Improving Data Privacy in Voice-Driven Technologies with Machine Learning

Dr. K.Phalguna Rao
Associate Professor
Malla Reddy University, Hyderabad.

ABSTRACT: In speech data publishing, users' data privacy is disclosed and thereby more privacy of users is breached since speech data contains a large amount of information about speakers. Existing work focused on sanitization in speech content, speakers' voice, and data descriptions, without considering the correlation of speech content and speaker's voice. Therefore, this existing work cannot protect speakers' data privacy when attackers utilize such correlation to identify speakers' speech data. To tackle this problem, in this article, we propose a protocol to decrease such potential risks in speech data publishing while keeping the balance of privacy preservation and data utility. Specifically, we define both the risks of privacy disclosure and the data utility loss in speech content, speaker's voice, and data set description. Moreover, we do the first attempt to formalize the correlation between speech content and speaker's voice and regard it as a new kind of privacy leakage risk. Thereafter, we utilize the classifier in machine learning and optimize speech data sanitization considering the defined risks of privacy disclosure and data utility loss. Finally, simulation results validate the effectiveness of the proposed protocol.

KEY WORDS : Big Data, Artificial Intelligence, Machine Learning, Privacy.

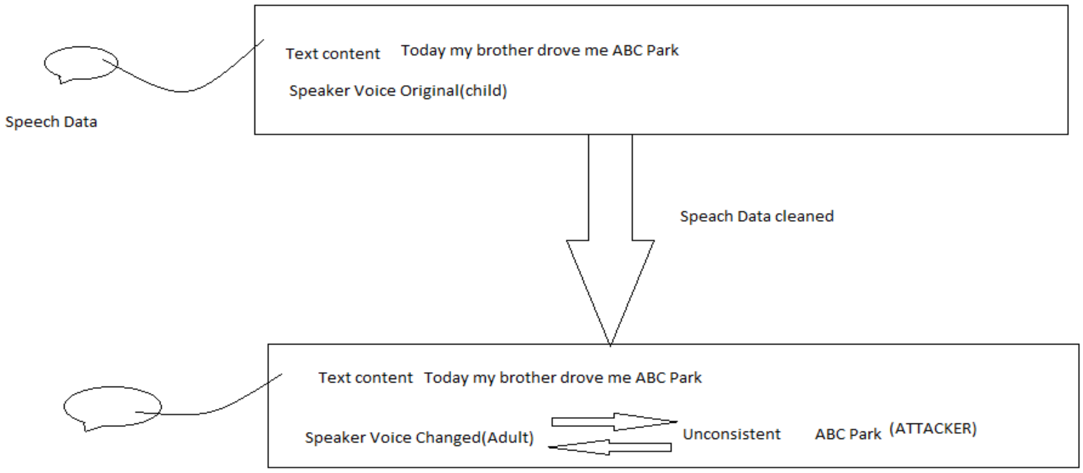
I. INTRODUCTION

The of artificial intelligence (AI) and machine learning (ML) across billions of devices has reinvented the way societies around the world learn, work and function. Voice-based services and technologies like Siri and Google Assistant continue to create seamless and hands-free experiences that transform our lives. However, in this ever-evolving digital world, the privacy obligations and security issues associated with speech data cannot be overlooked, and a team in China has introduced a new system that enhances privacy protection in speech data publishing.

II. METHODOLOGY

While major companies aim to delete all personally identifiable information (PII) of users' speech data, such as their name, phone number and email when improving voice-based services, cyber-attacks continue to increase. Partial anonymization is not enough. Current systems fail to consider the correlation between speech content – the text or words in the speech – and a speaker's voice, making it easier for attackers to filter out the sanitized – or altered – parts of speech data and identify sensitive information. An example of such a scenario is illustrated in Figure 1.

Figure1.
Speech content is not reliable with speaker’s voice once cleaned.


Speech Data
Text content
Today my brother drove me ABC Park
Speaker Voice Original(child)
Speach Data cleaned
Text content Today my brother drove me ABC Park
Speaker Voice Changed(Adult)
Unconsistent
ABC Park
(ATTACKER)

Companies or institutions can attempt to sanitize the data by altering a given voice from a child to an adult, in order to mask the speaker’s age, but the content itself remains the same. This allows attackers to spot inconsistencies in the conversation and narrow down the speaker’s actual age. Additionally, in today’s digital landscape, anonymous data is often not significantly anonymous. Companies and organizations that believe in anonymity can be susceptible to linkage attacks, whereby attackers attempt to de-anonymize data sets.As outlined in Figure 2, adversaries collect secondary information about an individual from multiple data sources and then comb through that data to form a full profile. The more a company attempts to preserve the analytical utility of the dataset, such as keeping “gender” and “date of birth” information in the dataset, the more they are prone to linkage attacks, even if they remove the PII from the data. Attackers could cross-reference public records with attributes obtained from analyzing speech data to match and determine the identity of the PII-removed individual.

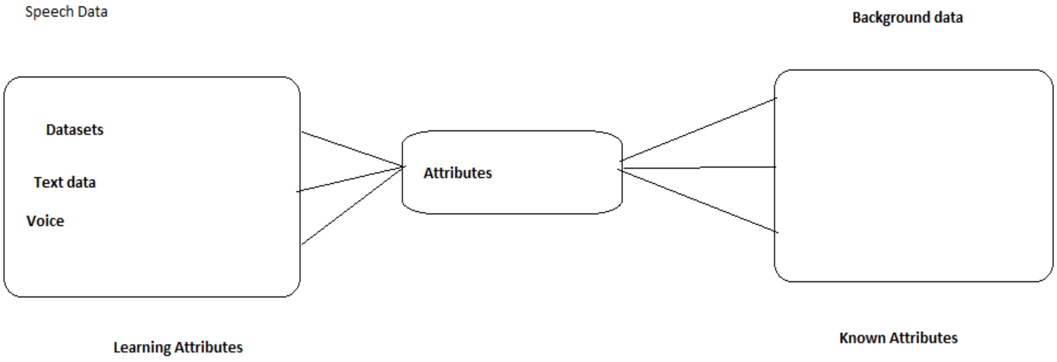


Figure 2: The process of Linkage attackes

It’s challenging to protect user information when publishing speech data. This data the last decade. There is trade-off between preventing privacy and minimize utility loss of speech data.

REFERENCES

1. R. Kotecha and S. Garg, "Preserving output-privacy in data stream classification", *Progress in AI*, vol. 6, no. 2, pp. 87-104, 2017.
2. IEEE Digital library.