Towards Privacy-Preserving Speech Data Publishing

Jianwei Qian*, Feng Han†, Jiahui Hou*, Chunhong Zhang§, Yu Wang‡, Xiang-Yang Li†
* Department of Computer Science, Illinois Institute of Technology
† School of Computer Science and Technology, University of Science and Technology of China
‡ Department of Computer Science, University of North Carolina at Charlotte
§ School of Information and Communication Engineering, Beijing University of Posts and Telecommunications

***Abstract*—Privacy-preserving data publishing has been a heated research topic in the last decade. Numerous ingenious attacks on users' privacy and defensive measures have been proposed for the sharing of various data, varying from relational data, social network data, spatiotemporal data, to images and videos. Speech data publishing, however, is still untouched in the literature. To fill this gap, we study the privacy risk in speech data publishing and explore the possibilities of performing data sanitization to achieve privacy protection while preserving data utility simultaneously. We formulate this optimization problem in a general fashion and present thorough quantifications of privacy and utility. We analyze the sophisticated impacts of possible sanitization methods on privacy and utility, and also design a novel method – key term perturbation for speech content sanitization. A heuristic algorithm is proposed to personalize the sanitization for speakers to restrict their privacy leak (p-leak limit) while minimizing the utility loss. The simulations of linkage attacks and sanitization on real datasets validate the necessity and feasibility of this work.**

I. Introduction

We have witnessed the pervasiveness of voice-based human-computer interaction: input keyboards, web search, voice assistants, and voice authentication. These applications have brought numerous benefits to our daily lives. The big data era has spawned many data trading platforms since the immense value of big data has been prominently manifested. The sharing or publishing of speech data is also going to be an irresistible trend. From search engine tech giants to telecom companies, their speech data are not just mined to improve the service but could be shared to third-parties for profit as well. For instance, Samsung and Apple have admitted voice data sharing to third-parties [1], [2]. Meanwhile, speech data may also be published to foster research on spoken language analysis, *e.g.* TIMIT and NIST SRE datasets.

Speech data contain a rich amount of information about the speakers that can be inferred by mining their search history and voice commands, including their demographics, preferences, online behaviors, living habits, and interpersonal relations. Some of such information might be sensitive to the speakers. For the sake of their privacy, all the personally identifiable information (PII) associated with the data to be published has to be removed, including names, phone numbers, and device IDs of the speakers. However, this is far from enough to prevent malicious third-parties (*attackers*) from undermining the speakers' privacy.

There is still a **privacy risk** called *linkage attack*, *i.e.* linking an anonymous speech recording to a real person to infer her sensitive information. This could be achieved by attackers with some background knowledge and reasoning ability. Linkage attacks at speech data may compromise the person's privacy from four aspects. *First*, the content of the voice recording conveys a lot of demographics and life details about the person. Demographics that can be inferred include, but are not limited to, gender, age, education level, ethnicity, geographic region [4], social status [6], and personality [35]. The life details may leak privacy too, such as schedules the person added to Google calendar, products she purchased on Amazon Alexa, and even text messages and emails she wrote by voice input. From these details, the attacker is able to extract various private attributes and paint an accurate profile of this person. *Second*, the attacker can learn the person's demographic categories by analyzing her voice solely. In fact, an intimidating number of demographics can be mined from voice, referred to as *voice attributes*, such as age, gender, ethnicity, geographic region (accent), height [17], emotion [19], and even health condition [27]. *Third*, the person's voiceprint is leaked. Voiceprint as a type of biometrics is widely applied in emerging systems for authentication. Unlike password that can be changed once stolen, voiceprint is unchangeable. Once it is leaked to a miscreant, we will never feel safe again to adopt voice authentication to secure our properties due to the fear of identity theft. *Finally*, the attacker gets to know the fact that the person belongs to this dataset, which might be sensitive, for example when the dataset is a collection of utterances of heart disease patients. This is known as *membership privacy*.

The leak of voiceprint further results in three **security risks**. The first is identity theft as aforementioned. The attacker may commit *spoofing attacks* to voice authentication systems [31]. Besides, the victim could suffer from *reputation attacks*. The attacker can fabricate recordings that sound like the victim's voice but has indecent or illegal content, to damage her reputation or frame her up, *e.g.* fake Obama speech[1]. Moreover, the victim may experience *fraud attacks*. The attacker may use her voice to agree on some terms to sign up for paid service and authorize bogus charges on a credit card.[2]

We seek to answer four questions. 1) What is the potential risk of privacy leak in speech data publishing? 2) How to design sanitization methods suitable for speech data? 3) How to quantify their influence on privacy level and data utility?

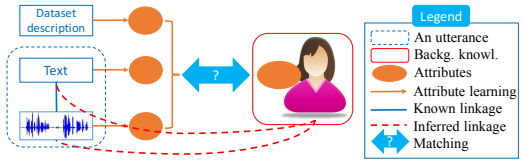
[1]Fake Obama speech, https://goo.gl/pnR3VK
[2]The 'Can you hear me?' fraud, https://goo.gl/Wy3e7u

4) How to select a good combination of sanitization methods and parameters to reach a balance of privacy and utility?

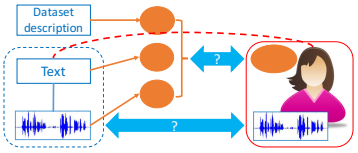
To this end, we are facing the following **challenges**. First, existing privacy definitions may not fit for speech data. Unlike tables, texts, or images, speech is not a pure data form but a tightly-coupled integration of voice and text (*i.e.* speech content). Privacy in text data has always been hard to define, because it is person-specific and context-dependent, not to mention privacy in text plus voice. Second, what is the utility of speech data is still unknown. It is very difficult to quantify the utility because it depends on multiple factors such as audio quality, speaker diversity, speech content relevance, etc. It is also up to what the consumer uses the data for. Third, it is challenging to find a "bliss point" of data perturbation such that both privacy and utility are well preserved. We want to make speakers' voices indistinguishable to defend against linkage attacks by voiceprint; meanwhile, we also hope to retain voice diversity in the published dataset. We want to remove the hidden demographics from the speaker's voice but also hope to preserve the detailed voice idiosyncrasies. We want to sanitize the speech content by truncating sensitive parts but we are reluctant to damage the semantics and devalue the data. Privacy and utility seem to be always contradictory here. A *naïve solution* is voice shuffling, that is, to change the voice of each person into another one's voice in the dataset. Unfortunately, it relies on targeted voice conversion, an immature technology [16]. The output audio usually has undesirable quality, and it cannot hide every single fine detail of a person's distinctive voice.

In this paper, we generally formulate the privacy-preserving speech data publishing problem and instantiate it with our definitions of privacy leak and utility loss (§III). We discuss possible data sanitization approaches in three directions and analyze how they restrict privacy leak and damage data utility (§III-D). As an example, we design a TF-IDF based key term sanitization method to perturb the private speech content (§IV), and analyze how voice conversion and speech synthesis influence privacy and utility (§V). Then, by applying a heuristic algorithm, we can simplify the optimization problem and find a near optimal selection of methods and parameters for speech data sanitization (§VI).

Our **contributions:** *First*, we highlighted the privacy risk of speech data publishing. We quantified privacy leak and utility loss from four aspects each and proposed a privacy notion *p-leak limit* to limit every speaker's privacy leak from the published data. This is formulated as an optimization problem, *i.e.* minimize utility loss, subject to p-leak limit. A heuristic algorithm is proposed for it. *Second*, besides exploring the possibility of using existing speech processing technologies for data sanitization, we proposed an original approach – TF-IDF based speech content sanitization. We analyzed the impact of all these approaches on privacy leak and utility loss. *Third*, we simulated linkage attacks on real datasets to show the vulnerability of the published speech data. We also simulated various data sanitization approaches to evaluate their effects on privacy and utility.



(a) Victim's voice is unknown



(b) Victim's voice is known

Fig. 1: **Linkage attacks.** (a): when the attacker does not know the victim's voice, he identifies her by matching the attributes. (b): when the attacker knows her voice, he can use the attributes to reduce the search range and then quickly identify her by voice.

II. Background

A. Scenario

In speech data publishing, there are three parties involved – speakers, the data publisher, and the data consumer. The publisher is also the voice-driven service provider, who owns a collection of utterances generated by a group of speakers. Since the speakers have agreed to the terms to provide their speech data, the publisher is assumed to be trusted. The publisher either releases the dataset to the public or shares it with a third-party after removing the PII associated with the data. Those who have access to the dataset are referred to as data consumers. An *attacker* is a malicious data consumer who attempts to de-anonymize the utterances and steal the speakers' private information. The attacker usually possesses background knowledge of some people (*candidates*) in real life, *e.g.* their voices or attributes like gender, occupation, and region. He aims to pinpoints the target speaker (*victim*) among the candidates.

B. Speech Data Model

We model the speech dataset as $D = (des, \mathbb{U})$. Here, des refers to the dataset description. It could be a brief introduction of the settings where the voice was recorded like "in the lab" or the speakers' attributes like "graduate students of ABC University in their 20's". The set of utterances (*i.e.* voice recordings) is denoted by $\mathbb{U}$. An utterance is comprised of two elements – *voice* and *text*. Voice refers to the voiceprint, that is, the distinctive voice characteristics. Text refers to the speech content of the recording, which may be a short sentence or a long passage. Accordingly, the set of utterances $\mathbb{U} = (\mathcal{V}, \mathcal{T}, F)$ is decomposed into three parts in concept: voice set $\mathcal{V}$, text set $\mathcal{T}$, and a mapping $F : \mathcal{T} \to \mathcal{V}$. Each voice $v \in \mathcal{V}$ belongs to a distinct speaker. We assume each speaker has only one utterance in the dataset. If a speaker has multiple utterances, the publisher concatenates them or makes them sound like different speakers by voice conversion.

C. Linkage Attacks

The publishing of speech data is vulnerable to linkage attacks, which is realized by matching the attributes and/or voice of the anonymous speaker and those of the candidates. As illustrated in Fig. 1, we introduce two cases of linkage attacks; the difference lies in whether the attacker possesses a voice sample of the victim beforehand. Suppose the attacker selects an utterance $u = (v, t)$ as a target and attempts to find the speaker's real identity.

Let's first look at the case when the attacker does not know the victim's voice prior to the attack (Fig. 1(a)). The attacker first extracts some attributes about the victim from des, t, v. The attribute sets are denoted by A_1, A_2, A_3 respectively. He exploits A_1 to determine the search range (candidate set) and A_2, A_3 to further pinpoint the speaker by matching her attributes. For example, if $A_1 = \{\text{ocupation: students, university: ABC, lab: DEF}\}$, the attacker will select this group as candidates. The speaker can be successfully de-anonymized when there are adequate attributes in des, t, v and in the background knowledge. Then the victim is linked to v and t, which leads to the disclosure of her voiceprint and other sensitive information.

The attack gets easier if the attacker already knows the voices of the candidates (Fig. 1(b)). Likewise, he first utilizes A_1, A_2, A_3 to reduce the search space, then compares v with the candidates' voices by speaker recognition. Then the victim is linked to t, which causes the disclosure of her private speech content. Speaker recognition is very accurate at present [10] and voice features can be extracted from very short samples (0.3 s) [11]. In our experiment, we can identify a person from over 800 people with the accuracy 100%.

III. Problem Formulation

A. Privacy-Preserving Speech Data Publishing Problem

We first focus on formulating this problem as a general framework. As mentioned in §I, there are four layers of **privacy leak** in each person's speech data: text leak, voice attribute leak, voiceprint leak, and membership leak. The amount of leak is denoted by P_t, P_{va}, P_{vp}, P_m, respectively (see detailed illustrations in §III-B). The total amount of privacy leak of an utterance u (*i.e.* the total amount of published information about a speaker u) is

$$P^u = F_P(P_t^u, P_{va}^u, P_{vp}^u, P_m), \tag{1}$$

where F_P is decided by the publisher, which could be a linear combination or a supermodular function for example. P_m is the same for any $u \in \mathbb{U}$ whereas the other three are utterance-specific. The **privacy** of a dataset D lies in its ability to limit the disclosure of private information for each speaker, so as to thwart linkage attacks and reduce the potential privacy leak for the speaker. Thus, we have the following privacy definition.

Definition 1 (p-Leak limit): We say $D = (des, \mathbb{U})$ is p-leak limited, if and only if $P^u \le p$, $\forall u \in \mathbb{U}$. In other words, the potential privacy leak for every speaker in the dataset is at most p. We refer to p as the privacy leak budget.

Sanitization actions (parameters)	Privacy leak after sanitization ↓				Utility loss caused by sanitization ↑			
	Text P_t	Voice attribute P_{va}	Voice print P_{vp}	Membership P_m	Voice diversity U_v	Text authenticity U_t	Speech quality U_q	Data use U_d
Data description sanitization ($\boldsymbol{w}$)				$e(\boldsymbol{w})$ Eq. 6				$f(\boldsymbol{w})$ Eq. 6
Key term perturbation (x)	$g^u(\boldsymbol{x}^u)$ Eq. 12					$h^u(\boldsymbol{x}^u)$ Eq. 13		
Voice conversion (y)		$i^u(\boldsymbol{y}^u)$ Eq. 9	$j(\boldsymbol{y}^u)$ Eq. 14		$k(\boldsymbol{y}^1,\dots,\boldsymbol{y}^n, \boldsymbol{z}^1,\dots,\boldsymbol{z}^n)$ Eq. 11		$l(\boldsymbol{y}^u)$ Eq. 15	
Speech synthesis (z)		0	0					

Fig. 2: **Influence of sanitization actions on privacy leak and utility loss.** The functions e-l take the parameters $\boldsymbol{w}$-$\boldsymbol{z}$ as input and calculate the influence of the sanitization actions on specific aspects of privacy leak and utility loss. Example formulas can be referenced by the Eq. No. Blank entries represent no influence. The superscript u means utterance-specific. P_m, U_v, U_d are dataset-level while others are utterance-level.

Inevitably, the cost of privacy protection is **utility loss**. There are four aspects of data utility: voice diversity, text authenticity, speech quality, and data use clarity. Their *losses* are denoted by U_v, U_t, U_q, U_d, respectively (see details in §III-C). The total utility loss for D is defined as

$$U = F_U(U_v, U_t, U_q, U_d), \tag{2}$$

where F_U is decided by the publisher, which could be the max function, a linear combination, or a supermodular function. U_v, U_d are dataset-level utility, whereas U_t, U_q are utterance-level so they need to be aggregated over all utterances: $U_t = \Sigma_t(\{U_t^u \mid u \in \mathbb{U}\}), U_q = \Sigma_q(\{U_q^u \mid u \in \mathbb{U}\})$. Here, Σ_t, Σ_q are the aggregation functions, like sum, average, max, etc.

There could be various **sanitization actions** for speech data. We list four feasible actions in Fig. 2: data description sanitization, key term perturbation, voice conversion, and speech synthesis (see details in §III-D). Each has a distinct influence on privacy leak and utility loss. After data description sanitization on des with parameters $\boldsymbol{w}$, the membership leak becomes $P_m = e(\boldsymbol{w})$ and the loss of data use clarity is $U_d = f(\boldsymbol{w})$. The rest actions are applied to every utterance, so their parameters have a superscript u. For instance, after applying key term perturbation on u with parameters $\boldsymbol{x}^u$, we get $P_t^u = g^u(\boldsymbol{x}^u)$. A function with a superscript u means it is utterance-specific. After speech synthesis, P_{va}, P_{vp} both turn zero because the voice attribute and voiceprint are wiped out. Calculating the voice diversity loss is very complex, as it depends on the parameters selected for each of the n utterances and it is jointly decided by voice conversion and speech synthesis. We will instantiate the general functions e-l with concrete formulas, which can be referenced by the equation numbers in Fig. 2 (notice the concrete examples have their own notation for the parameters). However, there are no universal formulas for these functions. They are up to the implementations of the actions and the specific speech dataset.

In the **optimization problem**, the publisher aims to sanitize D to guarantee p-leak limit before publishing it, while causing as little utility loss as possible. Therefore, the problem is to minimize U, subject to $P^u \le p, \forall u \in \mathbb{U}$. According to Fig. 2,

Eq. 1, and Eq. 2, we have

$$U = F_U(\ k(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n, \boldsymbol{z}^1, \ldots, \boldsymbol{z}^n), \Sigma_t(\{h^u(\boldsymbol{x}^u)\}), \Sigma_q(\{l(\boldsymbol{y}^u)\}), f(\boldsymbol{w})\ ).$$

$$P^u = F_P\ (\ g^u(\boldsymbol{x}^u), \delta(\boldsymbol{z}^u) i^u(\boldsymbol{y}^u), \delta(\boldsymbol{z}^u) j(\boldsymbol{y}^u), e(\boldsymbol{w})\ ),$$

where $\delta(\cdot)$ is an impulse-like function ($\delta(\boldsymbol{z}^u) = 1$ if $\boldsymbol{z}^u = 0$; $= 0$ if $\boldsymbol{z}^u \neq 0$). When $\boldsymbol{z}^u \neq 0$, speech synthesis is applied to u so P_{va}, P_{vp} turn 0. A general definition of the privacy-preserving speech data publishing problem is as follows.

$$\underset{\boldsymbol{w}, \boldsymbol{x}^u, \boldsymbol{y}^u, \boldsymbol{z}^u}{\text{Minimize}} \quad U$$

$$\text{Subject to} \quad \begin{cases} P^u \leq p, \forall u \in \mathbb{U}, \\ U_v \leq \theta_v, \\ U_t^u \leq \theta_t^u, \forall u \in \mathbb{U}, \\ U_q^u \leq \theta_q^u, \forall u \in \mathbb{U}, \\ U_d \leq \theta_d, \\ \boldsymbol{x}^u \boldsymbol{y}^u = 0, \forall u \in \mathbb{U}. \end{cases} \tag{3}$$

The purpose of the thresholds $\theta_v, \theta_t^u, \theta_q^u, \theta_d$ is to ensure not to sacrifice a single utility too much. Since it is pointless to apply speech synthesis and voice conversion to a single utterance simultaneously (explained in §V), we add a constraint $\boldsymbol{x}^u \boldsymbol{y}^u = 0$, meaning that at least one of $\boldsymbol{x}^u, \boldsymbol{y}^u$ is set to 0. This optimization is not a convex problem. We will discuss how to simplify it and how to make personalized sanitization for every utterance in §VI .

B. Privacy Leak

As illustrative examples, we provide formulas for each privacy leak. They may be defined differently in other scenarios.

Text leak: P_t^u is defined as the sum of TF-IDF (term frequency–inverse document frequency) of terms in a text, which reflects the sensitivity level of each term. We will present more details in §IV.

Voice attribute leak: $P_{va}^u = \sum_{i=1}^{n_a} a_i$, where n_a is the number of attributes that can be learned from voice solely and the weight a_i reflects the sensitivity level of attribute i.

Voiceprint leak: $P_{vp}^u = p_{vp} b$, where p_{vp} is set by the publisher or the data consumer and represents the amount of loss caused by voiceprint leakage, and $b \in [0, 1]$ represents how much of the voiceprint is leaked. There is no explicit definition of voiceprint today. We can quantify b as the success rate of speaker authentication. Given a speaker authentication system well trained for a speaker with his/her original utterance, if the original voice is used to pass the system, the success rate $b = 1$. If we use perturbed voice to verify the speaker, b will be smaller. Similarly, if the attacker acquires the unperturbed voice, he can extract an 100% accurate voiceprint and use it to perform spoofing attacks. If the published voice is noisy, he can only extract a fraction of voiceprint. The parameter b aims to capture such a fraction.

Membership leak: $P_m = \sum_{j=1}^{n_a'} a_j$, where n_a' is the number of attributes that can be learned from des and the weight a_j reflects the sensitivity level of attribute j.

A possible formula of the total privacy leak for u is

$$P^u = P_t^u + P_{va}^u + P_{vp}^u + P_m. \tag{4}$$

We do not need to assign weights because they are absorbed into each term.

C. Utility Loss

Since it is much harder to define utility, we opt to quantify utility loss instead. Notice we denote utility *loss* as U hereafter.

Voice diversity loss U_v: Usually, speech data is published to train a recognition system. Voice diversity is important because a system trained on people with various voices is less likely to overfit a certain voice. Voice diversity is up to the diversity of speakers' attributes like gender, age, and region. Let P, Q be the joint distribution of these attributes in the original/perturbed dataset respectively, we define U_v as the distance between P, Q: $U_v = \frac{1}{2}||P - Q||_1$. We have $U_v \in [0, 1]$. Alternative distance metrics include Hellinger distance and JS divergence.

Text authenticity loss U_t^u: Sometimes the publisher may modify the speech content to remove sensitive parts of the audio, which causes loss of speech content authenticity. A possible measure for U_t^u is word-level edit distance. In our scenario, one substitution does not necessarily count as one error but r errors ($0 \leq r \leq 1$), where the value of r depends on the compatibility of the replacement word. If it is syntactically and semantically compatible with the original sentence, then r is very small. So for an utterance u, $U_t^u = (r \cdot s + d + i)/N$, where s, d, i are the number of insertions, deletions, and substitutions respectively, and N is the number of words in u. We have $U_t^u \in [0, 1]$.

Speech quality loss U_q^u: We hope the data consumer can still get the speech content clearly when we distort an utterance to hide the speaker's voiceprint. The Perceptual Evaluation of Speech Quality (PESQ) from the ITU standard P.862 [25] is considered the standard objective measurement for speech quality. It measures the similarity of the reference audio and the noisy audio. Thus, we calculate U_q^u as one minus PESQ scaled to $[0, 1]$. We also use scaled PESQ as a similarity measure hereafter. An alternative metric for U_q^u is the decrease of speech recognition accuracy on the noisy audio compared to that on the reference. We found by experiment that the two metrics' Pearson correlation coefficient is 0.73.

Data use clarity loss U_d: D is usable only when the publisher provides clear metadata and a necessary description des, so that the data consumer knows whether D meets his needs (or what tasks can be done on D). This is referred to as *data use clarity*. Suppose there is a speech dataset collected from a group of heart disease patients and to be shared to a third-party who intends to study how heart disease influences voice. If the publisher hides the fact that the speakers are heart disease patients for privacy purpose, the data consumer would not know if the dataset is suitable, so the data use clarity is 0. It is 1 if that is clarified in des. When D is published for n_u uses with importance weights w_k for $k = 1 \ldots n_u$, the overall clarity is quantified by the sum of the weights of the

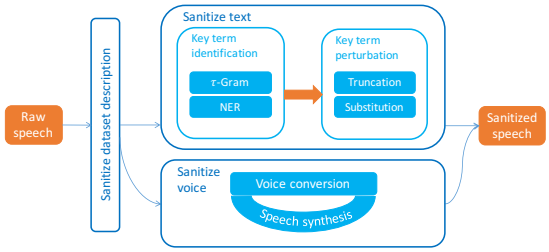


Fig. 3: **The workflow of speech data sanitization.**

clarified uses. The loss is $U_d = \sum_{k \in K} c_k$, where K is the set of uses that become unclear after sanitizing des. By setting $0 \leq c_k \leq 1, \sum_k c_k = 1$, we can scale U_d to $[0, 1]$.

As an example, we can define the overall utility loss as

$$U = \beta_v U_v + \beta_t U_t + \beta_q U_q + \beta_d U_d, \tag{5}$$

where $0 \leq \beta_v, \beta_t, \beta_q, \beta_d \leq 1$ and $\beta_v + \beta_t + \beta_q + \beta_d = 1$. The weights depend on the purpose of publishing and the data consumer's preference for each type of utility. U_t, U_q are averaged over all utterances, that is, $U_t = \overline{U_t^u}, U_q = \overline{U_q^u}$. As a result, we have $U \in [0, 1]$.

D. Possible Defenses

Basically, there are three categories of defenses, each of which may have several variations.

Sanitize dataset description, *i.e.* remove or modify any information in des that implies the speakers' attributes. This would reduce P_m and make the search range bigger. However, some information in des, like the settings where the utterances were recorded and the background of the speakers, may be critical to the data consumer. For instance, if the dataset is published for speech recognition on a specific dialect, the region or ethnicity of the speakers is inevitably disclosed. Thus, this method may harm the data use clarity. The publisher should sanitize des except those information that is essential to the data consumer.

Functions e, f (in Fig. 2). Let A_1 be the set of attributes about the speakers contained in des, let $A \subset A_1$ be the set of attributes that are removed/modified, and let $A' \subset A_1$ be the set of attributes that are critical to the data use clarity (suppose they have a bijective relation). The impacts of A on P_m, U_d are

$$P_m(A) = \sum_{j \in A_1 \setminus A} a_j, \quad U_d(A) = \sum_{j \in |A' \cap A|} c_j. \tag{6}$$

Sanitize text, *i.e.* modify the speech content to reduce the text leak P_t. However, there is a loss of text authenticity U_t. We will design a key term perturbation approach in §IV.

Sanitize voice, *i.e.* change the voice of every utterance to a different voice, so that the attacker is unable to infer some attributes or to steal the voiceprints. It can also thwart de-anonymization by voice when the attacker has the knowledge of the victim's voice. Voice change can be realized by speech synthesis or voice conversion. For the former, U_v would be damaged because it has limited choices for the output voice. For the latter, U_q will be damaged. (See details in §V.)

IV. Speech Content Sanitization

We adopt TF-IDF to detect terms in a text that might leak personal information. The observation is that if a person often uses a term that is not popular among others, it is usually highly related to her. The larger a term's TF-IDF is, the more private it is to her. Thus, we can define P_t as follows.

Definition 2 (Text leak): The amount of private information contained in the text tx of an utterance u, P_t^u, is the sum of scaled TF-IDF values of all terms tm in tx.

$$P_t^u = \sum_{tm \in tx} \text{tf-idf}(tx, tm). \tag{7}$$

We analyzed 8K Hillary Clinton emails and found most of the top terms are persons, locations, and organizations. Text sanitization on a given utterance is comprised of four steps: 1) identify key terms (defined later) in the transcript (can be obtained by speech recognition if not provided), 2) locate these terms in the audio using DTW based keyword spotting ([22]), 3) perturb the corresponding part in the audio, 4) perturb these terms in the transcript if it is to be published as well.

A. Key Term Identification

Based on how we define "term", there are two key term identification methods. Both have removed stop words at first.

τ-Gram based: This method only treats τ-grams as terms when identifying key terms and calculating P_t^u. A τ-gram refers to a contiguous sequence of τ words in a text. A text usually contains very few terms with large TF-IDF values (see the histogram in Fig. 4(a)). For the purpose of sanitizing text and preserving utility, we ought to reduce P_t^u by perturbing the top key terms first. Specifically, we select terms whose TF-IDF is greater than a threshold δ (referred to as *key terms*).

NER based: Named-entity recognition (NER), a natural language processing technology, aims to locate and classify named-entities in text into predefined categories. We utilize an LSTM-based NER [9] to extract the names of persons, locations, and organizations from the text as the terms. These names usually disclose the speaker's privacy, *e.g.* social connections, geographical region, occupation, and salary. Likewise, we select key terms by their TF-IDF values.

Comparison: The difference of the two definitions above lies in whether non-named-entities should be considered private information, For example, words not popular among the public but frequently used by a particular person might also reveal his/her attributes, *e.g.* terminologies in a field and slang in a small region. The first method counts them in P_t whereas NER based method does not. Besides, not all named-entities are highly related to the speaker. For instance, a speaker's quoting a celebrity or mentioning a foreign country or a famous organization does not imply she is that celebrity or she is in that country/organization. However, τ-gram based method is not perfect either. It is unable to identify a phrase as a whole that represents a single entity. For instance, it separates "David" from "David Koch", "Tea" from "Tea Party", and "Hill" from "Oak Hill". The separated words alone convey little information, plus perturbing them alone makes no sense.

Another limitation is that it cannot identify phrases longer than τ if only $\leq \tau$-grams are taken into account. A large τ will make the identification computationally expensive.

B. Key Term Perturbation

A way to perturb key terms is *substitution*, *i.e.* to replace them with other terms of the same category. If s terms are substituted, $U_t^u = rs/N$. If the substitution terms fit the context well, r is close to 0. However, the weakness of this method is that the publisher needs to build a categorized set of replacement words in advance. Another method is *truncation*, which is a special case of substitution with $r = 1$. It is easier to implement but results in more utility loss. Both methods have the same influence on the text leak because P_t^u is determined by δ. We will study the impact of δ on P_t^u, U_t^u in the experiment (Fig. 5). Besides, substitution may have a mild impact on audio quality if too many terms are replaced.

V. Voice Sanitization

Voice sanitization can be achieved by many approaches, but the most feasible ones are voice conversion and speech synthesis. The former modifies the voice of an utterance whereas the latter synthesizes fake voice according to the text.

A. Voice Conversion

Targeted voice conversion aims to make an utterance sound like a specific speaker, which requires parallel speech corpora for training and produces audio with undesirable quality [16], so it is inapplicable. Non-targeted voice conversion converts an utterance to an arbitrarily different voice. Generally, it produces audio with higher-quality than targeted voice conversion does. A popular non-targeted voice conversion paradigm is VTLN (vocal tract length normalization) based frequency warping. It consists of 6 steps: pitch marking, frame segmentation, FFT, VTLN, IFFT, and PSOLA (Pitch Synchronous Overlap and Add) [29]. VTLN distorts the voice by deforming the frequency axis of the signal according to a warping function. In this work, we adopt the commonly used bilinear function [29] as the warping function. Its formula is

$$f' = BI(f, \alpha) = \left| -i \frac{e^{\frac{i\pi f}{f_m}} - \alpha}{1 - \alpha e^{\frac{i\pi f}{f_m}}} \right| \cdot \frac{f_m}{\pi}, \tag{8}$$

where f, f', f_m are the original/new/maximum frequency, i is the imaginary unit, and α is a parameter tuning the distortion level. For this method, α is the parameter $\boldsymbol{y}$ in Fig. 2. Its influence on P_{vp}^u, U_q^u will be derived in the experiment (§VII-C). Its influence on U_v will be discussed in §V-B.

Function i^u. For illustration purpose, we focus on gender only and study the influence of α on the perceived gender. We classify gender by the mean pitch $\bar{f}$ of the speaker's utterance. The speaker is perceived as female if $\bar{f}$ is greater than the threshold θ_f (will be discussed later). Voice attribute is considered leaked if the perturbed voice and the original voice have the same perceived gender. Let a_g be the weight assigned for gender and $\mathbf{1}$ be the Heaviside step function, then

$$P_{va}^u(\alpha) = a_g \cdot \mathbf{1}((\bar{f}^u - \theta_f)(BI(\bar{f}^u, \alpha) - \theta_f)). \tag{9}$$

B. Speech Synthesis

Speech synthesis is also known as text-to-speech. It is composed of 3 major steps: tokenization, text-to-phoneme conversion, and waveform generation. Though a few speech synthesis systems like iSpeech and Microsoft Bing Speech can produce pretty natural voices, the voice diversity is very limited and the synthetic voice has invariant emotion and intonation contour. Speech synthesis decreases P_{va}, P_{vp} to 0 and preserves text authenticity and speech quality, so it is the best option when β_v is very small.

Function k. As an example, we suppose there is only one synthetic female voice to choose. The parameter of speech synthesis is $o^u \in \{1, 0\}$, where 0 represents keeping the speaker's original voice, and 1 represents replacing it with synthesized voice. The voice diversity loss depends on how α^u, o^u are chosen for each u in D. We take gender only as an example to study the joint influence of $\boldsymbol{\alpha}, \boldsymbol{o}$ on U_v. Recall that each u takes at most one action of voice conversion and speech synthesis ($\alpha^u o^u = 0$), so we can study their influence separately. For voice conversion, some speakers' perceived gender is changed from female to male, others from male to female. The number decrease of male-voiced utterances caused by voice conversion equals the male-to-female number minus the female-to-male number, *i.e.*,

$$\Delta n_1 = \sum_{u \in \mathbb{U}} \frac{1}{2} (\text{sgn}(BI(\bar{f}^u, \alpha^u) - \theta_f) - \text{sgn}(\bar{f}^u - \theta_f)).$$

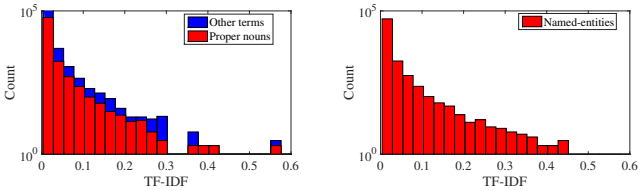
Let s^u be the gender of speaker u ($s^u = 1$ for male, 0 for female) and n_m, n_f be the number of male/female speakers in D. Then, the number decrease of male-voiced utterances caused by speech synthesis is $\Delta n_2 = \sum_u s^u o^u$. The total number decrease is $\Delta n = \Delta n_1 + \Delta n_2$. The original gender distribution is $P = (\frac{n_m}{n}, \frac{n_f}{n})$ and the new distribution is $Q = (\frac{n_m - \Delta n}{n}, \frac{n_f + \Delta n}{n})$. By the definition of U_v,

$$U_v(\boldsymbol{\alpha}, \boldsymbol{o}) = \frac{|\Delta n|}{n}. \tag{10}$$

VI. Heuristic Algorithm

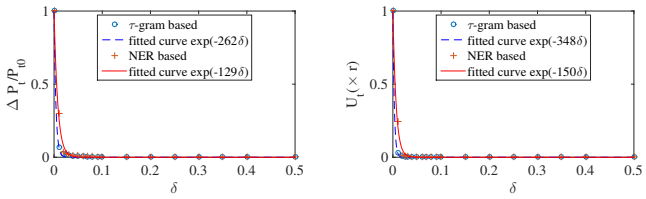
The optimization problem in Eq. 3 is not a convex problem, so we design a heuristic approach to simplify it. According to Fig. 2, we can group the actions into three sets – des sanitization, key term perturbation, and voice sanitization (voice conversion & speech synthesis). The sets are pairwise independent as they have independent influences on different aspects of privacy leak and utility loss. Suppose P^u, U are defined as in Eq. 4, 5. Given the total privacy leak budget p, we assign it to $P_t^u, P_{va}^u + P_{vp}^u, P_m$ in the ratio $\beta_t, \beta_v + \beta_q, \beta_d$. The intuition is that if the weight of a certain utility is large, it is important to preserve the utility by performing fewer actions that damage it, so the corresponding privacy leak budget should be loosened and set to a large value. Therefore, we split the optimization problem into three smaller ones:

- **P1:** For all $u \in \mathbb{U}$, minimize U_t^u, subject to $P_t^u < p\beta_t$.
- **P2:** Minimize $\beta_v U_v + \beta_q U_q$, subject to $P_{va}^u + P_{vp}^u < p(\beta_v + \beta_q), \boldsymbol{x}^u \boldsymbol{y}^u = 0, \forall u \in \mathbb{U}$.
- **P3:** Minimize U_d, subject to $P_m < p\beta_d$.



(a) τ-gram based (b) NER based

Fig. 4: **Histograms of TF-IDF values (Hillary emails).** They are on a log scale and they display a power-law like distribution.



(a) $\Delta P_t/P_{t0}$ - δ (b) U_t - δ

Fig. 5: **Impact of δ on text leak decrease ΔP_t and text utility decrease U_t (Hillary emails).** In (a), P_{t0} is the original text leak.

The **algorithms** for the three problems are as follows.

P1: We can easily solve P1 by Eq. 11, 12 and calculate the optimal δ and minimum U_t^u.

P2: We design a greedy algorithm for the simple case where only gender is taken into account for voice attributes and voice diversity. Recall we have two actions – speech synthesis and voice conversion– and two parameters o^u, α^u for each u. The constraint $o^u\alpha^u = 0$ requires us to select at most one action for u. Recall the formulas in Eq. 10, 13, 14. We have the budget $p_b = p(\beta_v + \beta_q)$. The algorithm selects o^u, α^u for each u as follows:

1) Calculate initial privacy leak $P = P_{va}^u + P_{vp}^u = a_g + p_{vp}$.
2) If $P \leq p_b$, return.
3) Initialize $o^u = \alpha^u = 0$.
4) If $a_g > p_b$ or u is female, set $o^u = 1$, return.
5) Get α by solving $P_{va}^u(\alpha) + P_{vp}^u(\alpha) = p_b$.
6) If $\beta_q \frac{1}{n} U_q^u(\alpha) \leq \beta_v \frac{1}{n}$, set $\alpha^u = -|\alpha|$, else set $o^u = 1$.
7) Return.

This algorithm guarantees optimum.

P3: We can greedily approximate P3. Suppose des contains n_a' attributes about the speakers, we arrange the attributes that are unimportant to data use clarity on the top, and those important at the bottom, which are further sorted by their weights a_j in descending order. Given the sorted list, we remove attributes from top to bottom until the sum of the rest ones' weights is lower than $p\beta_d$.

VII. Experimental Evaluations

Dataset: 1) *TED talks.* We downloaded the mp3 audio and transcripts (as .txt files) of 800 TED talks from ted.com. We removed talks that are too short and get audios of 562 speakers. We truncated applause, silence, and the speaker introduction at the beginning/end of the audios. 2) *LibriSpeech* [18]. The corpus contains 100 hours utterances spoken by 251 native speakers of American English. 3) *Hillary Clinton emails.* The dataset includes nearly 8K emails. 4) *US Census Data* [13]. It contains 2.46M people's demographics (68 categorical attributes).

A. Simulating Linkage Attacks

At first, we show it is possible for the attacker to greatly reduce the search range by filtering the candidates with several attributes he knows about the speaker. For Census, the search range can be cut out from 2.46M people to 1000 on average if six attributes are known. In the best case, the search space can be reduced to 40 when only one attribute is known. For TED talks, by using gender, age bracket, and ethnicity, we can group speakers into 35 sets with minimum size 1, max size 194, and average size 16. We see that it is hard but still possible to identify the victim if the attacker knows several attributes. If the attacker has knowledge of the victim's voice beforehand, he identifies her among the reduced search space by matching the voice, which is supported by speaker recognition. We utilized Bing Speaker Recognition API to perform the task on a dataset with over 800 speakers, obtained by combining TED talks and LibriSpeech. From each person's speech, we take a one minute sample for training and another for identification. The speakers of the testing utterances can be identified among 800 with a 100% success rate, which is surprisingly accurate. This simulation reveals the potential risk of linkage attacks on unprotected speech data.

B. Sanitizing Text

Key term identification: We implement and evaluate both the τ-gram based and NER based methods. Fig. 4 plots the histograms of TF-IDF values of the terms in every utterance. Their TF-IDF values display power-law distributions, *i.e.* the majority of terms have very small TF-IDF. The TF-IDF is scaled to $[0, 1]$ by dividing the max value. Very few terms have values greater than 0.6. If the count is 1, it cannot be displayed on the log scale, so we cut off the right part of the histograms where TF-IDF$>$ 6. In Fig. 4(a), we also present the portion of terms that are tagged by NER as proper nouns, though they are not necessarily complete named-entities.

Key term perturbation: By setting δ and removing the terms whose TF-IDF exceeds δ, we can greatly reduce the text leak while inducing the least modification on the text content. The impact of δ on P_t^u and U_t^u is depicted in Fig. 5. For P_t^u, the text length is a factor to be considered. Generally, a long passage contains more terms and thus has higher P_t^u than a short sentence. We assume their terms' TF-IDF values conform to the same distribution. To bypass the influence of text length, we opt to study the relation between δ and the ratio of text leak decrease $\Delta P_t^u/P_{t0}^u$. For U_t^u, we do not have this hassle because it is already normalized. U_t^u is up to the number of perturbed terms and the perturbation method (reflected by the ratio r). Fig. 5 shows the curves fitted for both key term identification methods.

Utility weights				Privacy leak budgets				Utility losses					Qualification rate
β_t	β_v	β_q	β_d	p	P_t^u	$P_{va}^u + P_{vp}^u$	P_m	$U_t(\times r)$	U_v	U_q	U_d	U	$\lvert\{u \mid P^u \le p\}\rvert/n$
0	0.333	0.333	0.333	60	0	40	20	1	0	0	0	0	100%
0.333	0	0.333	0.333	60	20	20	20	0.001	0.722	0	0	0 ∼3.33e-4	99.37%
0.333	0.333	0	0.333	60	20	20	20	0.001	0	0.579	0	0 ∼3.33e-4	99.37%
0.333	0.333	0.333	0	60	20	40	0	0.001	0	0	1	0 ∼3.33e-4	99.37%
0.25	0.25	0.25	0.25	60	15	30	15	0.008	0	0.494	0	0.124 ∼ 0.125	91.52%
0.25	0.25	0.25	0.25	50	12.5	25	12.5	0.024	0	0.545	0.5	0.261 ∼ 0.267	84.57%
0.25	0.25	0.25	0.25	40	10	20	10	0.071	0	0.579	0.5	0.270 ∼ 0.288	79.44%
0.25	0.25	0.25	0.25	30	7.5	15	7.5	0.209	0.722	0	0.5	0.306 ∼ 0.358	74.77%

TABLE I: **Utility loss caused by sanitization under various $\boldsymbol{\beta}, p$ settings (TED talks).** U increases with p getting smaller.

Functions g^u, h^u. The concrete formulas of these function are dataset-dependent. For Hillary emails, by Fig. 5, if we choose τ-gram based method, then the formulas are

$$P_t^u(\delta) = P_{t0}^u(1 - \Delta P_t^u / P_{t0}^u) = P_{t0}^u(1 - \exp(-262\delta)). \quad (11)$$

$$U_t^u(\delta) = r \cdot \exp(-348\delta). \quad (12)$$

C. Sanitizing Voice

Voice conversion: We study the impact of the VTLN-based voice conversion on privacy and utility by varying the parameter α. In Fig. 6 are the fitted curves that measure how α changes the speaker authentication success rate (denoted by sr) and the PESQ of the output voice. It shows they are both monotone descending when $|\alpha|$ increases. The impact on P_{vp}^u and U_q^u is derived as follows.

Functions j, l. According to Fig. 6, we have

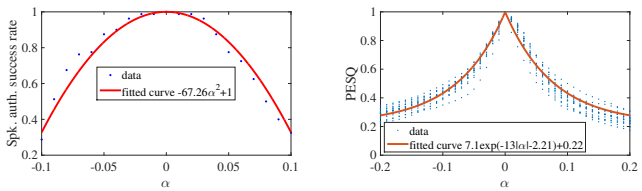
$$P_{vp}^u(\alpha) = p_{vp} \cdot sr = (-67.26\alpha^2 + 1)p_{vp}, \quad (13)$$

$$U_q^u(\alpha) = 1 - \text{PESQ} = 1 - 7.1exp(-13|\alpha| - 2.21 + 0.22). \quad (14)$$

Speech synthesis: We use Bing Text to Speech API to synthesize speeches with given transcripts and use Google Speech Recognition API to transcribe them. The accuracy of speech recognition on the real speech and synthesized speech is 81.1% and 85.9% respectively. In one sense, this suggests speech synthesis does not damage speech quality.

D. Solving Optimization Problem

We evaluate the heuristic algorithm on TED talks to make personalized sanitization for each speaker. Initially, mean text leak $\overline{P_t^u} = 10$. 97.4% talks have $P_t^u \le 20$, so we set $P_{vp}^u = p_{vp} = 20$, $P_{va}^u = a_g = 20$, $P_m = 20$ to keep the setting neat here. Our approach also applies to other settings. Then, the initial privacy leak $P^u = 70$ on average. The dataset we crawled does not have a description, so we synthesize a description with 3 attributes, *e.g.* $des = \{att_1, att_2, att_3\}$, with weights $\boldsymbol{a} = (5, 10, 5)$ (so $P_m = 20$). Each of att_2, att_3 is necessary for the clarity of a data use. The importance of the two data uses is set to $\boldsymbol{c} = (0.5, 0.5)$. There are 406 males and 156 females in TED talks. The averaged mean pitch we calculated is 144 Hz for males and 202 Hz for females, so we set the threshold $\theta_f = 173$ Hz to detect gender by voice. We ignore the constraint on each type of utility loss in Eq. 3 here, though publishers may have these constraints in reality.



(a) Spk. auth. sr. - α (b) PESQ - α

Fig. 6: **Impact of voice conversion on speaker authentication success rate and speech quality (LibriSpeech).**

Our algorithm successfully limits $P_{va}^u + P_{vp}^u + P_m$ within given budgets for every u, and it limits the P_t^u of 74.77% ∼ 99.37% of the utterances within budget if budget for P_t^u is $7.5 \sim 20$. For Hillary emails, the percentage is 98.65% ∼ 99.95%, which is much better. The reason is that TED talks has a much smaller corpus for us to learn an accurate formula for Eq. 11. Tab. I displays the consequent utility loss caused by sanitization under various settings of utility weights $\boldsymbol{\beta}$ and budget p. We refer to the fraction of utterances satisfying p-leak limit as *qualification ratio*. It can be seen that we induce only minimal utility loss while guarantee p-leak limit for most utterances (qualification ratio $\ge 74.77\%$). The outliers will be sanitized by further reducing δ to meet the budget.

Run time: The cost of action/parameter selection and dataset description sanitization is negligible. The time of speech content sanitization is $0.60ls$ seconds per utterance, where l is the audio length, and s is the number of terms perturbed (keyword spotting is the most expensive step). For voice sanitization, the cost is $0.42l$ and $0.20l$ seconds per utterance for voice conversion and speech synthesis respectively.

VIII. Related Work

Speaker recognition and voiceprint extraction. The research on speaker recognition dates back to 1937 [15]. In 2000, Reynolds *et al.* [26] introduced adapted Gaussian mixture models (GMMs) to represent each speaker's voice characteristics. Then Dehak *et al.* [3] utilized joint factor analysis to calculate an i-vector for each speaker and compare two speakers' identity according to the cosine similarity of their i-vectors. Variani *et al.* [30] trained a deep neural network (DNN) to classify speakers at the utterance level, then used the activations of the last hidden layer of the DNN as the d-vector for each speaker. It demonstrated that d-vector outperforms i-vector in representing voiceprint in some cases. The d-vector model has been followed by plenty of researchers who tried other neural networks including convolutional time-delay DNN [11], ResCNN and GRU [10].

Privacy in various data. In the past decades, researchers have thoroughly studied privacy issues in data of various forms, including relational data [5], [8], graph data [23], [24], [32], text [7], images [33], [34], location data [12], and sensor data [14]. There are, however, only several papers on privacy issues in speech/voice data. Smaragdis *et al.* [28] were the first to design a secure speech recognition scheme with secure multi-party computation. Pathak *et al.* [20], [21] utilized secure multi-party computation to achieve speaker verification while protecting the speaker's voice.

IX. Conclusion

This work is the first attempt to identify the privacy risk of speech data sharing and to come up with countermeasures to protect the speakers' privacy. Firstly, we discussed how the attacker compromises their privacy by performing linkage attacks. Then, we formulated the privacy-preserving speech data publishing problem, instantiated it with our quantifications of privacy leak and utility loss, each in four aspects, and proposed a heuristic algorithm. We studied the effects of a few possible sanitization approaches including voice conversion and speech synthesis, and designed a TF-IDF based key term perturbation approach to desensitize the speech content. The experiments on real datasets validate the privacy leak risk and the effectiveness of our sanitization approaches and heuristic algorithm. Potential future work includes a better privacy definition and quantification, a better problem formulation and approximation, and more defensive measures.

Acknowledgment

Xiang-Yang Li is the contact author. The work is partially supported by China National Funds for Distinguished Young Scientists with No. 61625205, Key Research Program of Frontier Sciences, CAS, No. QYZDY-SSW-JSC002, NSFC with No. 61520106007 and No. 61572347, NSF ECCS 1247944, NSF CNS 1526638, NSF CNS 1343355.

References

[1] Apple admitted Siri voice data sharing, https://goo.gl/rRHj4r.

[2] Samsung admitted voice data sharing, https://goo.gl/bQPUDj.

[3] 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):788–798, 2011.

[4] D. Gillick. Can conversational word usage be used to predict speaker demographics?. In *INTERSPEECH*, pages 1381–1384. Citeseer, 2010.

[5] J. Hou, X.-Y. Li, T. Jung, Y. Wang, and D. Zheng. CASTLE: Enhancing the utility of inequality query auditing without denial threats. *TIFS*, 2018.

[6] S. Johar. Psychology of voice. In *Emotion, Affect and Personality in Speech*, pages 9–15. Springer, 2016.

[7] T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su. Accounttrade: Accountable protocols for big data trading against dishonest consumers. In *INFOCOM*, pages 1–9. IEEE, 2017.

[8] T. Jung, X.-Y. Li, and M. Wan. Collusion-tolerable privacy-preserving sum and product calculation without secure channel. *IEEE Transactions on Dependable and secure computing*, 12(1):45–57, 2015.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[10] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[11] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang. Deep speaker feature learning for text-independent speaker verification. *arXiv preprint arXiv:1705.03670*, 2017.

[12] X.-Y. Li and T. Jung. Search me if you can: privacy-preserving location query service. In *INFOCOM, 2013 Proceedings IEEE*, pages 2760–2768. IEEE, 2013.

[13] M. Lichman. UCI machine learning repository, 2013.

[14] H. Liu, X.-Y. Li, L. Zhang, Y. Xie, Z. Wu, Q. Dai, G. Chen, and C. Wan. Finding the stars in the fireworks: Deep understanding of motion sensor fingerprint. In *INFOCOM*. IEEE, 2018.

[15] F. McGehee. The reliability of the identification of the human voice. *The Journal of General Psychology*, 17(2):249–271, 1937.

[16] S. H. Mohammadi and A. Kain. An overview of voice conversion systems. *Speech Communication*, 2017.

[17] I. Mporas and T. Ganchev. Estimation of unknown speaker's height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE, 2015.

[19] P. Patel, A. Chaudhari, R. Kale, and M. Pund. Emotion recognition from speech via boosted gaussian mixture models. *International Journal of Research In Science & Engineering*, 3, 2017.

[20] M. Pathak, J. Portelo, B. Raj, and I. Trancoso. Privacy-preserving speaker authentication. In *International Conference on Information Security*, pages 1–22. Springer, 2012.

[21] 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):397–406, 2013.

[22] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng. Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*, 2017.

[23] J. Qian, X.-Y. Li, C. Zhang, and L. Chen. De-anonymizing social networks and inferring private attributes using knowledge graphs. In *INFOCOM*, pages 1–9. IEEE, 2016.

[24] J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han. Social network de-anonymization and privacy inference with knowledge graph model. *TDSC*, 2017.

[25] P. Recommendation. 862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Feb*, 14:14–0, 2001.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[27] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The INTERSPEECH 2011 speaker state challenge. In *INTERSPEECH*, pages 3201–3204, 2011.

[28] P. Smaragdis and M. Shashanka. A framework for secure speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1404–1413, 2007.

[29] D. Sundermann and H. Ney. VTLN-based voice conversion. In *ISSPIT*, pages 556–559. IEEE, 2003.

[30] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP*, pages 4052–4056. IEEE, 2014.

[31] Z. Wu and H. Li. Voice conversion versus speaker verification: an overview. *APSIPA Transactions on Signal and Information Processing*, 3:e17, 2014.

[32] S. Xue, L. Zhang, A. Li, X.-Y. Li, C. Ruan, and W. Huang. Appdna: App behavior profiling via graph-based deep learning. In *INFOCOM*. IEEE, 2018.

[33] L. Zhang, T. Jung, K. Liu, X.-Y. Li, X. Ding, J. Gu, and Y. Liu. Pic: Enable large-scale privacy preserving content-based image search on cloud. *TPDS*, 28(11):3258–3271, 2017.

[34] L. Zhang, K. Liu, X.-Y. Li, C. Liu, X. Ding, and Y. Liu. Privacy-friendly photo capturing and sharing system. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 524–534. ACM, 2016.

[35] H. Zhao, Z. Yang, Z. Chen, and X. Zhang. Automatic chinese personality recognition based on prosodic features. In *International Conference on Multimedia Modeling*, pages 180–190. Springer, Cham, 2015.