factor analysis. The i-vector has been a SOTA technique in ASV systems for many years. In addition, HMM was one of the most widely used ASR modeling approaches before the deep learning approach gained popularity [169]. On the other hand, statistical descriptors of low-level speech features (e.g., pitch, intensity) were predominantly applied in emotion recognition and sentiment analysis tasks. OpenSMILE [57] is one of the widely adopted tools that enable abundant research works using this approach. However, the performance of these traditional modeling frameworks is largely impacted by the channel variations and utterance variations of the input speech signals.

2.3.2 Deep Learning Approach

In recent years, we have seen a wide variety of successes in applying deep learning to speech-centric systems. Deep speech [79] is one of the most recognized deep neural models based on the recurrent neural network (RNN). This model has been a strong baseline in the ASR domain for years. Until a few years ago, with the popularity of transformer architecture [173], researchers have proposed the self-supervised learning framework like Wav2Vec 2.0 [14]. Wav2Vec 2.0 has quickly become the SOTA machine learning model for ASR [14] and SER [34]. This model even achieves similar performance in the ASR task compared to humans. Around the same time, Google proposed a convolution-augmented transformer called Conformer [75] that reached competitive ASR performance to the Wav2Vec 2.0 model. More recently, OpenAI introduced its transformer-based ASR model called Whisper [136], which outperformed most existing works. Similar to the ASR task, many SOTA speaker verification systems are built upon the deep embedding, also known as x-vector, extracted from the deep neural network [155].

3 Related Surveys in Trustworthy Speech-centric Machine Learning

In this section, we describe several related survey papers that cover the privacy and adversarial defense in speech applications. However, to the best of our knowledge, there are no survey papers focusing on fairness risks in speech-centric applications.

3.1 Privacy

Cai *et al.* [27] provides a detailed literature review on generative models, while the author briefly describes the use of generative speech models to protect user privacy. Cai *et al.* [27] introduces 2 usage cases, remote health monitoring, and voice assistance, based on Generative speech models. In both applications, the author presents GAN-based speech models that either obfuscate the sensitive attribute or transform the user voice into a common speech signal. Cheng and Roedig [37] is another review paper that focuses on privacy in personal voice assistant applications. This survey paper discusses voice privacy preservation mechanisms including encryption schemes, voice anonymization, and distributed learning. However, this paper does not provide comprehensive reviews on privacy attacks and also does not summarize the taxonomy of privacy-preserving methods.

3.2 Safety

Apart from surveying privacy-related research topics, Cheng and Roedig [37] provides extensive reviews on security challenges in speech-centric applications. The author discusses a wide range of research works in mitigating adversarial attacks in ASR applications. This review covers popular adversarial attacks including PGD attacks, gradient decent attacks, etc. However, this review does not provide a categorization of adversarial attacks. Likewise, Huynh *et al.* [87] provides a survey on the safety aspect in speech-centric applications that categorize adversarial attacks based on types of attacking perturbations and threat models. Despite surveying different adversarial attacks in speech-centric applications, these survey papers do not discuss theories behind adversarial attacks and mitigations.

3.3 Fairness

Although many survey papers have discussed the fairness considerations in machine learning (e.g, [115]), there is a lack of comprehensive literature studying bias and fairness in speech-centric applications. Most previous work in the speech domain focus on specific applications such as ASR, ASV and SER using traditional evaluation schemes. For example, Peri *et al.* [131] highlight the need for fairness-centric metrics that better highlights the bias in existing methods and evaluation schemes to improve fairness in ASV. In this paper, we provide detailed reviews summarizing the recent works that target improving model fairness in ASR, ASV, and SER applications.

4 Safety in Speech-centric Machine Learning

Safety concerns in ML systems originated from adversarial attacks. Adversarial attacks involve a malicious actor ("adversary") who manipulates data samples with the intention of negatively affecting model performance. In a poisoning attack, the adversary manipulates data and/or model parameters during training, while in an evasion attack, they pass an adversarial sample to the model at evaluation time. In both scenarios, the adversary attempts to avoid detection by the model's users and/or administrators. Table 2 summarizes the key works on adversarial attacks.

4.1 Evasion Attacks and Defenses

4.1.1 Preliminaries

Evasion attacks can be classified both by the adversary's knowledge of the targeted model (black box or white box) and by the attack's objective (targeted or untargeted). A black-box attacker has no special knowledge of the model other than being able to observe its predictions. On the other hand, a white-box attacker possesses full visibility into model architecture, parameters, ex., and importantly, can perform backpropagation to extract loss function gradients. The white-box scenario is generally viewed as the more insidious threat model. An untargeted attack tries to produce incorrect predictions on a malicious sample without regard to the exact nature of the miss-classification. Mathematically, given benign input x with true label y, the untargeted attack optimizes for a perturbation δ that satisfies:

$$\delta = \text{argmax } \mathcal{L}(x + \delta, y) \text{ s.t. } \|\delta\| < \epsilon, \tag{1}$$

where $\mathcal{L}$ is a loss function, and $\|\cdot\|$ is some norm. The norm constraint is needed to minimize the chance of detection. In contrast, a targeted attack seeks miss-classification as a *specific* incorrect class. For benign input x with associated label y, the adversary solves

$$\delta = \text{argmin } \mathcal{L}(x + \delta, \hat{y}) \text{ s.t. } \|\delta\| < \epsilon, \tag{2}$$

where $\hat{y}$ is the targeted label for the (incorrect) class the adversary wants the model to predict. Once again the constraint $\|\delta\| < \epsilon$ ensures that the attack is relatively imperceptible to the human user.

Some of the earliest work on evasion attacks was presented by Biggio *et al.* [20]. They formulated the adversary's objective in terms of a constrained optimization problem in which the attacker searches for a perturbation that successfully "fools" the target model, while simultaneously ensuring that the attacked example remained within an ϵ-bound of the original. A gradient descent approach was proposed to generate these adversarial examples. A similar framework was suggested by Szegedy *et al.* [163], although L-BFGS was used in place of gradient descent. Additionally, Szegedy *et al.* [163] demonstrated the transferability of adversarial examples across deep learning models (i.e. attacks generated for one model can successfully fool a different one). Goodfellow *et al.* [71] link attack transferability to excessive model linearity, hypothesizing that successful perturbations are highly aligned with weight vectors, and thus tend to generalize well. They also introduce the fast gradient sign method "FGSM" for quickly constructing untargeted adversarial examples. For an FGSM attack with an ℓ_∞ norm constraint of size ϵ, the adversarial perturbation δ for sample x is given by $\delta = \epsilon \ \text{sign}(\nabla_x J(x, y; \theta))$. Here J is some loss function, y is the true label for x, and θ represents the model's parameters. Madry *et al.* [109] extended the FGSM methodology to introduce the Project Gradient Descent (PGD) attack. PGD employs an iterative approach to craft adversarial examples, with each step consisting of an FGSM perturbation followed by projection onto the ϵ-ball. PGD attacks can be used in both targeted and untargeted threat models.

4.1.2 Evasion Attacks in Speech-centric ML

A number of speech-specific attacks have also been proposed. Gong and Poellabauer [70] provided one of the first investigations of end-to-end gradient-based white box attacks on audio-modality classifiers (e.g. speaker recognition). They demonstrated that the attacks substantially degrade model accuracy while only minimally impacting human perceptual evaluation. Chen *et al.* [33] also presented an attack against speaker recognition models but used a black-box approach that relies on gradient estimation. Black-box approaches to attacking ASR systems have been introduced by [1], [8], and [93].

Meanwhile, Carlini and Wagner [29] presented a strong targeted white box attack against ASR, that obtains a perfect success rate (w.r.t. the ability of the attack to produce the targeted transcript) with greater than 30 dB mean adversarial SNR. Their approach specified the generic adversarial framework for the Connectionist Temporal Classification (CTC) loss commonly used for training ASR systems. They also quantify perturbation magnitude in decibels instead of linear units. Within the framework from Carlini and Wagner [29], attack generation equates to solving the following optimization problem

$$\min |\delta|_2^2 + \alpha \mathcal{L}_{\text{CTC}}(x + \delta, t) \text{ s.t. } \text{dB}(\delta) - \text{dB}(x) < \tau,$$

where t is the targeted transcription and $\mathcal{L}_{\text{CTC}}$ is the CTC loss function. By optimizing this objective with successively smaller values of τ, the adversary determines the smallest magnitude perturbation that achieves the targeted objective. This attack methodology was extended by Qin *et al.* [135] which introduced the "Imperceptible attack" for ASR. After finding a perturbation that fools the network, an additional update step was used to regularize the power spectrum of the attack such that it falls under the masking threshold of the clean speech. These two steps (perturbation update and spectral shaping) were repeated for a number of iterations. In this manner, the Imperceptible attacks leveraged audio-specific notions of perceptibility in constraining the attack generation process. Similar work on psychoacoustically motivated attacks has also been presented in Schönherr *et al.* [146].

4.1.3 Defenses against Evasion Attacks in Speech-centric ML

A variety of defenses have been proposed to counteract the insidious effects of evasion attacks. One method that has been shown to be broadly successful is adversarial training (AT), in which adversarial examples are generated during the training process and used to further tune the model weights [71, 109].[2] This approach is not unlike traditional data augmentation methods such as additive noise or image cropping, except that the adversarial samples are generated specifically for the model at hand (usually using a white-box attack such as FGSM). Unfortunately, despite its success, AT introduces a significant computational overhead since it requires the construction of adversarial examples on each training epoch, as well as additional weight updates.

Another widely known defense is Randomized Smoothing (RS), first introduced by Cohen *et al.* [42]. RS uses stochastic averaging to "wash-out" the effects of adversarial perturbations, which are relatively small in absolute magnitude. Specifically, given a classifier $f(\cdot)$ that is robust to additive Gaussian noise, a new "smoothed" classifier $g(\cdot)$ is produced by $g(x) = \text{argmax}_k\ P(f(x+\epsilon) = k)$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Random smoothing can provide provable robustness guarantees to ℓ_2 bounded attacks (see [42]).

Other defenses include changes to model architecture and training procedures. For example, a modified RELU activation function that constrains the maximum value of a given neuron was suggested by Zantedeschi *et al.* [190]. Similarly, Cisse *et al.* [40] introduced Parseval Regularization which was based on minimizing the Lipschitz constant. Other works, such as [98] and [126], have proposed the use of denoisers of other preprocessor methods to remove or mitigate the impact of adversarial perturbations.

Speech-specific defenses include the detection framework for ASR attacks from Hussain *et al.* [86] which leverages the sensitivity of adversarial examples to small changes. They identified adversarial samples by testing whether the predicted transcript varied substantially under alterations such as filtering and quantization. Yang *et al.* [186] used a similar approach for attack detection that tests the consistency between the transcript prediction for the full audio signal and the prediction for a partial segment. Several papers have also proposed audio resyntheses defenses, such as the GAN-based methods from [55] and [56].

[2] Note that [109] investigates both attacks (PGD) and defenses (AT).

Table 2: Summary of key works on adversarial attacks.

	Evasion Attacks	Poisoning Attacks
Attacks	Goodfellow *et al.* [71] Madry *et al.* [109] Carlini and Wagner [29]	Biggio *et al.* [21] Muñoz-Gonzãlez *et al.* [122] Liu *et al.* [103] Gu *et al.* [74]
Defenses	Goodfellow *et al.* [71] Cohen *et al.* [42] Cisse *et al.* [40]	Steinhardt *et al.* [159] Tran *et al.* [168] Gao *et al.* [65]

4.2 Poisoning Attacks and Defenses

4.2.1 Preliminaries

Whereas an evasion attack occurs at inference time, poisoning attacks involve corruption of the training data thus leading to an inaccurate or vulnerable model. Given the increasing use of publicly sourced data and federated learning, it is often impossible to guarantee the integrity of a given dataset. Barreno *et al.* [16] provided one of the earliest investigations of poisoning attacks. In particular, they distinguish between availability attacks that attempt to a broadly inaccurate model and integrity attacks that seek to produce a more specific vulnerability (e.g., misclassification in response to the presence of a specific trigger). In both scenarios, the attacker has the ability to manipulate the features and/or labels for some small fraction of the training data to decrease the performance of the learned model.

A variety of approaches have been proposed to generate availability-type poisoning attacks. In this case, the adversarial objective is to generate a set of poisoned samples that minimizes the learned model's performance on a withheld set of test samples. Biggio *et al.* [21] presented an attack on the SVM classifier that used gradient methods to generate the poisoned samples. SVM poisoning attacks have also been studied by Mei and Zhu [116] in addition to attacks against linear and logistic regression. The Karush-Kuhn-Tucker (KKT) optimal conditions to generate the poisoned samples. While the relative simplicity of these models makes direct optimization feasible, Deep Neural Networks require different approaches. For example, Muñoz-Gonzãlez *et al.* [122] introduced a method that estimates back-propagation through the entire training procedure to efficiently generated poisoned samples to maximize the validation loss in the final model. Another approach suggested by Yang *et al.* [184] was to generative poisoned samples with the Generative Adversarial Network (GAN) where the target model acted as the discriminator and a separate auto-encoder model was used as the generator that synthesized the poisoned samples.

A large number of integrity-type poisoning attacks attempt to produce a model that performs normally on clean test samples while miss-classifying those that contain a specific trigger. For example, Liu *et al.* [103] showed how a pre-trained facial recognition model can be "Trojaned" by creating a trigger-induced backdoor. They first used the model gradient to generate a trigger patch that produced a large-magnitude activation for some hidden layer neurons. Then they reconstructed faux training examples for each class by performing gradient ascent on the input image. These synthetic samples form the basis for a finetuning procedure, in which a subset of the samples have the trigger added and the label changed. The resulting model should then miss-classify future samples containing the trigger. Importantly, this is accomplished without direct knowledge of the training dataset. A similar threat model was introduced by Gu *et al.* [74], which demonstrated the efficacy of the approach by creating a poisoned model that incorrectly identified street signs when a small trigger was present.

While both Liu *et al.* [103] and Gu *et al.* [74] assumed some attack autonomy over model training, other work such as by Chen *et al.* [36], assumed that the attacker only has access to a small percentage of the training data. The authors proposed several approaches to generate backdoor samples, including an input-instance-key method in which the poisoned samples are randomly perturbed versions of a "key" input with the target label. The adversary hoped that the final model would thus misclassify test samples that are similar to the key input used to poison the training data. Another poisoning algorithm suggested by Liu *et al.* [103] was to create poisoned instances by adding a specific pattern to benign samples, and changing the label to the targeted class. A clean-label attack that manipulated only the input features was presented by Shafahi *et al.* [147]. Their approach optimized for an adversarial sample that is perceptually similar to the target class in input space but close to a target instance in feature space. If the target instance is then deployed on the trained model, it is more likely to be misclassified as the target class. Another clean label attack was the Witches Brew method [68], which used gradient matching to produce samples that modified the model's training trajectory such that it would misclassify a specific target instance at test time.

4.2.2 Poisoning Attacks in Speech-centric ML

So far, there has been relatively little work done on audio-specific poisoning attacks. The first targeted attack against hybrid ASR systems is the VenoMave algorithm [3]. For each x_i frame selected by the adversary for misclassification, that attacker constructs poisoned frames that "surround" x_i in feature space and assigns them the target label. Thus any linear model that correctly classifies the poisoned frames will also misclassify the target frame. Ge *et al.* [67] proposed a clean-label attack to protect user's speech data from use in downstream learning tasks such as speaker recognition or speech command recognition. This was accomplished by introducing perturbations that maximize the distance between the MFCC features of the clean and poisoned signal while simultaneously minimizing the perturbation's magnitude. Thus any model trained on the poisoned data that uses MFCCs is likely to perform unreliably.

4.2.3 Defenses against Poisoning Attacks in Speech-centric ML

Many of the defenses proposed for poisoning attacks attempt to identify and remove the poisoned samples from the training data. For example, Tran *et al.* [168] extracted learned representation for all training samples with a given label. They then compute a singular value decomposition (SVD) and use the right singular vectors to compute outlier scores for each sample. Samples with a high outlier score are removed, and the model is retrained. Model activations were also used to detect poisoned samples by Chen *et al.* [32], but the authors employ a clustering approach in place of the SVD. They also suggest fine-tuning the model on the re-labeled poisoned data rather than training from scratch on the filtered dataset. Gao *et al.* [65] leveraged the fact the sample contains a trigger that should be consistently classified as that target class even under perturbation. By randomly combining different images, the check for samples that exhibit low entropy in the distribution of predicted labels. Some defensive method focus on addressing backdoors in the model itself. For example, Wang *et al.* [177] tested for backdoors by computing the minimal perturbations required to change the classification result of all samples to a given target label. If a perturbation of a relatively small magnitude exists, then this may indicate the presence of a backdoor. The "Fine-pruning" defense presented by Liu *et al.* [102] first pruned neurons that were inactive on clean samples under the hypothesis that these extraneous neurons can be co-opted by the backdoor trigger. Then the network was fine-tuned on a clean (un-poisoned) dataset to avoid overall performance degradation. Steinhardt *et al.* [159] provided a method to upper bound the effects of a potential poisoning attack in the case where a data sanitation defense was used to remove outliers before training.

5 Privacy in Speech-centric Machine Learning

Despite the promises modern speech applications can deliver, they also raise significant concerns and risks, such as exposing sensitive information that people might wish to keep confidential. The sensitive information can be individual attributes (e.g., age, gender), states (e.g., health, emotions), or biometric fingerprints. This section presents a comprehensive review of the privacy and security challenges related to trustworthy speech processing.

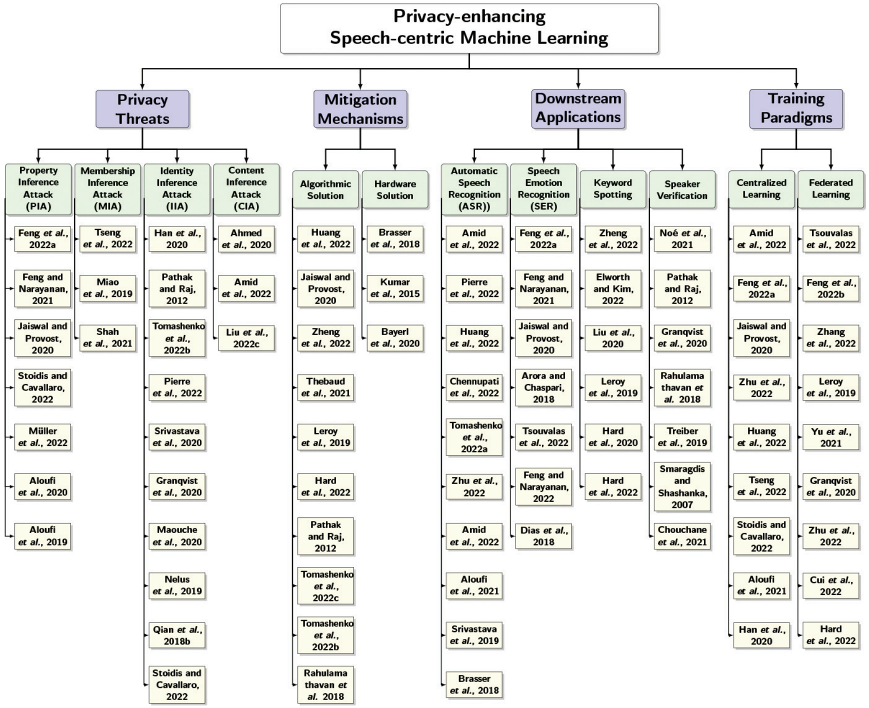

Figure 2: Taxonomy of privacy-related speech processing and modeling works.

5.1 Taxonomies of Privacy-related Speech-centric ML

In this subsection, we create a taxonomy of existing privacy-related topics on speech-centric ML in Figure 2. In the categorization shown in Figure 2, we organize papers based on privacy threats, mitigation mechanisms, downstream applications, and training paradigms. In the first place, we categorize privacy threats based on privacy attacks, including Property Inference Attacks (PIA), Membership Inference Attacks (MIA), Identity Inference Attacks (IIA), and Content Inference Attacks (CIA). For example, in PIA, the privacy attacker can obtain or infer speaker attributes like demographic information. On the other hand, we can divide privacy-related literature based on privacy mitigation methods. Specifically, the literature surrounding privacy-preserving speech processing can be categorized into algorithmic solutions and hardware solutions. Moreover, the most studied downstream speech applications related to privacy topics can be categorized into automatic speaker verification, keyword spotting, automatic speech recognition, and speech emotion recognition. Lastly, we discuss the privacy-related speech-centric ML based on the training paradigm: centralized learning and federated learning (FL).

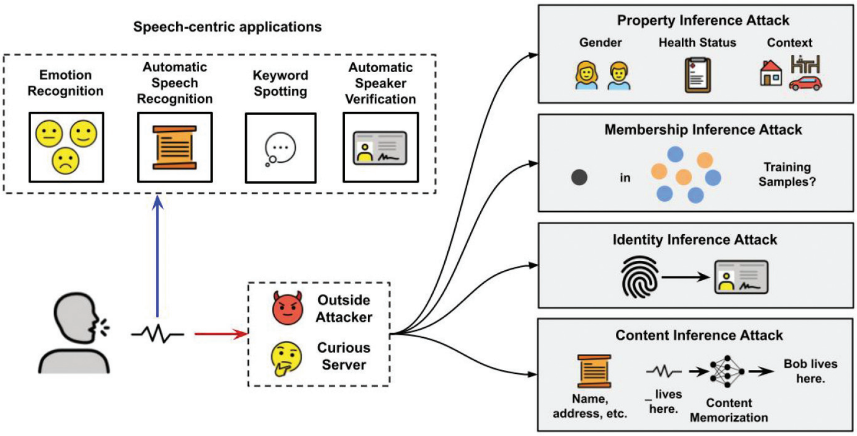

Figure 3: Overview of the privacy threats in speech-centric models.

5.2 Privacy Threats

In a privacy attack, the goal of an adversary is to acquire information that was not intended to be disclosed, including speech content, speaker demographics, and voice fingerprint. There are four mainstream privacy attacks against speech-centric applications as demonstrated in Figure 3: Property Inference Attacks (PIA), Membership Inference Attacks (MIA), Identity Inference Attacks (IIA), and Content Inference Attacks (CIA). We provide a brief overview of each privacy attack and highlight works that either identified new privacy risks or were the first to investigate privacy threats in their respective domains.

Property Inference Attacks (PIA): PIA occurs when the adversary attempts to infer private attributes which are unrelated to the primary learning task. A notable example of the PIA in speech-centric applications is to infer the gender attribute using a pre-trained gender classification model, while the target application is to classify emotions or transcribe text. Here, the speech data that is accessible to the attacker can either be the raw speech recordings or processed speech features like MFCCs. In addition to gender property, adversaries can perform classification to predict the age [145], the language used [107], or even the health status [6] of the speaker from the speech data. However, since most existing speech-related datasets only include the annotations of gender but not other properties, the majority of the PIA works in speech applications focus on gender classification. Furthermore, apart from using features derived from raw speech data, the adversaries could per-

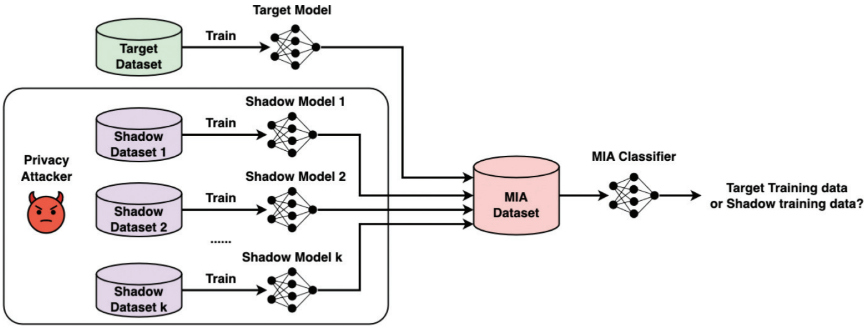

Figure 4: Overview of the Membership inference attack. The privacy attacker first trains a set of shadow models with shadow training datasets. Once the shadow training is finished, the attacker gathers the model outputs from the target model and the shadow model to train the MIA classifier. The MIA classifier infers the membership property given an input posterior from the target model.

form privacy attacks through training updates generated in the collaborative training process presented by Feng *et al.* [63].

Membership Inference Attacks (MIA): Membership inference aims to determine the participation of a data instance in training the target model. The idea of the MIA was first proposed by Shokri *et al.* [151], where the author assumed the attacker could access posterior estimations of a data sample by querying the target model. The attacker then used the posterior distribution to speculate whether the query data was in the training data. The general framework of MIAs is presented in Figure 4. In MIAs, the attacker typically starts with a training procedure called shadow training which emulates the target training procedure. To perform the shadow training, the attacker often gathers a collection of shadow training datasets that share similar data distribution or data format to the target training data. The attacker then trains a classifier to perform MIAs using a collection of posteriors from training and shadow data. Although MIAs have been widely studied in computer vision and natural language processing, there is a limited amount of work in speech modeling. In speech-centric ML, Shah *et al.* [148] were the first to investigate MIAs on ASR models and their results show that the attacker can infer membership of the speech data with a moderate precision score. Furthermore, it is important to point out that the success rate of MIAs depends largely on the speakers, while some speakers are more vulnerable to MIAs. However, the connection between the attack success rate of MIAs and the speaker remains to be determined. Recently, Tseng *et al.* [171] designed the MIA against pre-trained speech models trained using self-supervised learning

(SSL) in black-box settings. Their results indicate that both utterance and speaker-level MIAs are feasible against SSL-based speech models.

Identity Inference Attacks (IIA): The IIA against speech-centric applications is a class of privacy attacks where adversaries can extract personally identifiable information (PII) from speech data for re-identification or impersonation purposes. However, we highlight that such identifiers often have applications in speaker identification and automatic speaker verification tasks. Traditional speaker identification frameworks rely on the extraction of i-vector from MFCCs [46]. i-vector is a low-dimensional fixed-length representation of a speech utterance extracted using a data-driven approach. The state-of-art speaker identification systems in more recent years involve the extraction of the x-vector [155], which is the embedding extracted from Deep Neural Networks. Many papers [78, 111, 125, 130, 132, 158, 164] have proposed various privacy mitigation methods to prevent speaker re-identification. The vast majority of these works thus far applied adversarial training to disentangle the unique speaker information embedded in the speech recordings.

Content Inference Attacks (CIA): In addition to speaker properties, speaker memberships, and biometric identifiers, the textual contents of the speech can expose significant privacy concerns. For example, speech recordings often contain the name, address, and contact information of the speaker or the people around the speaker. Moreover, speech recorded in sensitive settings like business meetings can include proprietary information. Therefore, the content inference is to extract textual content from speech recordings. The naive approach to performing the content inference is through ASR itself. More recently, researchers have discovered that untended memorization in training deep speech models can also leak training data records [9, 104]. Specifically, the attacker can deliberately synthesize speech contents by guessing content patterns in the training data. For example, the crafted speech utterance can be "_ lives in the 1st street." where _ can be the silence audio snippet. The idea of the attack is that the ASR model would output the name along with the phrase "lives in the 1st street.", like "Bob lives in the 1st street.", as a consequence of model memorization.

5.3 Mitigation Mechanisms

In this paragraph, we continue our review of privacy-related literature on speech-centric ML based on mitigation strategies. We split the related papers into algorithmic and hardware solutions based on such criteria.

5.3.1 Algorithmic Solutions

There is a large body of work falling into algorithmic solutions. More concretely, differential privacy (DP), adversarial training, and encryption are frequently adopted mitigation methods in speech-centric applications.

Differential Privacy (DP): The idea of DP was first introduced by Dwork [51]. Essentially, the DP is a rigorous privacy definition that guarantees the exclusion or the inclusion of any particular data record in the dataset has a negligible impact on the original data distribution. In other words, the privacy attacker cannot distinguish the data distribution changes by including or excluding any particular data point in the original dataset. Mathematically speaking, we can define DP given privacy parameters ϵ and δ shown below:

Definition 1 ((ϵ, δ)-DP)**.** *A random mechanism $\mathcal{M}$ satisfies (ϵ, δ)-LDP, where $\epsilon > 0$ and $\delta \in [0, 1)$, if and only if for any two adjacent data sets $\mathcal{D}$ and $\mathcal{D}'$ in universe $\mathcal{X}$, we have:*

$$Pr(\mathcal{M}(\mathcal{D})) \leq e^{\epsilon} Pr(\mathcal{M}(\mathcal{D}')) + \delta \tag{3}$$

Here, $\epsilon > 0$ is defined as the privacy budget in DP, and a lower ϵ represents stronger privacy protection [53]. Specifically, ϵ provides the bound of all outputs on neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$, which differ by one sample in a database. The typical way to achieve differential privacy is through noise perturbation. When considering centralized speech applications, Qian *et al.* [134] proposed VoiceMask that conceals the voiceprints of the speaker using differential privacy. Unlike the DP implementation in VoiceMask, Preech [4] ensured DP by adding dummy words in the output transcripts. Besides the literature on centralized training, we have observed an increasing trend of DP research in Federated Learning. For example, Feng *et al.* [63] applied the DP to mitigate property inference attackers in training the FL-based speech emotion recognition model. Sotto Voce [150] was another recently proposed work that explored DP in Federated speech recognition.

Adversarial Training: As Mireshghallah *et al.* [120] pointed out, adversarial training is based upon information-theoretic privacy. The fundamental theory behind adversarial training is mutual information. For example, given a sensitive property or unique identifier z and the associated speech data $\mathbf{x}$, we want to learn the perturbation $h(\cdot)$ that maximizes the mutual information $I(h(\mathbf{x}); z)$. This learning objective is typically turned into the following adversarial training objectives as suggested by Song *et al.* [156]:

$$\min_{\boldsymbol{\psi}} \max_{\boldsymbol{\phi}} \; \mathcal{L}(adv_{\boldsymbol{\psi}}(h_{\phi}(\mathbf{x})), z). \tag{4}$$

Essentially, we would like to train an adversary that is able to infer z accurately, and meanwhile, we aim to improve the quality of the perturbation that confuses the adversary classifier. This training objective is normally combined with the target training objective in the learning phase. As we described in PIAs, many papers used adversarial training to disentangle the gender attribute from the speech signal. Jaiswal and Provost [89] were the first ones to propose to use of adversarial training to remove gender property in the SER task. Following Jaiswal and Provost [89], Feng *et al.* [58] combined adversarial training with feature selection to greatly reduce the gender inference risks in SER. Another common approach to conducting adversarial training is through generating adversarial examples. For instance, in ASR training, Stoidis and Cavallaro [160] presented a generative adversarial network that fed gender-ambiguous training samples to train the ASR model. This design attempted to disentangle gender from training utterances. Aside from unlearning demographics from speech signals, Chennupati *et al.* [38] tried to unlearn PII such as x-vector by synthesizing speech utterances from a large pool of x-vector.

Encryption: The last popular privacy-defending mechanism in this category is encryption. Many papers exploit Homomorphic Encryption (HE) for secure training of speech-centric models. Elworth and Kim [54] and Zheng *et al.* [193] investigated the use of HE in keyword spotting systems, and Dias *et al.* [49] were the first to investigate HE in SER applications. On top of HE integrations, Pathak and Raj [130] adapted secure multiparty computation (SMC) protocols that substantially reduce the computation overhead needed to satisfy the privacy constraints. However, in our literature search, we cannot find related papers investigating HE in the ASR system. The lack of ASR research in this direction can be caused by the heavy computation required in encryption computation. Meanwhile, the modern ASR system demands a significant amount of computing resources.

5.3.2 Hardware Solutions

Compared with algorithmic mitigation in speech-centric applications, fewer research papers work on hardware solutions. Based on the sampled literature, we divide the hardware solution into sensing strategies and trust computing environments. In the context of audio sensing, Kumar *et al.* [95] aimed to improve the privacy of audio data recorded from wearables by audio subsample and audio shredding on the device. Specifically, audio shredding was to randomize the sequence of recorded audio features. These two sensing strategies promised to provide useful audio features that were secure from reconstruction attacks. Additionally, Feng *et al.* [60] introduced a wearable audio solution that enhances privacy through sampling low-level acoustic characteristics instead of raw audio samples to study workplace stress (Mundnich *et al.* [121] and Yau *et al.* [187]). On the other hand, there has been a growing interest in recent years in using a trusted computing environment (TEE) in speech-centric applications. For example, VoiceGuard [23] was one of the first works demonstrating the use of Intel SGX, a widely available TEE implementation, on the ASR application. Last but not least, Bayerl *et al.* [17] built a TEE architecture called Offline Model Guard (OMG) that allowed running KWS tasks on the pre-dominant mobile computing platform ARM.

5.4 Downstream Speech Applications

Here, we review the privacy-related speech literature based on the downstream applications. In this review, we select popular speech applications, including automatic speech recognition (ASR), speech emotion recognition (SER), automatic speaker verification (ASV), and keyword spotting (KWS).

ASR: As shown in Figure 2, ASR is the most studied speech application in the context of privacy. These works implement privacy-enhancing features by removing gender property [7], biometric identifiers [132], and sensitive content [9] from speech signals. Specifically, the scientific community has also introduced VoicePrivacy 2020 [166] and VoicePrivacy 2022 challenges [165], with the target to evaluate privacy-preserving ASR modeling frameworks that suppress biometric identifiers in the speech signal. Currently, most of the privacy-enhancing ASR systems are proposed in the centralized setting, and Federated ASR training remains a challenge in speech-centric ML research. In this survey, we find only a few presented works [188, 195] that focus on ASR modeling using federated learning. As Yu *et al.* [188] concludes, ASR models suffer significant utility loss using Federated learning due to the nature of the complexity and high variability residing in speech data. The lack of ASR modeling works in the FL domain can also be caused by the expensive computation requirements of the ASR models. Unlike the algorithmic solutions to reduce privacy risks, Brasser *et al.* [23] introduced the VoiceGuard architecture that protects user privacy using inside a trusted execution environment (TEE). Last, we also want to highlight that the Librispeech [128] dataset is commonly used in privacy-related ASR works.

SER: In our review, we identify that many papers [58, 61, 89] in this domain focus on disentangling the gender attribute from the speech signal using adversarial training. Among all these papers, Jaiswal and Provost [89] was the only work that considers multi-modal learning with other modalities. As opposed to gender obfuscation works mentioned above, Arora and Chaspari [12] was the first to explore speaker anonymization using siamese neural network architecture. Apart from centralized speech emotion recognition, many recent works [62, 63, 172] explored privacy risks in federated learning

settings. Specifically, the IEMOCAP dataset [26] is one of the most used datasets in conducting these experiments.
KWS: The literature of privacy-preserving keyword spotting systems [54, 193] are mainly based on homomorphic encryption (HE) solutions. However, as we discussed earlier, this method has major constraints in computation efficiency and is extremely challenging to deploy in the field. Alternatively, DataMix [106] improves privacy by generating data samples through the mixup approach [191]. The mixup data is a mixture of data samples that can effectively prevent privacy attacks like IIA while preserving the target model utility.

ASV: Interestingly, most privacy-centered speech modeling papers treat speaker identity as sensitive information. Most of these papers heavily studied the obfuscation of the speaker identity, or in other words, reduce the performance of the speaker verification system, where the target application is often ASR, SER, or other speech-related applications. Since the i-vector or the x-vector already carries the biometric identifier of the speaker, many works that attempt to preserve privacy in the speaker verification systems focus on redesigning the system using homomorphic encryption solutions [130, 153]. However, the homomorphic encryption frameworks require heavy computations and are frequently impractical for real-life deployment. In contrast to these holomorphic proposals, Rahulamathavan *et al.* [137] designed a randomization algorithm that significantly reduced the computation overhead for privacy-enhancing speaker verification using the i-vector. With the popularity of Federated Learning in more recent years, Granqvist *et al.* [73] investigated the on-device learning schemes for local speaker verification.

5.5 Training Paradigms

5.5.1 Centralized

Centralized learning requires collecting the raw speech data. In this setting, the speech signal is normally sampled at the client device and is then transferred to the service provider's server for post-processing. The collection of speech data often draws substantial privacy concerns as speech signals encapsulate demographics, health information, PII, or sensitive speech content. If service providers are untrusted, they may not only infer the mentioned private information from speech data but also even render a person identifiable information.

To decrease the privacy risk of gathering speech signals, many privacy regulations have been issued in recent years, such as *European Union's General Data Protection Regulation (GDPR)* law [175]. Noticeably, *GDPR* does not explicitly consider the machine learning models as personal data, but recent works imply that ML models themselves could be covered by *GDPR* as models may memorize sensitive information during the training process. With regard to the research community, a large amount of effort has been made to decrease

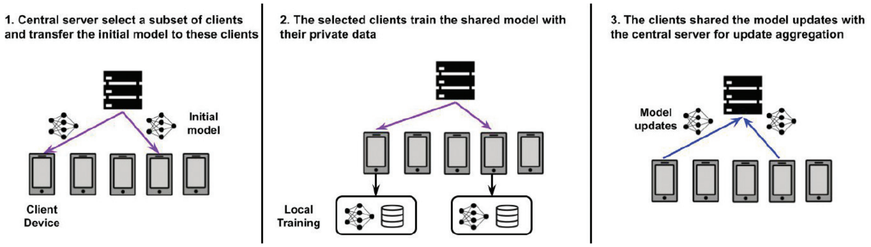
1. Central server select a subset of clients and transfer the initial model to these clients
Initial model
Client Device
2. The selected clients train the shared model with their private data
Local Training
3. The clients shared the model updates with the central server for update aggregation
Model updates

Figure 5: The training process of Federated Learning.

the privacy risks in centralized speech systems like ASR, SER, and ASV. Many of these papers, as summarized earlier, focus on perturbing the speech data or intermediate speech features to disentangle private information. Widely used speech data perturbation methods are differential privacy, adversarial training, or generative methods. We would also stress that most current works emphasize hiding/removing PII, but fewer works studied the topic of preventing the rendering of PII.

5.5.2 Federated Learning

Section 5.5.1 has introduced the plethora of privacy risks centralized computing poses among which a significant attack concept is PII being transferred to a centralized server, wherein it is prone to cybersecurity attacks or curious server attacks. As an alternative to the traditional methods of training machine learning on a single server, federated learning (FL) employs a server-client model such that the data never leaves the client. Instead, the objective of the server is to aggregate all the model updates from the clients. Unlike centralized training, the clients train the model on a local dataset and transmit the updated parameters instead of the raw data. The general learning procedure of FL was shown in Figure 5.

FL Optimization Techniques: The earliest optimization algorithm proposed to ensure convergence of the global model was done by McMahan *et al.* [113], known as Federated Averaging (*FedAvg*). *FedAvg* is similar to SGD, wherein the client model performs multiple iterations of updates before communicating the updates to the server. Although *FedAvg* has been shown to have great success, it has convergence issues in heterogeneous data settings, i.e., when the data distribution is non-IID and in cases wherein a limited number of clients participate in every global update. One method proposed to improve optimization for heterogeneous data is the Stochastic Controlled Averaging Algorithm [92] (SCAFFOLD) by using control variates which reduce the client drift away from the global optima. More recently, Asad *et al.* [13] proposed three adaptive optimization algorithms, FedAdaGard, FedAdam, and FedYOGI, which are the FL versions of AdaGard, Adam, and YOGI, respectively. From empirical estimates, FedOpt outperforms most federated optimization strategies. Federated learning for speech processing tasks has been explored for numerous applications, mainly in keyword spotting, automatic speech recognition, and speech emotion recognition. Specifically, Zhang *et al.* [192] provided a comprehensive benchmark for various audio-related tasks.

KWS: With the ubiquitous usage of smart assistants such as Siri, Google Voice, and Alexa, keyword spotting is an essential downstream speech processing task. Furthermore, it requires a relatively low parameter count ($\sim$ 200k) to achieve state-of-the-art performance, making it ideal for federated learning. Hard *et al.* [81] and Leroy *et al.* [97] proposed the FL approach for keyword spotting tasks. Both papers provided models with comparable performance to a centrally trained model on their respective datasets. Hard *et al.* [81] observed that the performance of the FL model depends on the strategies employed to deal with the non-IID data. For example, data augmentation through SpecAug [129] or the usage of adaptive optimization methods such as ADAM in the local clients have been observed to be effective in preserving performance. In more recent work, Hard *et al.* [80] demonstrated a real-time on-device training of an FL model for Keyword Spotting. Also, it introduces semi-supervised learning in an FL scenario and self-correcting labels based on metadata during the recordings.

ASR: Dimitriadis *et al.* [50] was one of the earliest works to propose FL for Speech Recognition tasks. In order to train with heterogeneous data, a two-level hierarchical optimization strategy was proposed, which involved a local client optimization followed by a global optimization and retraining of the global model on the client side held out dataset. This helps the training process better adjust to client drift. In addition, a weight model averaging is proposed, which helps improve the convergence speed. In order to account for data heterogeneity, the addition of variational noise has been proposed by Guliani *et al.* [76] wherein each client model is added with a local random variational vector. The author named this method as federated variational noise (FVN). FVN has been shown to improve the relative performance of the federated learning models in a non-IID data scenario.

With the constraints on the computational costs and the model sizes, there has been a focus on employing a cross-silo FL framework. Since the number of clients is smaller and each client has much larger computing power than a cross-device setting, usage of large-scale ASR models is justifiable. Cui *et al.* [43] proposed a cross-silo FL framework that contained a detailed analysis of training an ASR system with multi-domain data, i.e., each client has a specific domain exclusive data such as read speech, conversational data, meeting data etc. It introduces the client-adaptive federated learning (CAFT) method which accounts for the differing domain modalities across clients and adapts the client's data using a transform. Nandury *et al.* [124] experimented a modification to FedAvg termed as FedAvg-DS wherein DS stands for Diversity scaling. This modification emphasizes accounting for the variability in gradient directions of the local client updates.

Recent efforts have enabled cross-device FL-based ASR training by [77], which employed federated dropout [28] in order to reduce the model sizes in the clients and at the same time obtain a fully trained ASR at the server side. Additionally, it has been shown that training with federated dropout allows sub-models of the fully trained model to have comparable performance allowing for deployment on devices with varying computing capacities. Yang *et al.* [185] introduced partial variable training (PVT) which involves freezing layers in clients and training a specific set of layers per client and aggregating them per layer for the server updates.

SER: In order to deploy these models for real-time usage, we have to note that the availability of labeled data to the clients is extremely low, if not non-existent. Hence, semi-supervised FL frameworks are increasingly popular to train SER models with limited labeled data points per training client and a larger unlabelled set of data. The unlabelled sets of data are used in the supervised training by predicting pseudo labels. Tsouvalas *et al.* [172] generated the pseudo labels for the unlabelled models and retains them based on the confidence measure of the label. Whereas Feng and Narayanan [62] used multiview pseudo-labeling [181] coupled with uncertainty-aware pseudo-labeling selection process [140] to generate the pseudo labels.

5.5.3 Challenges in FL

Despite considerable improvements in privacy owing to the transition from a centralized to a decentralized training approach in FL, significant challenges remain for the ubiquitous adoption of FL models in speech processing, as shown in Figure 6. Although some speech processing tasks can attain the state of the art performance within parameter counts of 300k, tasks such as speech recognition and speech generation require about 100M parameters to obtain performance that is equivalent to state of art. However, clients in an FL are constrained by low computational power and model sizes. This provides an opportunity for future research into better optimization techniques or smaller models which can enable these tasks to be trained in a federated setting.

Another issue that arises while training models at clients in a supervised manner is the lack of clean labels. Therefore, semi-supervised and unsupervised techniques are of interest since they do not have to deal with the lack of clean labeled data. Some methods to tackle this issue in current works include employing a student-teacher framework providing weak labels on the local datasets and using external metadata to infer the labels based on user actions.

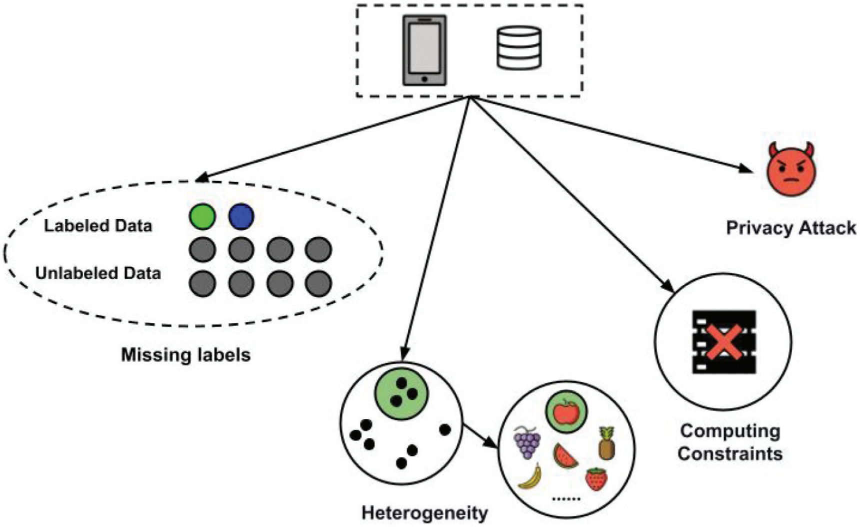

Figure 6: Challenges in Federated speech-centric applications.

Moreover, despite user data not being transferred to the server, it has been shown that the gradients which are transferred are susceptible to a multitude of privacy attacks, such as data reconstruction attacks, property inference attacks, membership inference attacks, and poisoning attacks which adversely affect the privacy-preserving nature of FL. Although there have been numerous works on privacy attacks and defenses in federated learning, most of them focus on image or text tasks. Privacy attacks for speech processing tasks are a relatively under-explored area so are the defenses. However, it is an increasing field of interest owing to the large-scale adoption of voice assistants and their privacy concerns.

5.5.4 Privacy Attacks in FL

Tomashenko *et al.* [164] proposed two attacks on ASR models in order to infer the speaker identity. One is purely statistical, and the other is a neural network-based attack to infer speaker identity based on the outputs of hidden layers from speaker models transmitted from the clients to the server. It has been demonstrated that this attack is successful with EER values obtained about 1–2%. Similar property inference attacks were performed on SER models by Feng *et al.* [59], wherein the gender of the client was inferred from the model updates transmitted to the central server, which could either be the gradients or weight parameters. The attack is performed by shadow training [151] on an open dataset similar to the target dataset, followed by using the model updates obtained during the shadow training to infer the gender of the speaker. A point to note is that in both the previous attacks, it is observed that the first hidden layer of the model provides the most information about the speaker identity.

5.5.5 Defences for FL

Defenses for FL models mainly include the use of DP [51], or Homomorphic Encryption [127]. In the speech processing domain, Feng *et al.* [63] deployed User-Level Differential Privacy (UDP) to mitigate property inference attacks from SER models. However, one limitation of the proposed defense is that when the adversary obtains multiple updates of the models, the performance of the defense degrades. Chang *et al.* [31] proposed a two-layered defense mechanism i.e., a combination of randomization and adversarial training in a federated setting to defend against FGSM, PGD, and Deepfool attacks for SER tasks.

6 Bias and Fairness in Speech-centric Machine Learning

6.1 Fairness in Machine Learning

The advance of machine learning technology has led to its ubiquitous application spanning several domains including healthcare, travel, and also the judicial system. While ML can alleviate human effort and promote automation, it has to be used cautiously to avoid potential biases to infiltrate the automatic decision-making process [69]. For example, a popular study investigated Recidivism Prediction Instruments (RPI) – ML technology used to predict if a person who had committed a criminal offense in the past is likely to commit an offense in the future, and found that the popular COMPAS RPI was biased against black defendants [10].

Similarly, evidence of bias has been found in face recognition [141, 157], natural language applications [108] as well as voice assistants [48]. As explained below, such bias can originate from either the training data, features/model or incorrect application of algorithm. Hence, mitigation methods have been proposed at each of these stages of the pipeline, from data curation to model deployment. In this section, we briefly delineate individual and group fairness, introduce different causes of bias, and finally delve into mitigation methods proposed in the speech domain.

6.1.1 Notions of Fairness

The fairness of machine learning algorithms is measured along one of two dimensions:

(a) *Individual fairness:* tracks fairness/bias at the level of an individual member of a population, with the assumption being that similar individuals will be treated similarly [52, 96].

(b) *Group fairness:* measures relative bias between different subgroup populations of interest [19, 82]. These sub-groups are typically divided along the lines of sensitive attributes of individuals such as gender, race, or age.

6.1.2 Causes of Bias

Unfair machine learning applications can be largely attributed to one of two main causes:

(a) *Data Bias*: Sources of data bias can be broadly categorized into measurement bias, omitted variable bias, representation bias, sampling bias and aggregation bias [115]. These biases creep in either due to (a) incorrect sampling of data from different subgroups, (b) biased feature representations used for modeling, or (c) misinterpretation of population statistics for subgroup statistics. A common mechanism used to mitigate data bias due to imbalance in subgroups is to over-sample data from minority subgroups during training [48, 64].

(b) *Algorithmic Bias*: Algorithmic bias arises when an ML algorithm introduces bias in the system or amplifies existing bias in the data. Such bias permeates in the absence of carefully designed ML algorithms. Algorithmic bias can be classified based on training data bias, bias in designing an algorithm or bias in deploying an algorithm [44]. Training data bias can originate from an algorithm that propagates existing bias in the data [157]. Bias in the design of an algorithm can be either a focus bias, wherein the algorithm uses features that are biased towards specific sub-groups, or processing bias, wherein the algorithm itself introduces bias as in the case of a statistically biased estimator. Finally, bias can occur in the deployment or interpretation of an algorithm. For example, using an algorithm outside the context it is developed for can lead to unfair results. Similarly, using incorrect performance metrics that do not reflect the distribution of the data can lead to misinterpretation of the results. Methods used to mitigate algorithmic bias include adversarial training and joint multi-task training [131, 162, 170].

Different formulations have been proposed based on the fairness objective being targeted [174]. For example, some fairness metrics such as statistical parity and equal acceptance rate only consider the predicted outcome of a model. More commonly used metrics, including equalized odds, predictive equality, predictive parity and equal opportunity further consider the true label in their fairness definition.

In speech processing applications, however, most work still use standard performance metrics to obtain fairness metrics (e.g., word error rate (WER) for ASR, equal error rate (EER) for ASV). As shown by Peri *et al.* [131], these

Table 3: Speech datasets for fairness evaluation.

Dataset	Task	Sensitive attributes (# classes)	No. Hours
Fairvoice [64]	ASV	Gender (2), Age (9), Language (6)	1700
Casual Conversations [100]	ASR	Gender (2), Skin Type (6)	572
TedTalk [2]	ASR	Gender (3), Race (4)	564
Artie Bias Corpus [118]	ASR	Gender (3), Age (8), Accent (3)	2.4
Speech Accents [180]	ASR	Accents (7), Gender (2), Age	<1

metrics do not always hold, especially when applied to subgroups of population. In the following subsection, we outline different fairness related work including mitigation strategies in speech processing applications and present the different data resources that have been curated for fairness research.

6.2 Fairness in Speech-centric Applications

Compared with the flourishing of fairness research in other domains, such as facial analysis and natural language processing, the importance of addressing bias issues in speech processing has been underestimated for a long time. Up to today, there is still a limited amount of studies focusing on speech-related fairness. However, the accuracy degradation of a biased speech model on a specific demographic group not only leads to users' inconvenience but also makes the group believe the product is not designed for them. Therefore, it is essential to systematically evaluate the speech-related models' performance on fairness and investigate effective methods to mitigate the bias in speech models. Several datasets have been proposed to advance fairness research in the speech domain (see Table 3). In addition to the task-specific labels (speaker/transcript for ASV/ASR), these datasets include labels for sensitive demographic attributes such as gender, race, age, accent, and language.

6.2.1 Fairness in Automatic Speaker Verification

ASV systems suffer from bias in different demographic attributes, such as gender, age, language, and ethnic. For example, Stoll [161] analyzed the performance difference of statistical speaker models regarding gender and age. However, the analysis is based on the verification scores while lacking deeper insights into the bias of the decision model and the distribution of speaker embeddings. As discussed earlier, these two are essential factors in clearly uncovering bias in ML models. More recently, Peri *et al.* [131] investigated bias in ASV, showing the importance of distribution scores and also proposed fairness metrics for ASV instead of standard EER. They also explored mitigation strategies of adversarial training and multi-task training to reduce gender bias in ASV systems.

In addition to gender and age, ASV systems have also been found to be biased across languages. A study of UN-meetings [84] investigated the effect of language in speaker verification results and found that ASV degraded for Russian language as compared to English. Jin *et al.* [91] forced the ASV learner to focus on poorly performing instances by weighting samples with an adversarial reweighting network and demonstrated that the reweighting method significantly improves the performance of ASV across different subgroups of gender and nationality.

NIST introduced new languages in their Speaker Recognition Evaluation protocol in 2016 [144] and 2018 [143], to investigate the influence of language in the speaker verification system. In the 2021 VoxCeleb Speaker Recognition Challenge (VoxSRC) [24], the language attribute was added to the speaker verification track, aiming to encourage researchers to solve the performance degradation issue in the multi-lingual setting and boost the fairness in speaker verification. It is worth noting that the majority of the Fairness studies on the ASV task provide the evaluations using the VoxCeleb dataset series [39, 123]. Chen *et al.* [35] explored the bias in speaker identification systems across different race groups and found that latinxs performed significantly worse than Caucasian speakers. Recently, Fenu *et al.* [64] explored the fairness in deep learning-based ASV systems by collecting a speaker dataset – Fairvoice, conducting performance analysis on EER and score (FAR, FRR) distributions, and providing more understanding of how diverse speaker verification is correlated with demographic attributes. To standardize the ASV fairness evaluation, Toussaint and Ding [167] devised a framework to provide comprehensive fairness evaluation metrics and visualization methods to present model's fairness across subgroups.

6.2.2 Fairness in Automatic Speech Recognition

In the past decade, due to the development of deep learning and the availability of large-scale speech and language databases, the word error rate (WER) of ASR models has decreased to a satisfactory level in many languages. However, the fairness of ASR systems still raises the interests of researchers from psychological, sociology, and engineering backgrounds [117, 138]. Mengesha *et al.* [117] investigated the ASR failure on African American Vernacular English and demonstrated the detrimental impact on African American users from the psychological perspective. In addition to proposing a set of methodologies to model the users' feelings and experience in fairness research, Mengesha *et al.* also encourages researchers to spot more light on fairness AI. Rajan *et al.* [138] introduced an automated testing framework (AEQUEVOX) for evaluating the fairness of ASR systems. By conducting extensive fairness experiments on four datasets with three commercial ASRs, Rajan *et al.* [138] validated the ASR fairness violation on non-native English, female, and Nigerian English speakers. With the recent advent of self-supervised learning, approaches such as wav2vec2 [14] are being widely used in a variety of speech recognition related tasks. Boito *et al.* [22] investigated the impact of pretrained data distribution on the fairness performance across subgroups. By pretraining the wav2vec 2.0 with gender-specific and different proportion of gender data, it is demonstrated that the fairness is related to downstream integration and balanced-gender pretraining data does not necessarily reduce bias.

To mitigate the bias in ASR on the groups across geographic locations and demographic attributes, Dheram *et al.* [48] proposed an initial method, oversampling under minority groups and undersampling majority groups, to reduce the performance gap between different cohorts. Liu *et al.* [100] presented results from multiple ASR models on the Casual Conversations dataset and observed the significant WER difference across gender and ethic. To accurately evaluate the ASR fairness issue on racial demographics, Liu *et al.* [105] adopted mixed-effects Poisson regression to mitigate the negative influence from nuisance factors, such as speaker, context, phoneme, prosody, etc.

6.2.3 Fairness in Speech Emotion Recognition

Fairness in SER is important across multiple areas because the performance differences resulting from gender and race are significant in most of emotion recognition scenarios [149]. Gorrostieta *et al.* [72] investigated gender-based bias in speech emotion recognition and mitigated unwanted bias through adversarial training and additional weight for the objective function. In recent years, the large-scale self-supervised learning (SSL) model has become popular in computer vision, language processing, and audio processing. With the widespread use of SSL, Wagner *et al.* [176] have shone light on the performance variance and demonstrated that transformer-based SSL models have moderate fairness scores in the SER area.

7 Future Directions

In this section, we discuss the main challenges and potential research opportunities for speech-centric trustworthy machine learning in order to inspire readers to research this field further.

7.1 Privacy

In recent years, self-supervised learning speech models such as Wav2Vec 2.0 [14] and Whisper [136] have established the SOTA performance for many downstream speech tasks such as ASR. However, privacy-related topics, such as MIAs, on these emerging model techniques have not been explored extensively. As we discussed earlier, Tseng *et al.* [171] was the only one to investigate MIA threats to SSL speech models. Their results imply that the latent speech representation of SSL models holds the membership information of the input speech signals. However, they only perform some preliminary defenses on MIAs and findings from these defenses are limited. Therefore, it is valuable and critical to extend the work presented in Tseng *et al.* [171] to broader SSL speech models, downstream speech tasks, and MIA mitigation strategies.

In addition, there exist more property inferences where current PIAs have not been explored but are of demand in speech-centric ML, e.g., age, health status, etc. As our review points out, the majority of literature focuses on gender obfuscation, while none of the studies attempts to generalize the existing approaches to demographics like age or race. Apart from the lack of exploration of broader demographics, many studies only evaluate their works on several clean-audio benchmarks like Librispeech [128] and IEMOCAP [26], while the robustness and efficacy of many proposed privacy-enhancing approaches on the more dynamic recording conditions are unknown.

As we also presented in the Privacy review section, Federated Learning has become an emerging research topic in almost every field of ML. Nevertheless, compared to NLP and CV domains, the FL on speech-centric tasks stays largely unexplored. One of the significant challenges is to enable the training of the ASR system in the FL setting. Due to the nature of the ASR modeling, expensive computations are typically mandatory, while most mobile devices cannot afford to train and run these heavy computing models. Therefore, it is urgent and essential to investigate efficient FL training approaches for ASR models. Last but not least, most literature to date focuses on dealing with missing labels and decoupling heterogeneity conditions in FL, and fewer works are targeting privacy risks in federated speech learning. As a result, an exciting but critical research direction is systematically studying the privacy risks in federated speech modeling.

7.2 Safety

While consideration has been given to speech-specific evasion attacks, there has been far less work on audio-modality defenses. Speech signals possess unique structural and temporal properties that distinguish them from images. Traditional defense methods from the computer vision domain fail to fully leverage these attributes in aid of adversarial robustness. Furthermore, there have been some works [135, 183] that consider "over the air attacks" in which the adversarial audio travels to the model by means of an acoustic channel (as opposed to being directly fed to the model input). More work is needed in this area to gauge the feasibility and threat level of evasion attacks in this more realistic setting. Finally, there is relatively little work on poisoning attacks and defenses for speech systems. Given that large-scale datasets are increasingly procured from unverified sources (i.e. internet posts, client data in FL), it is critical that we better understand the potential risks and how to mitigate them.

7.3 Fairness

As outlined in the previous section, research on fairness and bias mitigation in the speech domain is limited to a few preliminary works. These are typically restricted to fairness evaluations on small attribute-balanced datasets for either gender or accent. With the introduction of newer datasets (Table 3), we hope to see more fairness studies along the demographic attributes of ethnicity, language, etc., as well as intersectional attributes (e.g, female Spanish vs male Spanish). Furthermore, there has been little work exploring the extent of bias in recent SOTA models such as Wav2Vec2.0 [14] and Whisper [136] for both ASR and ASV. There has also been little to no work in other speech tasks such as SER. Finally, the need for fairness-specific metrics is highlighted by Peri *et al.* [131], with most of the existing literature using standard metrics such as EER and WER.

7.4 Balance between Fairness, Privacy, and Safety

Despite the tremendous effort in designing trustworthy machine learning techniques over the last decade, most of these applications attempt to isolate one single aspect of trustworthiness in their modeling work. Consequently, a limited amount of work explores the impact of mitigating one trustworthy risk over other trustworthy challenges. Recently, Chang and Shokri [30] demonstrated that privacy and fairness are often opposed to each other in trustworthy machine learning. The author showed that increasing model fairness requires optimizing objectives that typically constrain the model to perform equally on every subgroup, leading the model to memorize training samples from the unprivileged subgroups. Hence, enhancing model fairness frequently raises significant privacy concerns in exposing the private information of the training data.

Apart from investigating the relationship between fairness and privacy, Xu *et al.* [182] found that safety-awareness learning poses a disparate impact on the fairness risk of subgroups. The author also proposed a Fair-Robust-Learning (FRL) framework to enhance the model fairness while performing adversarial defenses. However, both of these works solely investigate computer vision applications, while the interplay between privacy, fairness, and safety in speech-centric applications remains largely unexplored. It is, therefore, critical for future researchers to investigate the interactions between different trustworthiness aspects in speech-centric applications.

7.5 Trustworthy Multimodal Applications

Besides speech-centric applications that we discussed in this article, speech signals have broader usage cases in diverse multimodal applications, such as multimedia applications, audio-visual understanding, and multimodal sentiment analysis [90]. Typically, multimodal applications are reported with a higher system performance than unimodal models. Consequently, one promising future research direction is to explore speech-based multimodal applications that attempt to achieve more satisfactory balances between fairness, privacy, and safety.

8 Conclusion

As machine learning becomes more prevalent in speech-centric modeling, the scientific community has also become more aware of concerns and risks over the trustworthiness of these systems. This survey paper aims to provide a comprehensive and systematic summary of the recent efforts made to protect privacy, ensure fairness, and defend against adversarial attacks in these speech-centric systems. Increasingly, we identify several open problems of importance to address, such as investigating the extent of bias in recent SOTA models. Through this review, we hope to provide the necessary knowledge for future research in speech-centric trustworthy ML.

Acknowledgement

This work was supported by DARPA (Grant No: HR00112020009) and USC Amazon Center for Secure and Trusted Machine Learning.

References

[1] H. Abdullah, M. S. Rahman, W. Garcia, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, “Hear “no evil”, see “kenansville”*: Efficient and transferable black-box attacks on speech recognition and voice identification systems,” in *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2021, 712–29.

[2] R. Acharyya, S. Das, A. Chattoraj, and M. I. Tanveer, “FairyTED: A fair rating predictor for TED talk data,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, 2020, 338–45.

[3] H. Aghakhani, L. Schönherr, T. Eisenhofer, D. Kolossa, T. Holz, C. Kruegel, and G. Vigna, “VenoMave: Targeted poisoning against speech recognition,” *arXiv preprint arXiv:2010.10682*, 2020.

[4] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, “Preech: A system for Privacy-preserving speech transcription,” in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, 2703–20.

[5] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, 116, 2020, 56–76.

[6] A. Akman, H. Coppock, A. Gaskell, P. Tzirakis, L. Jones, and B. W. Schuller, “Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges,” *Frontiers in Digital Health*, 4, 2022.

[7] R. Aloufi, H. Haddadi, and D. Boyle, “Privacy-preserving voice analysis via disentangled representations,” in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, 1–14.

[8] M. Alzantot, B. Balaji, and M. Srivastava, “Did you hear that? adversarial examples against automatic speech recognition,” *arXiv preprint arXiv:1801.00554*, 2018.

[9] E. Amid, O. D. Thakkar, A. Narayanan, R. Mathews, and F. Beaufays, “Extracting targeted training data from ASR models, and how to mitigate it,” in *Proc. Interspeech 2022*, 2022, 2803–7, DOI: 10.21437/Interspeech.2022-10895.

[10] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias,” in *Ethics of Data and Analytics*, Auerbach Publications, 2016, 254–64.

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.

[12] P. Arora and T. Chaspari, “Exploring siamese neural network architectures for preserving speaker identity in speech emotion classification,” in *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, 2018, 15–8.

[13] M. Asad, A. Moustafa, and T. Ito, "FedOpt: Towards communication efficiency and privacy preservation in federated learning," *Applied Sciences*, 10(8), 2020, 2864.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, 33, 2020, 12449–60.

[15] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[16] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, 81(2), 2010, 121–48.

[17] S. P. Bayerl, T. Frassetto, P. Jauernig, K. Riedhammer, A.-R. Sadeghi, T. Schneider, E. Stapf, and C. Weinert, "Offline model guard: Secure and private ML on mobile devices," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2020, 460–5.

[18] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for AI," *Communications of the ACM*, 64(7), 2021, 58–65.

[19] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, 50(1), 2021, 3–44.

[20] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, 387–402.

[21] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.

[22] M. Z. Boito, L. Besacier, N. A. Tomashenko, and Y. Estève, "A study of gender impact in self-supervised models for speech-to-text systems," in *INTERSPEECH*, ISCA, 2022, 1278–82.

[23] F. Brasser, T. Frassetto, K. Riedhammer, A.-R. Sadeghi, T. Schneider, and C. Weinert, "VoiceGuard: Secure and private speech processing," in *Proc. Interspeech 2018*, 2018, 1303–7, DOI: 10.21437/Interspeech.2018-2032.

[24] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "VoxSRC 2021: The third VoxCeleb speaker recognition challenge," *CoRR*, abs/2201.04583, 2022, arXiv: 2201.04583.

[25] A. Brown, J. Huh, A. Nagrani, J. S. Chung, and A. Zisserman, "Playing a part: Speaker verification at the movies," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6174–8.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, 42(4), 2008, 335–59.

[27] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Computing Surveys (CSUR)*, 54(6), 2021, 1–38.

[28] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

[29] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*, IEEE, 2018, 1–7.

[30] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2021, 292–303.

[31] Y. Chang, S. Laridi, Z. Ren, G. Palmer, B. W. Schuller, and M. Fisichella, "Robust federated learning against adversarial attacks for speech emotion recognition," *arXiv preprint arXiv: 2203.04696*, 2022.

[32] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.

[33] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2021, 694–711.

[34] L.-W. Chen and A. Rudnicky, "Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.

[35] X. Chen, Z. Li, S. Setlur, and W. Xu, "Exploring racial and gender disparities in voice biometrics," *Scientific Reports*, 12(1), 2022, 1–12.

[36] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[37] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—a survey," *Proceedings of the IEEE*, 110(4), 2022, 476–507.

[38] G. Chennupati, M. Rao, G. Chadha, A. Eakin, A. Raju, G. Tiwari, A. K. Sahu, A. Rastrow, J. Droppo, A. Oberlin, *et al.*, "ILASR: privacy-preserving incremental learning for automatic speech recognition at production scale," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, 2780–8.

[39] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[40] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *International Conference on Machine Learning*, PMLR, 2017, 854–63.

[41] L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, J. Edwards, *et al.*, "The state of speech in HCI: Trends, themes and challenges," *Interacting with Computers*, 31(4), 2019, 349–71.

[42] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*, PMLR, 2019, 1310–20.

[43] X. Cui, S. Lu, and B. Kingsbury, "Federated acoustic modeling for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6748–52.

[44] D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems.," in *Ijcai*, Vol. 17, 2017, 4691–7.

[45] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, *et al.*, "The YouTube video recommendation system," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010, 293–6.

[46] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 2010, 788–98.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[48] P. Dheram, M. Ramakrishnan, A. Raju, I. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, "Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities," in *INTERSPEECH*, ISCA, 2022, 1268–72.

[49] M. Dias, A. Abad, and I. 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, 2057–61.

[50] D. Dimitriadis, K. Kumatani, R. Gmyr, Y. Gaur, and S. E. Eskimez, "A federated approach in training acoustic models," in *Interspeech*, 2020, 981–5.

[51] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, Springer, 2008, 1–19.

[52] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, 214–26.

[53] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, Springer, 2006, 265–84.

[54] D. L. Elworth and S. Kim, "HEKWS: Privacy-Preserving convolutional neural network-based keyword spotting with a ciphertext packing technique," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2022, 1–6.

[55] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Class-conditional defense GAN against end-to-end speech attacks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 2565–9.

[56] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Cyclic defense gan against speech adversarial attacks," *IEEE Signal Processing Letters*, 28, 2021, 1769–73.

[57] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, 1459–62.

[58] T. Feng, H. Hashemi, M. Annavaram, and S. S. Narayanan, "Enhancing privacy through domain adaptive noise injection for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 7702–6.

[59] T. Feng, H. Hashemi, R. Hebbar, M. Annavaram, and S. S. Narayanan, "Attribute inference attack of speech emotion recognition in federated learning settings," *arXiv preprint arXiv:2112.13416*, 2021.

[60] T. Feng, A. Nadarajan, C. Vaz, B. Booth, and S. 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, 33–8.

[61] T. Feng and S. Narayanan, "Privacy and utility preserving data transformation for speech emotion recognition," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2021, 1–7.

[62] T. Feng and S. Narayanan, "Semi-FedSER: Semi-supervised Learning for Speech Emotion Recognition On Federated Learning using Multiview Pseudo-Labeling," *arXiv preprint arXiv:2203.08810*, 2022.

[63] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition on federated learning," in *Proc. Interspeech 2022*, 2022, 5055–9, DOI: 10.21437/Interspeech.2022-10060.

[64] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *International Conference on Computational Science and Its Applications*, Springer, 2020, 77–93.

[65] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, “Strip: A defense against trojan attacks on deep neural networks,” in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, 113–25.

[66] J. Garofolo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) Sennheiser LDC93S6B,” *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[67] Y. Ge, Q. Wang, J. Zhang, J. Zhou, Y. Zhang, and C. Shen, “WaveFuzz: A clean-label poisoning attack to protect your voice,” *arXiv preprint arXiv:2203.13497*, 2022.

[68] J. Geiping, L. Fowl, W. R. Huang, W. Czaja, G. Taylor, M. Moeller, and T. Goldstein, “Witches’ brew: Industrial scale data poisoning via gradient matching,” *arXiv preprint arXiv:2009.02276*, 2020.

[69] B. van Giffen, D. Herhausen, and T. Fahse, “Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods,” *Journal of Business Research*, 144, 2022, 93–106.

[70] Y. Gong and C. Poellabauer, “Crafting adversarial examples for speech paralinguistics applications,” *arXiv preprint arXiv:1711.03280*, 2017.

[71] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[72] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, “Gender De-Biasing in speech emotion recognition,” in *INTERSPEECH*, 2019, 2823–7.

[73] F. Granqvist, M. Seigel, R. van Dalen, Á. Cahill, S. Shum, and M. Paulik, “Improving on-device speaker verification using federated learning with privacy,” in *Proc. Interspeech 2020*, 2020, 4328–32, DOI: 10.21437/Interspeech.2020-2944.

[74] T. Gu, B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” *arXiv preprint arXiv:1708.06733*, 2017.

[75] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv: 2005.08100*, 2020.

[76] D. Guliani, F. Beaufays, and G. Motta, “Training speech recognition models with federated learning: A quality/cost framework,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 3080–4.

[77] D. Guliani, L. Zhou, C. Ryu, T.-J. Yang, H. Zhang, Y. Xiao, F. Beaufays, and G. Motta, “Enabling on-device training of speech recognition models with federated dropout,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 8757–61.

[78] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice- indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, 1–6.

[79] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[80] A. Hard, K. Partridge, N. Chen, S. Augenstein, A. Shah, H. J. Park, A. Park, S. Ng, J. Nguyen, I. Lopez-Moreno, R. Mathews, and F. Beaufays, "Production federated keyword spotting via distillation, filtering, and joint federated-centralized training," in *Proc. Interspeech 2022*, 2022, 76–80, DOI: 10.21437/Interspeech.2022-11050.

[81] A. Hard, K. Partridge, C. Nguyen, N. Subrahmanya, A. Shah, P. Zhu, I. L. Moreno, and R. Mathews, "Training keyword spotting models on non-iid data with federated learning," *arXiv preprint arXiv:2005.10406*, 2020.

[82] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, 29, 2016.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[84] R. Hebbar, K. Somandepalli, R. Peri, R. Travadi, T. Tuplin, F. Rivera, and S. Narayanan, "A computational tool to study vocal participation of women in UN-ITU meetings," in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, 2021, 1–4.

[85] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, 37, 2020, 100270.

[86] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "{WaveGuard}: understanding and mitigating audio adversarial examples," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, 2273–90.

[87] N. D. Huynh, M. R. Bouadjenek, I. Razzak, K. Lee, C. Arora, A. Hassani, and A. Zaslavsky, "Adversarial attacks on speech recognition systems for mission-critical applications: A survey," *arXiv preprint arXiv:2202.10594*, 2022.

[88] A. Irum and A. Salman, "Speaker verification using deep neural networks: A," *International Journal of Machine Learning and Computing*, 9(1), 2019.

[89] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, 2020, 7985–93.

[90] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Information Fusion*, 73, 2021, 22–71.

[91] M. Jin, C. Ju, Z. Chen, Y. Liu, J. Droppo, and A. Stolcke, "Adversarial reweighting for speaker verification fairness," in *INTERSPEECH*, ISCA, 2022, 4800–4.

[92] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, PMLR, 2020, 5132–43.

[93] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," *arXiv preprint arXiv:1811.01312*, 2018.

[94] A. Köhn, F. Stegen, and T. Baumann, "Mining the spoken wikipedia for speech data and beyond," 2016.

[95] S. Kumar, L. T. Nguyen, M. Zeng, K. Liu, and J. Zhang, "Sound shredding: Privacy preserved audio sensing," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, 2015, 135–40.

[96] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Information Processing Systems*, 30, 2017.

[97] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6341–5.

[98] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 1778–87.

[99] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, 54(2), 2021, 1–36.

[100] C. Liu, M. Picheny, L. Sari, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, IEEE, 2022, 6162–6, DOI: 10.1109/ICASSP43922.2022.9747501.

[101] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy AI: A computational perspective," *ACM Transactions on Intelligent Systems and Technology*, 14(1), 2022, 1–59.

[102] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International Symposium on Research in Attacks, Intrusions, and Defenses*, Springer, 2018, 273–94.

[103] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," 2017.

[104] Y. Liu, A. Kapadia, and D. Williamson, "Preventing sensitive-word recognition using self-supervised learning to preserve user-privacy for automatic speech recognition," in *Proc. Interspeech 2022*, 2022, 4207–11, DOI: 10.21437/Interspeech.2022-85.

[105] Z. Liu, I. Veliche, and F. Peng, "Model-Based Approach for Measuring the Fairness in ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, IEEE, 2022, 6532–6.

[106] Z. Liu, Z. Wu, C. Gan, L. Zhu, and S. Han, "Datamix: Efficient privacy-preserving edge-cloud inference," in *European Conference on Computer Vision*, Springer, 2020, 578–95.

[107] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, 5337–41.

[108] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," in *Logic, Language, and Security*, Springer, 2020, 189–202.

[109] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[110] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, 80(6), 2021, 9411–57.

[111] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A comparative study of speech anonymization metrics," in *Proc. Interspeech 2020*, 2020, 1708–12, DOI: 10.21437/Interspeech.2020-2248.

[112] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," *Interspeech 2020*, 2020.

[113] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, 1273–82.

[114] N. Mehlman, A. Sreeram, R. Peri, and S. Narayanan, "Mel frequency spectral domain defenses against adversarial attacks on speech recognition systems," *JASA Express Letters*, 3(3), 2023, 035208.

[115] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, 54(6), 2021, 1–35.

[116] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[117] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""I don't think these devices are very culturally sensitive." – Impact of automated speech recognition errors on African Americans," *Frontiers in Artificial Intelligence*, 4, 2021, 725911.

[118] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, 6462–8.

[119] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, 19(6), 2018, 1236–46.

[120] F. Mireshghallah, M. Taram, P. Vepakomma, A. Singh, R. Raskar, and H. Esmaeilzadeh, "Privacy in deep learning: A survey," *arXiv preprint arXiv:2004.12254*, 2020.

[121] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, *et al.*, "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, 7(1), 2020, 354.

[122] L. Muñoz-Gonzãlez, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, 27–38.

[123] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[124] K. Nandury, A. Mohan, and F. Weber, "Cross-silo federated training in the cloud with diversity scaling and semi-supervised learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 3085–9.

[125] A. Nelus, S. Rech, T. Koppelmann, H. Biermann, and R. Martin, "Privacy-preserving siamese feature extraction for gender recognition versus speaker identification," in *Proc. Interspeech 2019*, 2019, 3705–9, DOI: 10.21437/Interspeech.2019-1148.

[126] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Pérez-Cabo, "No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation," *IEEE Transactions on Information Forensics and Security*, 12(11), 2017, 2640–53.

[127] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, 1999, 223–38.

[128] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, 5206–10.

[129] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[130] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2), 2012, 397–406.

[131] R. Peri, K. Somandepalli, and S. Narayanan, "To train or not to train adversarially: A study of bias mitigation strategies for speaker recognition," *arXiv preprint arXiv:2203.09122*, 2022.

[132] C. Pierre, A. Larcher, and D. Jouvet, "Are disentangled representations all you need to build speaker anonymization systems?" In *Proc. Interspeech 2022*, 2022, 2793–7, DOI: 10.21437/Interspeech.2022-10586.

[133] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[134] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, 82–94.

[135] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International Conference on Machine Learning*, PMLR, 2019, 5231–40.

[136] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *OpenAI Blog*, 2022.

[137] Y. Rahulamathavan, K. R. Sutharsini, I. G. Ray, R. Lu, and M. Rajarajan, "Privacy-preserving iVector-based speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3), 2018, 496–506.

[138] S. S. Rajan, S. Udeshi, and S. Chattopadhyay, “AequeVox: Automated Fairness Testing of Speech Recognition Systems,” in *Fundamental Approaches to Software Engineering - 25th International Conference, FASE 2022, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Munich, Germany, April 2-7, 2022, Proceedings*, ed. E. B. Johnsen and M. Wimmer, Vol. 13241, *Lecture Notes in Computer Science*, Springer, 2022, 245–67.

[139] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, 10(1-3), 2000, 19–41.

[140] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,” *arXiv preprint arXiv: 2101.06329*, 2021.

[141] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, “Face recognition: too bias, or not too bias?” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 0–1.

[142] A. Rousseau, P. Deléglise, and Y. Esteve, “TED-LIUM: an Automatic Speech Recognition dedicated corpus.,” in *LREC*, 2012, 125–9.

[143] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, “The 2018 NIST Speaker Recognition Evaluation,” in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, ed. G. Kubin and Z. Kacic, ISCA, 2019, 1483–7.

[144] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, “The 2016 NIST Speaker Recognition Evaluation,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, ISCA, 2017, 1353–7.

[145] S. Safavi, M. Russell, and P. Jančovič, “Automatic speaker, age-group and gender identification from children’s speech,” *Computer Speech & Language*, 50, 2018, 141–56.

[146] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, “Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding,” *arXiv preprint arXiv:1808.05665*, 2018.

[147] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, “Poison frogs! targeted clean-label poisoning attacks on neural networks,” *Advances in Neural Information Processing Systems*, 31, 2018.

[148] M. A. Shah, J. Szurley, M. Mueller, A. Mouchtaris, and J. Droppo, “Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models to Membership Inference Attacks,” in *Proc. Interspeech 2021*, 2021, 891–5, DOI: 10.21437/Interspeech.2021-1188.

[149] G. Sharma and A. Dhall, “A survey on automatic multimodal emotion recognition in the wild,” in *Advances in Data Science: Methodologies and Applications*, Springer, 2021, 35–64.

[150] M. Shoemate, K. Jett, E. Cowan, S. Colbath, J. Honaker, and P. Muthukumar, “Sotto Voce: Federated speech recognition with differential privacy guarantees,” *arXiv preprint arXiv: 2207.07816*, 2022.

[151] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2017, 3–18.

[152] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, 529(7587), 2016, 484–9.

[153] P. Smaragdis and M. Shashanka, “A framework for secure speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), 2007, 1404–13.

[154] N. A. Smuha, “The EU approach to ethics guidelines for trustworthy artificial intelligence,” *Computer Law Review International*, 20(4), 2019, 97–106.

[155] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5329–33.

[156] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, “Learning controllable fair representations,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, 2164–73.

[157] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, “Face recognition algorithm bias: Performance differences on images of children and adults,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, 0–4.

[158] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, “Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?” In *Proc. Interspeech 2019*, 2019, 3700–4, DOI: 10.21437/Interspeech.2019-2415.

[159] J. Steinhardt, P. W. W. Koh, and P. S. Liang, “Certified defenses for data poisoning attacks,” *Advances in Neural Information Processing Systems*, 30, 2017.

[160] D. Stoidis and A. Cavallaro, “Generating gender-ambiguous voices for privacy-preserving speech recognition,” *arXiv preprint arXiv: 2207.01052*, 2022, DOI: 10.21437/Interspeech.2022-11322.

[161] L. L. Stoll, “Finding Difficult Speakers in Automatic Speaker Recognition,” *PhD thesis*, EECS Department, University of California, Berkeley, 2011.

[162] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, “Domain adversarial training for accented speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 4854–8.

[163] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[164] N. Tomashenko, S. Mdhaffar, M. Tommasi, Y. Estève, and J.-F. Bonastre, “Privacy attacks for automatic speech recognition acoustic models in a federated learning framework,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6972–6.

[165] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, “The VoicePrivacy 2022 Challenge Evaluation Plan,” *arXiv preprint arXiv:2203.12468*, 2022.

[166] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’Brien, *et al.*, “The VoicePrivacy 2020 Challenge: Results and findings,” *Computer Speech & Language*, 74, 2022, 101362.

[167] W. Toussaint and A. Y. Ding, “SVEva Fair: A Framework for Evaluating Fairness in Speaker Verification,” *CoRR*, abs/2107.12049, 2021.

[168] B. Tran, J. Li, and A. Madry, “Spectral signatures in backdoor attacks,” *Advances in Neural Information Processing Systems*, 31, 2018.

[169] E. Trentin and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” *Neurocomputing*, 37(1-4), 2001, 91–126.

[170] A. Tripathi, A. Mohan, S. Anand, and M. Singh, “Adversarial learning of raw speech features for domain invariant speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5959–63.

[171] W.-C. Tseng, W.-T. Kao, and H.-y. Lee, “Membership Inference Attacks Against Self-supervised Speech Models,” in *Proc. Interspeech 2022*, 2022, 5040–4, DOI: 10.21437/Interspeech.2022-11245.

[172] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, “Privacy-preserving speech emotion recognition through semi-supervised federated learning,” in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, 2022, 359–64.

[173] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 30, 2017.

[174] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, IEEE, 2018, 1–7.

[175] P. Voigt and A. Von dem Bussche, "The EU General Data Protection Regulation (GDPR)," *A Practical Guide,* 1st Ed., Cham: Springer International Publishing, vol. 10(3152676), 2017, 10–5555.

[176] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *CoRR*, abs/2203.07378, 2022.

[177] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2019, 707–23.

[178] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv: 2101.00390*, 2021.

[179] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[180] S. H. Weinberger and S. A. Kunath, "The Speech Accent Archive: towards a typology of English accents," in *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, Brill, 2011, 265–81.

[181] B. Xiong, H. Fan, K. Grauman, and C. Feichtenhofer, "Multiview pseudo-labeling for semi-supervised learning from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 7209–19.

[182] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International Conference on Machine Learning*, PMLR, 2021, 11492–501.

[183] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *arXiv preprint arXiv:1810.11793*, 2018.

[184] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks," *arXiv preprint arXiv: 1703.01340*, 2017.

[185] T.-J. Yang, D. Guliani, F. Beaufays, and G. Motta, "Partial variable training for efficient on-device federated learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 4348–52.

[186] Z. Yang, B. Li, P.-Y. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," *arXiv preprint arXiv:1809. 10875*, 2018.

[187] J. C. Yau, B. Girault, T. Feng, K. Mundnich, A. Nadarajan, B. M. Booth, E. Ferrara, K. Lerman, E. Hsieh, and S. Narayanan, "TILES-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit," *Scientific Data*, 9(1), 2022, 536.

[188] W. Yu, J. Freiwald, S. Tewes, F. Huennemeyer, and D. Kolossa, "Federated learning in ASR: Not as easy as you think," in *Speech Communication; 14th ITG Conference*, VDE, 2021, 1–5.

[189] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, 2236–46.

[190] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, 39–49.

[191] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv: 1710.09412*, 2017.

[192] T. Zhang, T. Feng, S. Alam, S. Lee, M. Zhang, S. S. Narayanan, and S. Avestimehr, "FedAudio: A Federated learning benchmark for audio tasks," *arXiv preprint arXiv:2210.15707*, 2022.

[193] P. Zheng, Z. Cai, H. Zeng, and J. Huang, "Keyword spotting in the homomorphic encrypted domain using deep complex-valued CNN," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, 1474–83.

[194] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 920–4.

[195] H. Zhu, J. Wang, G. Cheng, P. Zhang, and Y. Yan, "Decoupled Federated Learning for ASR with Non-IID Data," in *Proc. Interspeech 2022*, 2022, 2628–32, DOI: 10.21437/Interspeech.2022-720.