Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

You Are What You Say: Exploiting Linguistic Content for VoicePrivacy Attacks

Ünal Ege Gaznepoglu[1], *Anna Leschanowsky*[2], *Ahmad Aloradi*[3], *Prachi Singh*[2], *Daniel Tenbrinck*[3], *Emanuël A. P. Habets*[1], *Nils Peters*[4]

[1]International Audio Laboratories Erlangen, FAU Erlangen-Nürnberg, Germany
[2]Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[3]Department of Data Science, FAU Erlangen-Nürnberg, Germany
[4]Department of Electrical and Electronics Engineering, Trinity College Dublin, Ireland

ege.gaznepoglu@audiolabs-erlangen.de

Abstract

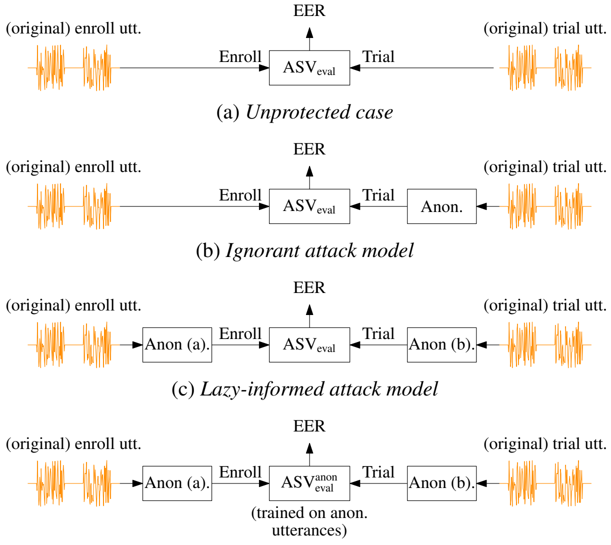
Speaker anonymization systems hide the identity of speakers while preserving other information such as linguistic content and emotions. To evaluate their privacy benefits, attacks in the form of automatic speaker verification (ASV) systems are employed. In this study, we assess the impact of intra-speaker linguistic content similarity in the attacker training and evaluation datasets, by adapting BERT, a language model, as an ASV system. On the VoicePrivacy Attacker Challenge datasets, our method achieves a mean equal error rate (EER) of 35%, with certain speakers attaining EERs as low as 2%, based solely on the textual content of their utterances. Our explainability study reveals that the system decisions are linked to semantically similar keywords within utterances, stemming from how LibriSpeech is curated. Our study suggests reworking the VoicePrivacy datasets to ensure a fair and unbiased evaluation and challenge the reliance on global EER for privacy evaluations.

Index Terms: voice privacy, speaker anonymization, automated speaker verification, language models, explainable AI

1. Introduction

The field of speaker anonymization has emerged in response to the risks associated with advances in speech-processing technology, such as the inadvertent disclosure of personal information (e.g., age, health) when using cloud-enabled voice interfaces [1]. Speaker anonymization systems protect the speaker's identity while preserving important information for downstream tasks such as automatic speech recognition (ASR) and emotion recognition [2]. Recently, a VoicePrivacy Attacker Challenge [3] was held for the first time. Its aim is to develop techniques to compromise the privacy of speakers that have been processed by seven anonymization systems. These include the top three baseline systems from the Voice Privacy Challenge (VPC) 2024: B3, B4, B5, and the top four participant-submitted systems.

Works on both anonymization [4]–[6] and attacks to anonymization systems **zhang˙attacking˙2024**, [7], [8] use the publicly provided VPC datasets and evaluation protocols. These protocols consist of attacks based on automatic speaker verification (ASV) systems for privacy evaluation. As shown in Fig. 1, in *ignorant* and *lazy-informed* attack models, a pre-trained ASV_{eval} is used, while the *semi-informed* attack model uses an $\text{ASV}_{\text{eval}}^{\text{anon}}$ trained on anonymized data. The *unprotected* case serves as a reference. ASV_{eval} and $\text{ASV}_{\text{eval}}^{\text{anon}}$ are based on ECAPA-TDNN [9], and more information can be found in the challenge evaluation plan [3]. The speaker embeddings extracted by these systems are used for enrollment and trial on corresponding `libri-dev` and `libri-test` utterances, using cosine similarity as a similarity measure. Finally, equal error rate (EER) is calculated for male and female speakers. Lower EERs correspond to a better de-anonymization and hence a successful attack.



(a) *Unprotected case*

(b) *Ignorant attack model*

(c) *Lazy-informed attack model*

(d) *Semi-informed attack model*

Figure 1: *VPC attack models define how much information is available to an attacker [10].*

The literature on the analysis of attacker ASV scores is rather limited. The Zero Evidence Biometric Recognition Assessment (ZEBRA) framework provides worst-case privacy disclosure for an individual per anonymization system but no disaggregated scores on a speaker-level [11], [12]. In [13], Williams et al. explored speaker-level distributions of ASV scores obtained through ignorant and lazy-informed attack models by identifying subpopulations with distinct behaviors, using the methodology first introduced in [14]. In [15] the authors used an ignorant attack model on anonymized pathological speech while evaluating the benefit of speaker anonymization for vulnerable subgroups. They found the standard deviation of EERs across healthy speakers to be very low, e.g., $32.26\% \pm 0.31$, for the DSP-based anonymization system [16].

To the best of our knowledge, this study is the first to analyze semi-informed attacker scores on a speaker level to address the following questions: How does speaker-level performance vary for semi-informed attacks applied to various systems, and are there any identifiable patterns in the ASV score distributions that can be exploited?

10.21437/Interspeech.2025-681

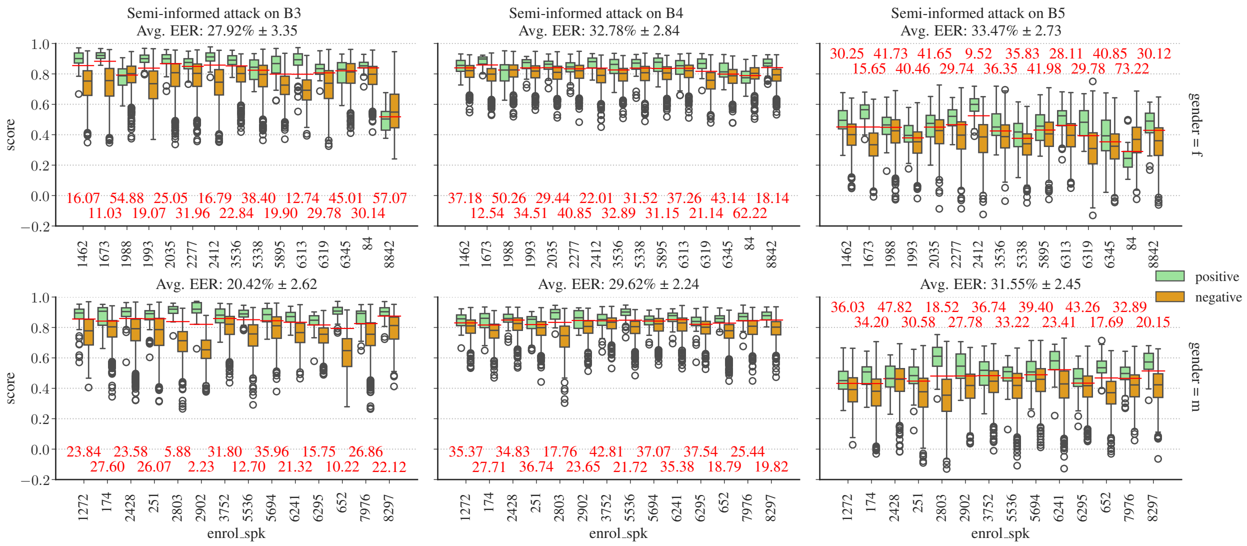


Figure 2: *Speaker-level breakdown of ASV$_{eval}^{anon}$ (ECAPA-TDNN) scores on the* `libri-dev` *dataset. The top row shows female speakers, while the bottom row shows male speakers. Columns correspond to the attacked anonymization system. Bar plots denote the cosine similarity score distributions, where light green bars indicate the positive pairs (enrollment and trial speakers matching) and orange bars indicate the negative pairs (different enrollment and trial speakers). Red lines denote the threshold where the difference between Type 1 and Type 2 errors are minimal for each speaker, and the corresponding EERs (in %) are shown by the red text.*

2. Semi-informed attack: status quo

This section presents an analysis of the ASV$_{\text{eval}}^{\text{anon}}$ scores on the speaker level. The anonymization systems to be attacked are chosen such that their architectures and intermediate representations are diverse. B3 performs any-to-any voice conversion via an ASR-TTS pipeline, where a Wasserstein GAN generates a target pseudo-identity [4]. In contrast, B4 [5] and B5 [6] perform any-to-few voice conversion to real speakers. We have run the VPC 2024 codebase[1] without modification to obtain anonymized utterances and the corresponding ASV$_{\text{eval}}^{\text{anon}}$ scores.

The scores, as well as the speaker-level thresholds and the corresponding EERs are visualized in Fig. 2 for several variables. We observe a notable variability in the EERs across different speakers for each system, some going as low as 2% yet the mean EERs are between 20% and 35%. Although for perfect anonymization a 50% EER is sufficient [17], some speaker-system pairs attain EERs greater than 50%, e.g., speaker `84` exhibits 62% EER when anonymized with B4 and 73% when anonymized with B5. Higher than necessary EERs attained by some speakers may obfuscate others with low EERs when averaged, giving a false sense of privacy. Therefore, when computing mean EERs, we propose clipping speaker EERs exceeding 50% such that they fall into the interval $[0\%, 50\%]$, using

$$f(x) = \min(50, x). \tag{1}$$

Closer inspection of Fig. 2 reveals an odd behavior: none of the considered systems can effectively anonymize some speakers (`1673` and `652`), shown by the EER values less than 20%. This is quite counter-intuitive, given that the considered anonymization systems have very different architectures. What could have caused these speakers to be de-anonymized by ASV$_{\text{eval}}^{\text{anon}}$ in all considered cases?

Table 1: *Training hyperparameters*

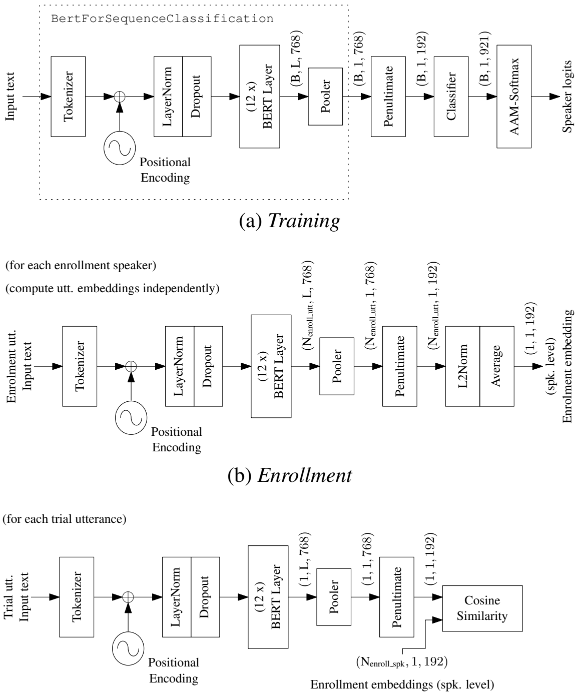
Hyperparameter	**Value**
Num. epochs	20
Batch size	256
Optimizer	`AdamW`, `lr`: 1×10^{-4}
`lr` scheduler	`LinearWithWarmup`
Train-validation split	90%, 10%
AAM parameters	Margin: 0.2, Scale: 30
Dropout probability	0.1

3. Proposed text-based attack

Consistent de-anonymization of specific speakers suggests the existence of persistent features that are invariant to different anonymization strategies. Upon manual inspection, we found that the texts read by speakers `1673` and `652` were on specific and unique topics, and some words recurring across their utterances. So, ASV$_{\text{eval}}^{\text{anon}}$ could be exploiting the intra-speaker linguistic content similarity in LibriSpeech. If true, this would confound the evaluations, because the anonymization systems are expected to preserve the linguistic content.

To test the feasibility of this hypothesis, we built a novel system that imitates ASV$_{\text{eval}}^{\text{anon}}$, but operating only on textual content. We use this 'text-based attack' on the ground truth transcriptions, i.e., the information that an ideal anonymization system would preserve. Our attack is based on the HuggingFace implementation [18] of BERT$_{\text{BASE}}$, first introduced by [19]. Fig. 3 outlines the training, enrollment, and trial phases, which are designed to be as similar as possible to ASV$_{\text{eval}}^{\text{anon}}$, and to align with the VPC evaluation protocol. In particular, the balance between the batch size, the embedding dimensionality and the loss hyperparameters are crucial. We utilize the class `BertForSequenceClassification`, which comprises the BERT architecture and a pooling layer selecting the output token corresponding to the [CLS] input token. Then, a

[1] `https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/`



(a) *Training*

(b) *Enrollment*

(c) *Trial*

Figure 3: *Training, enrollment, and trial phases of our system.*

linear layer (called Penultimate) reduces the 768-dimensional hidden representation to 192 dimensions to ensure additive angular margin (AAM) interacts with the embeddings in a similar fashion to $\text{ASV}_{\text{eval}}^{\text{anon}}$. Then, the classifier, a linear layer with L2-normalized weights and no bias, computes the logits using the embeddings after L2 normalizing them. Pre-existing dropout layers are kept, but the newly added layers do not use any dropout.

We initialize our network with the checkpoint available online[2], and we use normal initialization for new layers. Then, we fine-tune on `train-clean-360`, using Speechbrain's implementation [20] of AAM softmax loss [21], and setting the involved hyperparameters as summarized in Table 1. The training-validation split is performed such that both subsets contain utterances from each speaker and each session (called `spk-diverse-sess` in the VPC 2024 codebase). During training, we track the loss and the classification accuracy on the hold-out validation split. We reserve `libri-dev` and `libri-test` for evaluation purposes. After six epochs of training, the validation accuracy converges to 54%, while the validation loss starts to increase. We interpret this as the model starting to overfit, and to avoid that, we use the checkpoint after six training epochs for evaluations in the upcoming sections.

During our experiments, we noticed that the L2 norms of the utterance-level embeddings differ, even though the network encodes information exclusively via the direction of the embeddings. Therefore, unlike VPC 2024 $\text{ASV}_{\text{eval}}^{\text{anon}}$, our text-based attack normalizes the utterance-level enrollment vectors to unit length (see Fig. 3b). This is to facilitate equal contribution of each utterance towards the speaker-level enrollment vector. The source code to reproduce our work will be released at [3].

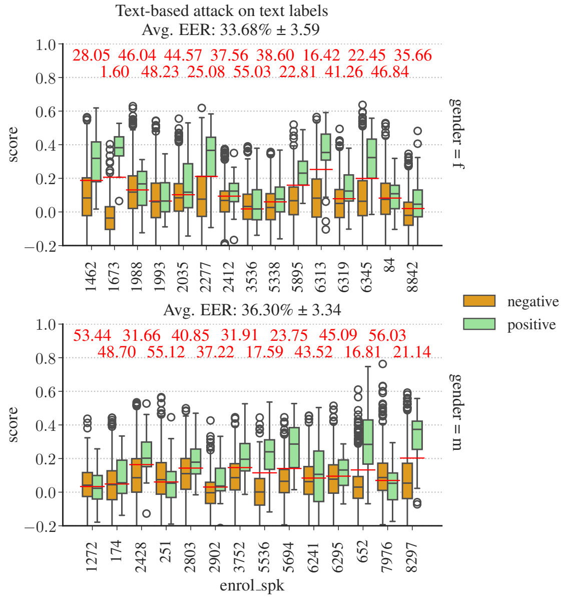


Figure 4: *Score distributions of our attack on `libri-dev`.*

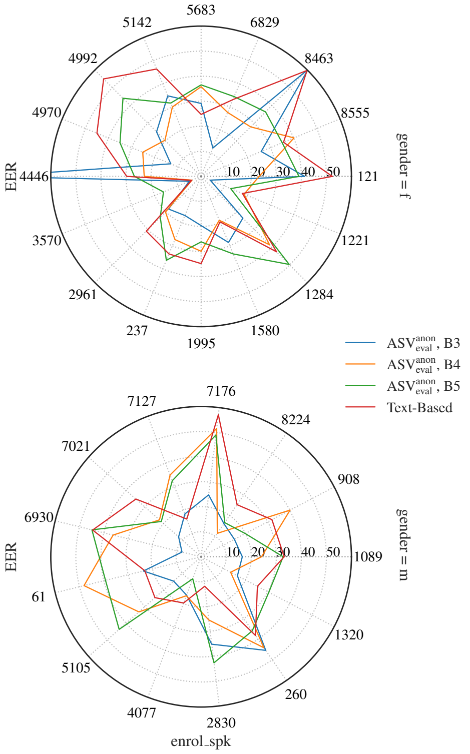


Figure 5: *Radar plots comparing our text-based attack to ASV_{eval}^{anon} on `libri-test` dataset. Spokes indicate enrollment speaker IDs; circular axes show corresponding speaker EERs*

4. Results and discussion

Fig. 4 shows the score distributions of our text-based attack; please refer to Fig. 2 for comparison and interpretation. On average, our attack achieves an EER of 33.68% for female speakers and 36.30% for male speakers, performing only slightly worse than $\text{ASV}_{\text{eval}}^{\text{anon}}$ despite the limited available information.

Turning to speaker-level performance, speakers `1673` and `652`, the ones that were identified in Sec. 2, achieved EERs of

[2] `https://huggingface.co/google-bert/bert-base-uncased`

[3] `https://doi.org/10.5281/zenodo.15526086`

1.60% and 16.81%, respectively, indicating that their anonymity was in fact compromised by the text-based attack. In Fig. 5, we include a radar plot to visualize how our attack compares to $\text{ASV}_{\text{eval}}^{\text{anon}}$ on `libri-test` subset. There, our attack achieved an EER of 4.10% and 11.86% for speakers `3570` and `2830`.

Also, we found that L2-normalizing utterance-level embeddings for enrollment slightly improved the performance of our system, reducing the mean EER by 0.39 and 0.32 percentage points for female and male speakers, respectively. We think the effects of normalization are worth exploring for $\text{ASV}_{\text{eval}}^{\text{anon}}$.

Explainability analysis

We use `transformers-interpret` [22], [23], a Python package, to get insight into our model decisions. Specifically, we apply Layer Integrated Gradients [24] to the cosine similarities in reference to EER thresholds (see Fig. 4). We selected five random trial utterances for three speakers, and used the corresponding speaker-level embedding for enrollment. The resulting attribution scores and word importance scores are visualized in Fig. 6 for two successfully attacked speakers (`1673` and `652`) and one failed attack (`7976`). Overall, the successful attacks are linked to semantically similar keywords, such as religious terms for speaker `1673` (church, Vatican, heretics, ...), and culinary terms for speaker `652` (meat, salad, casserole, ...). In contrast, for speaker `7976`, while some utterances with military terms (regiment, officer, ...) were recognized, many utterances did not have these, so the attack failed (EER: 50.11%).

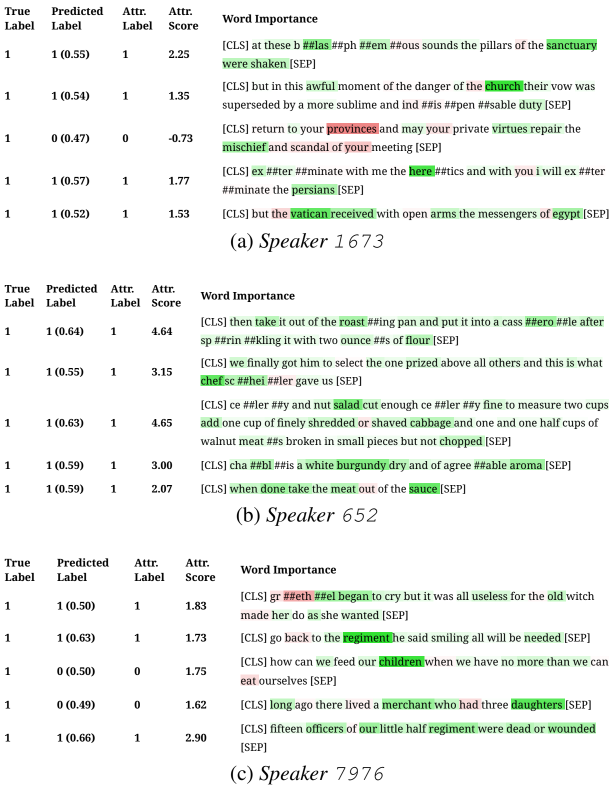
Discussion

Existence of speakers for which the text-based attack is successful suggests that the dataset has exploitable intra-speaker linguistic content similarity. This finding is particularly significant as our model achieves comparable verification performance to $\text{ASV}_{\text{eval}}^{\text{anon}}$ despite only having access to the text in `libri-test`, while $\text{ASV}_{\text{eval}}^{\text{anon}}$ using the anonymized speech.

Beyond global EER values, our text-based attack and $\text{ASV}_{\text{eval}}^{\text{anon}}$ show different behaviors at the speaker level. For certain speakers, e.g., `6829` in `libri-test`, we have observed that the text-based attack failed, but $\text{ASV}_{\text{eval}}^{\text{anon}}$ has been successful. The other case is, e.g., `1284` or `2830` in `libri-test`, where we see that the text-based attack was successful, but (at least some) anonymization systems still managed to attain decent EERs. Besides, the EERs vary for successful and failed attacks, such as our text-based attack achieves 1.60% EER for speaker `1673` but the semi-informed attacks on B3, B4 and B5 attain 11.03%, 12.54% and 15.65%, respectively.

Several factors can explain these variations. First, the success of $\text{ASV}_{\text{eval}}^{\text{anon}}$ in cases where our text-based attack failed may be attributed to the additional information available in speech signals, such as speaking rate and fundamental frequency.

Conversely, our text-based attack's occasional superior performance is likely due to its more sophisticated linguistic understanding, enabled by using a pre-trained BERT model. Notably, attempts to train our attack without pre-trained BERT weights failed to converge. Similarly, our initial experiments using classic NLP methods such as TD-IDF and CountVectorizer also failed, even after preprocessing to remove overly common words. In contrast, $\text{ASV}_{\text{eval}}^{\text{anon}}$, powered by ECAPA-TDNN, is unlikely to learn and utilize semantic similarity of the thematic words to the extent of the text-based attack. Second factor is the possible changes in linguistic content caused by the anonymization process, e.g., speech degradation leading to increased word error rates. We think investigating the effects of anonymization on linguistic content would constitute an interesting follow-up study. Nevertheless, our findings highlight an important vulnerability: speakers can be recognized through their corresponding text, and the risk of attackers exploiting this grows as their architectures become more advanced.



True Label	Predicted Label	Attr. Label	Attr. Score	Word Importance
1	1 (0.55)	1	2.25	[CLS] at these b ##las ##ph ##em ##ous sounds the pillars of the sanctuary were shaken [SEP]
1	1 (0.54)	1	1.35	[CLS] but in this awful moment of the danger of the church their vow was superseded by a more sublime and ind ##is ##pen ##sable duty [SEP]
1	0 (0.47)	0	-0.73	[CLS] return to your provinces and may your private virtues repair the mischief and scandal of your meeting [SEP]
1	1 (0.57)	1	1.77	[CLS] ex ##ter ##minate with me the here ##tics and with you i will ex ##ter ##minate the persians [SEP]
1	1 (0.52)	1	1.53	[CLS] but the vatican received with open arms the messengers of egypt [SEP]

(a) *Speaker* `1673`

True Label	Predicted Label	Attr. Label	Attr. Score	Word Importance
1	1 (0.64)	1	4.64	[CLS] then take it out of the roast ##ing pan and put it into a cass ##ero ##le after sp ##rin ##kling it with two ounce ##s of flour [SEP]
1	1 (0.55)	1	3.15	[CLS] we finally got him to select the one prized above all others and this is what chef sc ##hei ##ler gave us [SEP]
1	1 (0.63)	1	4.65	[CLS] ce ##ler ##y and nut salad cut enough ce ##ler ##y fine to measure two cups add one cup of finely shredded or shaved cabbage and one and one half cups of walnut meat ##s broken in small pieces but not chopped [SEP]
1	1 (0.59)	1	3.00	[CLS] cha ##bl ##is a white burgundy dry and of agree ##able aroma [SEP]
1	1 (0.59)	1	2.07	[CLS] when done take the meat out of the sauce [SEP]

(b) *Speaker* `652`

True Label	Predicted Label	Attr. Label	Attr. Score	Word Importance
1	1 (0.50)	1	1.83	[CLS] gr ##eth ##el began to cry but it was all useless for the old witch made her do as she wanted [SEP]
1	1 (0.63)	1	1.73	[CLS] go back to the regiment he said smiling all will be needed [SEP]
1	0 (0.50)	0	1.75	[CLS] how can we feed our children when we have no more than we can eat ourselves [SEP]
1	0 (0.49)	0	1.62	[CLS] long ago there lived a merchant who had three daughters [SEP]
1	1 (0.66)	1	2.90	[CLS] fifteen officers of our little half regiment were dead or wounded [SEP]

(c) *Speaker* `7976`

Figure 6: *Explainability study results. Tokens that contribute positively to a model decision are highlighted in green, and red stands for negative contributions. The intensity of the highlight signifies the strength. Attribution score is the sum of all word importance scores, used as a measure of confidence. The characters '##' occur when a word is represented by multiple tokens to show that surrounding tokens are part of the same word.*

5. Conclusion

In this work, we explored the speaker-level behavior of $\text{ASV}_{\text{eval}}^{\text{anon}}$ on speaker anonymization systems. In our analysis, we identified that reporting global EERs, which is a common practice in evaluating ASV systems for speaker verification, can obfuscate the shortcomings of speaker anonymization systems by overestimating their effectiveness. To tackle this issue, we proposed reporting average EER after clipping speaker-level EERs exceeding 50%. Furthermore, we identified some speakers in the evaluation datasets, whose anonymity were repeatedly compromised. To investigate if VPC speakers can be identified solely by their linguistic content, we repurposed BERT as an ASV model. Our system achieves EERs less than 20% for 4 / 29 enrolled speakers in `libri-dev` subset and 6 / 29 enrolled speakers in `libri-test`, showing that the linguistic content similarity in the utterances of these speakers is sufficient to verify their identities. Our explainability analysis suggests model decisions are influenced by thematically similar keywords, such as culinary or religious terms. Further work is needed to develop clean datasets for VPC attack training and evaluations and to quantify how much attacks in the literature exploit this.

6. Acknowledgements

This work is partially supported by the German Ministry of Science and Technology (BMBF) under grant agreement No. 01IS24072A (COMFORT).

7. References

[1] N. Tomashenko *et al.*, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech Conf.*, 2020.

[2] P. Champion *et al.*, *3rd VoicePrivacy Challenge Evaluation Plan (Version 2.1)*, 2024. [Online]. Available: https://www.voiceprivacychallenge.org/vp2024/docs/VoicePrivacy_2024_Eval_Plan_v2.1.pdf.

[3] N. Tomashenko, X. Miao, E. Vincent, J. Yamagishi, and N. Evans, *The First VoicePrivacy Attacker Challenge evaluation plan (version 2.2)*, 2024. [Online]. Available: https://www.voiceprivacychallenge.org/attacker/docs/Attacker_Challenge_Eval_Plan.pdf.

[4] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[5] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker Anonymization Using Neural Audio Codec Language Models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[6] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques," PhD thesis, Universite de Lorraine, 2024.

[7] P. Champion, T. Thebaud, G. Le Lan, A. Larcher, and D. Jouvet, "On the Invertibility of a Voice Privacy System Using Embedding Alignment," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[8] P. Champion, D. Jouvet, and A. Larcher, "Evaluating X-Vector-Based Speaker Anonymization under White-Box Assessment," in *Proc. Speech and Computers*, 2021.

[9] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech Conf.*, 2020.

[10] B. M. Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[11] A. Nautsch *et al.*, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Proc. Interspeech Conf.*, 2020.

[12] P.-G. Noé, "Representing evidence for attribute privacy: Bayesian updating, compositional evidence and calibration," PhD thesis, Université d'Avignon, 2023.

[13] J. Williams, K. Pizzi, N. Tomashenko, and S. Das, "Anonymizing Speaker Voices: Easy to Imitate, Difficult to Recognize?" In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[14] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 1998.

[15] S. Tayebi Arasteh *et al.*, "Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech," *Nature Communications Medicine*, vol. 4, no. 1, 2024.

[16] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Proc. Interspeech Conf.*, 2021.

[17] M. Panariello *et al.*, "The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 32, 2024.

[18] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[20] M. Ravanelli *et al.*, "Open-Source Conversational AI with SpeechBrain 1.0," *Journal of Machine Learning Research (JMLR)*, vol. 25, no. 333, 2024.

[21] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition," in *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conf.*, 2019.

[22] C. Pierse, *Transformers-Interpret*, 2021. [Online]. Available: https://github.com/cdpierse/transformers-interpret.

[23] Y. Zhu *et al.*, "Using natural language processing on free-text clinical notes to identify patients with long-term COVID effects," in *Proc. ACM Intl. Conf. on Bioinformatics, Computational Biology and Health Informatics (BCB)*, 2022.

[24] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proc. Intl. Conf. on Machine Learning (ICML)*, 2017.