Dementia Attribute Obfuscation in Speech

Dominika Woszczyk[1]

[1]Imperial College London, United Kingdom

d.woszczyk19@imperial.ac.uk

Abstract

Millions utilize voice services to interact with various types of sensors and services such as smart assistants, smart vehicles, or any other IoT device. However, these voice-activated devices pose privacy risks, as speech analysis can unveil sensitive attributes like gender, emotional state, or health conditions. This research expands attribute obfuscation beyond speaker identity, targeting audio and text domains, and focuses on concealing Alzheimer's disease (AD) in speech, identifiable by both speech and content characteristics. We aim to obfuscate attributes through privacy-preserving text-rewriting and text-to-speech, preserving privacy and concealing the disability.

Index Terms: dementia detection, speech synthesis, diverse paraphrase generation, attribute obfuscation

1. Introduction & Motivation

As users engage with voice-enabled services daily, their interactions generate a continuous stream of new voice recordings. However, these recordings contain far more information than necessary to perform the required task. Over the years, various studies have presented voice conversion techniques aimed at anonymizing speakers in speech, as well as attribute obfuscation methods that attempt to hide identifying information, using techniques such as variational auto-encoders (VQ-VAE) [1], CycleGAN [2] or Gradient Reversal Layer [3, 4, 5]. However, these approaches have primarily focused on manipulating acoustic features while disregarding linguistic content. This oversight leaves room for potential privacy breaches, as malicious agents can still exploit linguistic cues to compromise users' privacy. To illustrate this issue, we consider the case of Alzheimer's disease (AD), a cognitive decline condition characterized not only by observable changes in articulation rate and increased pauses but also by lexical and syntactical changes such as simpler and more generic vocabulary or lessened usage of the passive form. In this work, we want to perform changes on attributes to hide the health characteristics of the speaker, while preserving the identity. Related work in the attribute and authorship obfuscation in text attempts to fool authors and style classifiers to create more robust models or to hide attributes. Several works have examined text sanitising for sensitive information, topic, style or author obfuscation. Common approaches include rule-based methods [6, 7], paraphrasing [8], style transfer [9, 10] and back-translation [11]. Recently, other works have also investigated Differential Privacy (DP) as a means of obfuscating the word distribution [12, 13, 14, 15].

Challenges Available dementia datasets are limited in size due to the difficulty of accessing patients and their diagnoses. This forces us to look at low-resource approaches and to look at the bigger picture to understand key transformations. Another challenge is the different stages of dementia and its progressive impact on speech, which needs to be taken into account in our approach. The choice of downstream tasks is also critical for the evaluation of our system. Given our dataset, we can only try to preserve semantics as the samples do not reveal any sentiment nor are labelled for topics. We tackle this challenge by evaluating our models on auxiliary datasets with clearly defined utility tasks. Finally, we cannot formally prove the privacy gain from our approach. Hence, in our evaluation, we must consider various classifiers for both scenarios where an adversary is oblivious to the obfuscation strategy (static) and where it has access to it (adaptive).

Research Questions & Contributions. Our work aims to perform Privacy Preserving Transformations (PPT) in speech, such that samples with dementia characteristics are converted to healthy speech, and health attributes are hidden from potential adversaries. To this end, we propose an approach that performs obfuscation on the text and feeds it to a text-to-speech system that re-synthesizes it into a waveform preserving the source speaker identity, while also obfuscating acoustic attributes. The research questions we aim to answer are the following: 1) Can we neutralize Alzheimer's disease (AD) samples, such that AD classifiers cannot successfully detect them in text and speech?; 2) Can our system preserve the utility task of sentiment/topic classification and automatic speaker verification (ASV)? The expected contributions are as follows:

1. We propose a novel obfuscation system that alters both audio and text features for attribute obfuscation in speech.
2. We demonstrate the performance of our system on a new type of attribute, Alzheimer's disease.
3. We show that our obfuscation approach is effective against a variety of adversaries (traditional and neural approaches, static and adaptive).

2. Framework for Dementia Attribute Obfuscation in Speech

We consider the scenario where individuals provide audio recordings shared with voice-based services and stored for further analysis. Those speech utterances contain sensitive information both in their form and content. The aim of our system (Figure 1) is to modify speech samples in their acoustic and linguistic characteristics such that the sensitive attribute can no longer be detected but the semantics and utility tasks (sentiment analysis and automatic speaker verification) are preserved. To tackle this challenge we divide our work into three stages, further described in Section 3.

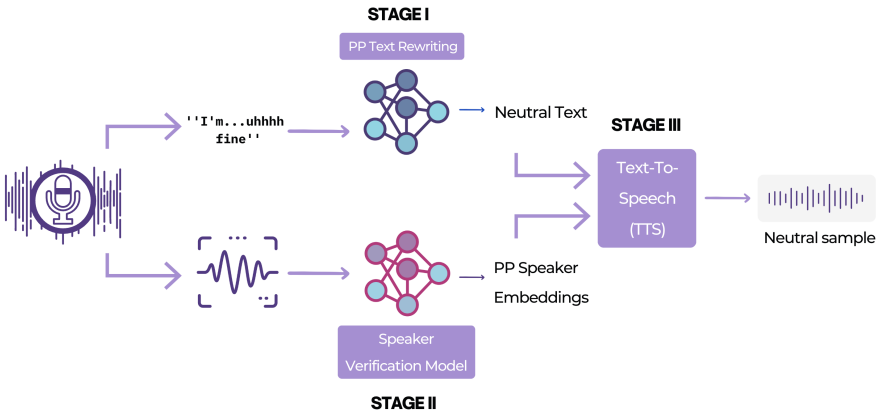

Figure 1: *Overall privacy-preserving voice conversion system. The audio sample and its transcription are fed to their specific models where the privacy-preserving (PP) mechanism is applied, resulting in obfuscated text and speaker embeddings. The new samples are then synthesized back to an obfuscated waveform.*

3. Progress

3.1. Timeline

With **Stage I** completed, **Stage II** being finalized, the current focus is on the final step, **Stage III**. The work has already begun and is expected to be completed over the next 6 months.

3.2. Stage I: Detecting Dementia from Spoken Language

Problem Statement: Previous research has explored the automatic classification of Alzheimer's disease (AD), utilizing both acoustic and linguistic features. However, they require large amounts of labelled data which is not easily available for the task of dementia detection. We investigate the automatic classification of AD and evaluate various data augmentations techniques to tackle the challenges of sparse data.
Methodology: Firstly, we analyze the differences in characteristics between dementia and control patients. We then implemented baseline and state-of-the-art dementia detection models (audio, text and fusion) on the ADReSS (ADR) [16] and DementiaBank (DB) [17] datasets. We then devise and compare the impact on classification accuracy and label preservation for several data augmentations techniques across text and audio.
Results & Conclusion: In our analysis, we observe changes for AD patients in terms of longer syntactical dependencies, simpler vocabulary and higher rates of filler words. Acoustic changes include a slower articulation rate, mispronunciations, and longer silences. The most successful data augmentations techniques were back-translation and voice conversion, and our augmented models on text (Acc. 85%) and audio (Acc. 74%) achieved accuracies comparable to larger and more complex models (Text: 85%, Audio: 74%).

3.3. Stage II: Dementia-Attribute Obfuscation in Text

Problem Statement: Current works for attribute obfuscation in text consider style transfer from one attribute to another (which requires training data and a new model for each variation of an attribute), more classical stylometry techniques or paraphrasing with poor semantic reconstruction. Given earlier analyses of dementia in text indicating that syntax plays an important role in identifying Alzheimer's patients, we explore a low-resource approach that obfuscates dementia and possibly other attributes by generating syntactically diverse paraphrases.
Methodology: We design and evaluate a model, which given a syntax distance, can generate a sentence fulfilling that condition, trained on ParaNMT50m dataset [18]. We also filter out syntactically and lexically similar sentences in our training set to further enhance syntactical distances and standardize the vocabulary. We then compare our model to neural paraphrasing models [19, 20], rule-based [6], style transfer models [9, 21, 10] on their obfuscation abilities against our classifiers from **Stage I** on the ADR and DB datasets. We also evaluate those models in terms of semantic preservation.
Results & Conclusion: We show that paraphrasing models do not disturb the syntax enough to obfuscate dementia and that more drastic techniques lose semantics (0.53 similarity). Our approach shows that by creating syntactically diverse sentences, we can remove dementia characteristics while preserving semantics (0.86 similarity). Our strategy is effective both against static (Acc. 54%) and adaptive adversaries (Acc. 62%). In our next experiments, we will further finetune our approach and evaluate our model for age and gender attributes.

3.4. Stage III: Privacy-Preserving Text-to-Speech for Dementia Speakers

Problem Statement: Given obfuscated text data, free of the dementia attribute, we can re-synthesize samples using a text-to-speech model. However, to preserve the original speaker's voice and perform speaker verification, the model needs to be either trained on a large dataset of that speaker or conditioned on the speaker's representation. Attribute leakage from the speaker and acoustic embeddings has been shown for accent, age, and gender. We hypothesize that it is also the case for dementia attributes and explore strategies and disentanglement techniques to sanitize speaker embeddings while preserving speaker similarity and quality.
Methodology and Future Work: We first investigate attribute leakage by extracting speaker embeddings for DB and ADR samples and performing classification on the embeddings. We are now exploring various ways to disentangle and modify speaker embeddings such that we remove the dementia features but preserve speaker identity. Next, we will compare it to naive mean-shift and disentanglement approaches.
Results: Our initial results have shown that dementia can be leaked from speaker embeddings.

4. References

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

[2] ——, "Emotion filtering at the edge," in *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*, 2019, pp. 1–6.

[3] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," *arXiv preprint arXiv:2012.04454*, 2020.

[4] L. Benaroya, N. Obin, and A. Roebel, "Beyond voice identity conversion: Manipulating voice attributes by adversarial learning of structured disentangled representations," *arXiv preprint arXiv:2107.12346*, 2021.

[5] C. Emmery, E. Manjavacas, and G. Chrupała, "Style obfuscation by invariance," *arXiv preprint arXiv:1805.07143*, 2018.

[6] T. Gröndahl and N. Asokan, "Effective writing style transfer via combinatorial paraphrasing." *Proc. Priv. Enhancing Technol.*, vol. 2020, no. 4, pp. 175–195, 2020.

[7] J. Bevendorff, T. Wenzel, M. Potthast, M. Hagen, and B. Stein, "On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification," *it-Information Technology*, vol. 62, no. 2, pp. 99–115, 2020.

[8] J. Mattern, B. Weggenmann, and F. Kerschbaum, "The limits of word level differential privacy," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 867–881. [Online]. Available: https://aclanthology.org/2022.findings-naacl.65

[9] R. Shetty, B. Schiele, and M. Fritz, "{A4NT}: Author attribute anonymity by adversarial training of neural machine translation," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1633–1650.

[10] F. Mireshghallah and T. Berg-Kirkpatrick, "Style pooling: Automatic text style obfuscation for improved classification fairness," *arXiv preprint arXiv:2109.04624*, 2021.

[11] Q. Xu, L. Qu, C. Xu, and R. Cui, "Privacy-aware text rewriting," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 247–257.

[12] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, "Preech: A system for {Privacy-Preserving} speech transcription," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2703–2720.

[13] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. Chow, "Differential privacy for text analytics via natural text sanitization," *arXiv preprint arXiv:2106.01221*, 2021.

[14] N. Fernandes, M. Dras, and A. McIver, "Author obfuscation using generalised differential privacy," *arXiv preprint arXiv:1805.08866*, 2018.

[15] O. Feyisetan, B. Balle, T. Drake, and T. Diethe, "Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 178–186.

[16] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[17] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[18] J. Wieting and K. Gimpel, "Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations," *arXiv preprint arXiv:1711.05732*, 2017.

[19] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[20] E. Bandel, R. Aharonov, M. Shmueli-Scheuer, I. Shnayderman, N. Slonim, and L. Ein-Dor, "Quality controlled paraphrase generation," *arXiv preprint arXiv:2203.10940*, 2022.

[21] A. Mahmood, F. Ahmad, Z. Shafiq, P. Srinivasan, and F. Zaffar, "A girl has no name: Automated authorship obfuscation using mutant-x." *Proc. Priv. Enhancing Technol.*, vol. 2019, no. 4, pp. 54–71, 2019.