Hiding Sensitive Information in Desensitized Voice Sequences*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname, 2nd Given Name Surname, 3rd Given Name Surname
dept. name of organization (of Aff.), City, Country
dept. name of organization (of Aff.), City, Country if needed
email address or ORCID of corresponding author(s)

Abstract—**Voice data is acquired and utilized by a variety of consumer services, voice assistants and voice-enabled mobile games for example. During the process of such services, raw recordings of user voices is, despite the presence of offline applications, mostly believed to be sent to web servers associated with them. This imposes a significant privacy problem regarding to the nature of raw voices, which contain style and content information that is believed to be sufficient to identify the speaker or to reveal sensitive data related to the speaker. In order to mitigate such problem, Desensitization methods are employed as an intermediary between user and the cloud services. However, these methods may prove not to be secure enough. This paper discusses and demonstrates a possibility of hiding potentially sensitive information of user in seemingly desensitized voices that may be used for online services with deep learning technologies, utilizing desensitization models with disguised stenography functions.**

Index Terms—**privacy, voice, desensitization, stenography**

I. Introduction

Voice is already an important mean of human-machine interaction since the last century. Voice recognition technologies enable human to communicate with machines with voice, the medium of human-human interaction. Voice identification technologies enables computers to perform authorization solely with certain features in human voices without involving passwords. In recent years deep learning drastically advanced thus it's now being employed by a great portion of voice recognition services. These services include voice search such as Google App and Taobao and input methods such as Gboard and iFly Chinese Pinyin IME.

These recognition services consume large amount of computational power which makes it not feasible to perform the process locally on end user's device. In order to mitigate such problem, service providers resort to processing voice data in their servers. This imposes a serious privacy concern since the raw recording from user is known to contain much more data than merely for content recognition. Such additional data may be used by malicious service providers or attackers to identify the speaker and cause a data breach incident.

It's commonly believed that the solution to mitigate such problem is two fold: One can choose to perform all processing solely on user's device or to use certain algorithm to remove user's identity from voices before submitting to remote servers. Due to the limitations of capabilities and energy consumption, it's often likely not viable to run a large model on smartphones, laptops or other battery-powered devices. Because of that, voice desensitization algorithms are being actively developed.

Conventional voice desensitization methods are believed to focus on two radically different areas: Content and Style. Those which focus on content desensitization employ certain methodologies to remove or replace voice segments that are detected to potentially contain sensitive information while others apply obfuscation and distortion on entire voices to decrease the possibility of associating certain features in voices with corresponding speakers. The detection of sensitive content is often handled by one or more deep learning based classification algorithms and the content wiping is also likely to be handled by a deep learning based transformation algorithm. Deep learning based methodologies are also employed to distort the voice features despite the presence of a variety of traditional algorithms. It's notable that many approaches to bring privacy to remotely handled voices combine these 2 ideas to provide better performance.

While evaluation of sensitive content removal is straightforward since researchers are enabled to use Text-To-Speech services with voices and compare the results with sources, the analysis of voice feature is of significantly more complexity. Obfuscation algorithms designed to disable certain recognition algorithms may not disable others and it's possible those algorithms would not confuse humans if the modification is not drastic.

However, the absence of sensitive data in desensitized voices can not be perfectly guaranteed. It's not only because of the incompleteness and weakness of desensitization models, but also because of the potential existence of an arbitrary algorithm that attempts to hide the data to recover the original voices or even the raw recording in desensitized voices by performing a slight modification. Modern deep-learning based stenography methodologies is more likely to get over a variety of stenography detection algorithms [1], thus make the protection against such attacks more difficult.

In this paper, we propose a novel, adversarial utilization of various types of voice transformer framework that desensitize user's voices as a desensitization model, but also attempts to hide information of original voice with certain stenography

Identify applicable funding agency here. If none, delete this.

technologies for another model to recovery the original one from it after the usage. The rest of the paper consists an detailed explanation of the proposed methodology, including the discussion of potential use cases.

II. Methodology

Our adversary framework works as an add-on to a benign (non-adversarial) voice desensitization model, and consists of three models: One benign model that handles conventional desensitization tasks, one stenography model that embeds the information of original voice into desensitized one generated by the prior model and one extraction model that takes in the product of the previous model and recovers the original voice, thus fulfills the adversarial purpose.

A. Benign model

As two slightly different scenarios, the benign model could be either a in-house solution designed by the attacker itself (and may be even trained with adversarial models) or a production model with predefined weights acquired from a third-party organization or the public. The source of benign model shows no significant difference in terms of the overall structure of our model, but the desensitization performance of different models may have an impact on the quality of stenography-added voices and recovered ones. For our experiments we used an StarGAN-VC [3] based model, the StarGanVCDialectConversion project hosted by user ”Didnelpsun” on GitHub [2] in particular.

Specifically, this specific StarGAN-VC implementation employs a transformer-based, multi-layered architecture for its generation model, which closely resemble the approach mention in the original StarGAN-VC publication. There are 5 sets of layers in both the encoder and decoder with each set of layer consisting of two convolution layers and two instance normalization layers. It's notable that this particular model is not strictly designed for desensitization, but we consider the fact that this model is capable to erase the original identity from the voice sequence and give it another one to be the ability to desensitize.

B. Stenography model

The stenography model works as an intermediate between the user and the benign model to embed information of original voice into clean product of the benign model. In order to achieve such a task to hide sufficient information in desensitized ones while minimizing the differences made, we employed the Hide and Speak model [4], an voice-centralized stenography model that could be considered as state of the art as it handles different lengths of carrier data and target data and produce high quality results.

It's clear that, users of this adversarial product are not intended to notice the presence of such stenography model that attempts to tamper the clean output. In order to achieve such stealthiness, the stenography model should contain following characteristics:

1) Integration with benign model: To prevent users from discovering the stenography logic, one can attempt to package the entire product into a black box that users get convincing results, hoping they have no questions about potential security risks from a non-transparent process. However, such actions may rise suspicion among users, which may not be desired in some cases. Another robust solution to this problem is to embed the stenography model into the victim, as an additional set of layers for ”post-processing”, thus integrated with it. In this way, the presence of stenography model could no longer be easily detected without dedicated research of behaviour or, even, source code.

2) Sufficient performance: The benign desensitization model attempts to replace a certain amount of features of original voices with features that does not belong to the owner of them while the stenography model attempts to embed a certain amount of information of original features into the already-desensitized voices. It's very likely that during this process the textual content of the voice or the overall audio quality will experience degradation to a degree. For this reason, the stenography model chosen for this purpose should be high quality enough to avoid excessively modifying the voice output.

C. Extraction model

The extraction model accepts the final product from the combination of benign and stenography model, which is likely to be the voices user trusted to be desensitized, and attempts recovery of original voice from it. As a model trained in conjunction with the stenography model, the one we employed in our experiment is also from the Hide and Speak mentioned before.

III. Experiment

A. Setup

As mentioned before, we use StarGANVCDialectConversion, a StarGAN-VC implementation as the benign model to produce desensitized voices, Hide and Speak for stenography and extraction process. All our experiments are performed on a x86-64 based container-enabled Linux server with a NVIDIA Tesla V100 GPU. We used software packages required by each models, but with Python 1.10 and PyTorch 1.13, which are slightly newer.

For convenience of further processes, we used voice samples from TM1, TM2, SF1 and SF2, speakers in the VCC2016 data set, as preferred by the configuration of benign model. The voices of TM1 is considered the original voices that contain sensitive information and ones of SF1, on the other hand, is considered desensitized. With this setup in mind, the experiment can be described as: Voices of TM1 are transformed into ones of SF1 by the benign model, and the stenography model takes the output and embed voice data of TM1, generating the final output. Extraction model then use the final output to recover the voices of TM1.

B. Training

We trained the models locally on the server. In order to streamline the experiment process while preserving the most accurate possible results, we avoid excess modifications to the models. Particularly, the benign model is trained to 200000 steps (60 epochs), as the default settings. Similar approach is applied on the stenography model, that it is trained to preferred settings by the model authors.

C. Generation

It's necessary to clarify that, in order to achieve the maximum quantity of samples and prevent the quality of benign model from having excess impact on our overall process, we use the whole VCC2016 training set as the original voices. Each identity contains 162 samples, and we get 648 samples in total.

As a StarGAN-based model, our benign model is capable to transform a voice sequence from any known speakers to another. We consider the fact that the transformed voice belongs to other identity than the original speaker to be a form of desensitization, as mentioned before. Since each sample can be used to generate 3 sequences targeting different speakers, we get 1944 in total.

The remaining steps are straightforward. We used each sample generated by benign model to get a stenography one with the original sample associated with it. Finally we used the extraction model to recover each voices.

D. Evaluation

In order to generate creditable numbers for our voice samples that correctly reflect the amount of sensitive data, the identity of original speaker, we employed cloud-based voiceprint analysis solution provided by iFlyTek. After learning about the identity of four speakers in our domain, this service is able to generate 4 numbers for each voice sample, denoting the probability of the voice to belong to a particular speaker. iFlyTek suggests in the official manual that a score that is higher than 0.6 meant that the identity of a sample can be confirmed.

IV. Analysis

Table I to IV presents the essential statistics of our results.

TABLE I
Statistics of Original Voices

	SF1	SF2	TM1	TM2
Mean	0.853889	0.837284	0.84537	0.837963
Trust Ratio	1.0	0.987654	1.0	1.0
Best	0.95	0.94	0.94	0.94
Worst	0.6	0.56	0.6	0.61
Class Ratio	1.0	1.0	1.0	1.0

TABLE II
Statistics of Desensitized Voices

	SF1	SF2	TM1	TM2
Mean	0.407654	0.440206	0.460556	0.417078
Trust Ratio	0	0.022634	0.047325	0.004115
Best	0.58	0.64	0.62	0.62
Worst	0.22	0.23	0.27	0.21
Class Ratio	0.125514	0.236626	0.195473	0.012346

TABLE III
Statistics of Stenography Voices

	SF1	SF2	TM1	TM2
Mean	0.387695	0.366070	0.412119	0.394115
Trust Ratio	0	0.002058	0	0
Best	0.52	0.61	0.56	0.56
Worst	0.21	0.15	0.17	0.20
Class Ratio	0.119342	0.183128	0.1893	0.014403

V. Discussion

According to our experiments, it is safe to consider conventional acoustical-based or NN-based audio transforming solutions not sufficient for voice desensitization. Besides the StarGAN-VC solution we used, there exists many more such ”voice changer” services on the Internet available for public use. One would consider these solutions secure because of the vast audible differences they made on its voice sequences. However, these solution exhibits potential of adversary with such method we demonstrated in this paper, which is not negligible. It's not likely that human ears can pick up subtle changes a stenography program made to certain parts of a voice sequence.

It's possible to avoid or mitigate risks of being attacked by such method. The most straightforward way to go is to avoid desensitization models from unknown or unsound sources. Due to the black-box nature of various proprietary services, it is not likely possible for users to have practical means to test these services for potential adversaries. When possible, users could train their own desensitization models. If the models are acquired from third-party, users should pay attention to the behaviour of model and the choice of training data set.

TABLE IV
Statistics of Extracted Voices

	SF1	SF2	TM1	TM2
Mean	0.681975	0.681728	0.695123	0.702593
Trust Ratio	0.901235	0.864198	0.950617	0.938272
Best	0.81	0.81	0.81	0.85
Worst	0.44	0.47	0.53	0.46
Class Ratio	0.993827	1	1	0.993827

TABLE V
Statistics of Differences Between Desensitized and Stenography Voices

Mean	0.041373
Variance	0.010478
Maximum	0.37
Minimum	-0.29
Maximum Absolute Value	0.37
Minimum Absolute Value	0

VI. Future Work

In our experiment, we simply directed the voice sequence generated by the benign model to the stenography model. As mentioned before, this approach may not be sufficient in terms of stealthiness. Also, the overall storage consumption of this black box will increase significantly and the processing performance may not be ideal.

We believed that turning the stenography model into extra layers of benign model, eliminating the redundant audio encoding and decoding processes, can mitigate such problems while potentially increase the overall adversarial quality. In this way, the benign model can train in conjunction with the adversary model, taking advantage of intermediate representations of it. Moreover, this creates the potential of reducing the overall storage consumption and be less suspicious.

As another way to mitigate such problem, users can use various traditional or ML-based methods to apply inaudible perturbation on processed voices to attempt erasure of potential stenography while preserving high audio quality.

VII. Conclusion

Targeting voice desensitization models based on generative NN models, we designed an adversary scheme that attempts to recover original voices from desensitized ones via stenography means, thus cause a privacy bleach. Our experiments prove this idea to be viable and the conventional solutions to be vulnerable to this type of adversary.

VIII. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IX. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IX-A–IX-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m^2” or “webers per square meter”, not “webers/m^2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

D. LaTeX-Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

BibTeX does not work by magic. It doesn’t get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can’t read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [11].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE VI
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks …”. Instead, try “R. B. G. thanks…”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [5]. The sentence punctuation follows the bracket [6]. Refer simply to the reference number, as in [7]—do not use “Ref. [7]” or “reference [7]” except at the beginning of a sentence: “Reference [7] was the first …”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [8]. Papers that have been accepted for publication should be cited as “in press” [9]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [10].

References

[1] N. Subramanian, O. Elharrouss, S. Al-Maadeed and A. Bouridane, ”Image Steganography: A Review of the Recent Advances,” in IEEE Access, vol. 9, pp. 23409-23423, 2021, doi: 10.1109/ACCESS.2021.3053998.

[2] https://github.com/Didnelpsun/StarGanVCDialectConversion

[3] arXiv:1806.02169 [cs.SD]

[4] arXiv:1902.03083 [cs.SD]

[5] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[6] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[7] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[8] K. Elissa, “Title of paper if known,” unpublished.

[9] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[10] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[11] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.