Controllable Voice Anonymization for Privacy-Preserving Disease Detection from Speech

Ben Luks[1,2,3], *Francisco Teixeira*[1], *Alberto Abad*[1,2], *Sebastian Möller*[3], *Isabel Trancoso*[1]

[1]INESC-ID, Portugal
[2]Instituto Superior Técnico, University of Lisbon, Portugal
[3]Technische Universität Berlin, Germany

benluks@inesc-id.pt

Abstract

Voice anonymization systems have the potential to achieve greater voice privacy, although existing systems suffer from a lack of use-case specificity and controllability [1]. One such use-case is the detection of speech-affecting diseases from voice data.

Emerging research shows promising results in early prediction of such diseases [2], but the required information is inherently sensitive. Speech conveys private attributes beyond linguistic content, including health, emotional state, and identity [3]. Privacy-preserving processing can therefore be achieved by transforming speech to limit disclosure of private attributes while preserving task-relevant information [4], aligning with information-theoretic formulations of the privacy–utility trade-off [5].

This work studies disease detection in anonymized speech. State-of-the-art anonymization pipelines constrain information through bottlenecks such as quantized SSL features [6] or phone transcriptions [7]. Utility-focused anonymization is achieved by controlling the flow of information through these bottlenecks. My research therefore focuses on: (1) **acoustic analysis** of disease-related phenomena under anonymization, (2) **feature disentanglement**, and (3) **controllable synthesis** for selectively preserving or suppressing health-related information.

Acoustic Analysis

Speech biomarkers include both acoustic and prosodic characteristics. Prior work has identified measurable correlates between speech and disease [8, 9], but their behavior under anonymization remains underexplored beyond limited studies on Parkinson's disease [10, 11].

Prosodic features such as pitch, energy, and unit durations are also informative [12, 13, 14]. In anonymization pipelines, linguistic units (e.g., VQ tokens from ASR-BN [6]) provide a discrete structure analogous to phones, enabling duration modeling as mappings between units and frame lengths. I therefore propose modifying unit durations via duration predictors (DPs), provided durations modulation would not destroy the desired biomarkers.

Preliminary results indicate that interpolating toward predicted durations consistently increases EER, while incurring only gradual degradation in ASR performance. These findings are consistent with prior work showing that phone durations alone can leak speaker identity [15, 16], motivating their explicit control as a component of anonymization.

Feature Disentanglement

Speech models encode multiple attributes in entangled representations. Disentanglement separates information in these representations such that task-relevant attributes can be extracted. Common approaches use adversarial classifiers with gradient reversal [17] to suppress attributes such as speaker identity, age, or emotion [18, 19, 20], sometimes combined with mutual information minimization [21].

These methods assume separable structure in latent space, which is often imperfect in practice. Bottleneck-based approaches instead suppress information globally and reintroduce desired attributes. The ASR-BN pipeline [22] uses a low-capacity VQ codebook to attenuate speaker information, while models such as emotion2vec [23] similarly enforce compact representations.

Recent work combines these approaches via residual vector quantization (RVQ) [24], enabling hierarchical disentanglement of attributes [25]. This suggests that biomarker-related attributes can be associated with specific levels of the hierarchy, enabling structured and interpretable control.

Controllability

Disentangled representations enable controllable synthesis. Duration modeling, for instance, could allow replacement of source durations with those characteristic of a target speaker. Current results primarily obscure rather than precisely control durations, suggesting the need for more expressive or categorical duration models [26].

Representation-level control can also be achieved through interpolation. Prior work demonstrates linear structure in speech representations, enabling transformations such as accent conversion [27] and phonetic manipulation [28]. These findings support the possibility of selectively combining disentangled biomarker-related features.

Modern TTS systems provide further insight. Flow-matching [29] models reconstruct masked spectrograms [29, 30, 31], while mechanisms such as classifier-free guidance (CFG) [32] and feature scaling enable controllable generation [33]. These approaches can preserve paralinguistic features such as laughing and breathing [34], suggesting a pathway for controlling biomarker-relevant attributes in anonymized speech.

This work aims to bridge privacy and utility by developing controllable anonymization systems that preserve clinically relevant speech information. By combining acoustic analysis, disentangled representations, and controllable synthesis, it supports the development of privacy-preserving disease detection from speech.

1. Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme grant No

101168193 (DOI: https://doi.org/10.3030/101168193) and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025).
Index Terms: voice anonymization, speech privacy, speech-based disease detection, feature disentanglement

2. References

[1] S. Meyer and N. T. Vu, "Use Cases for Voice Anonymization," in *5th Symposium on Security and Privacy in Speech Communication*. ISCA, pp. 73–84. [Online]. Available: https://www.isca-archive.org/spsc_2025/meyer25_spsc.html

[2] S. A. Sheikh, M. Sahidullah, and I. Kodrasi, "Overview of Automatic Speech Analysis and Technologies for Neurodegenerative Disorders: Diagnosis and Assistive Applications," pp. 1–20. [Online]. Available: http://arxiv.org/abs/2501.03536

[3] T. Bäckström, "Privacy in Speech Technology," vol. 113, no. 7, pp. 668–692. [Online]. Available: https://ieeexplore.ieee.org/document/11261339/

[4] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques." [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[5] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the Information Bottleneck to the Privacy Funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*, pp. 501–505. [Online]. Available: https://ieeexplore.ieee.org/document/6970882/

[6] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques." [Online]. Available: http://arxiv.org/abs/2308.04455

[7] 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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096607/

[8] C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Speech as a Biomarker for Disease Detection," vol. 12, pp. 184 487–184 508. [Online]. Available: https://ieeexplore.ieee.org/document/10767227/

[9] S.-I. Ng, L. Xu, I. Siegert, N. Cummins, N. R. Benway, J. Liss, and V. Berisha, "An End-to-End Overview of Clinical Speech AI," vol. 34, pp. 1016–1048. [Online]. Available: https://ieeexplore.ieee.org/document/11371361

[10] M. Baas, B. Van Niekerk, and H. Kamper, "Voice Conversion With Just Nearest Neighbors," in *INTERSPEECH 2023*. ISCA, pp. 2053–2057. [Online]. Available: https://www.isca-archive.org/interspeech_2023/baas23_interspeech.html

[11] C. Franzreb, F. Teixeira, B. Luks, S. Möller, and A. Abad. Evaluating Parkinson's Disease Detection in Anonymized Speech: A Performance and Acoustic Analysis. [Online]. Available: http://arxiv.org/abs/2603.07544

[12] S. Skodda, W. Visser, and U. Schlegel, "Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms," vol. 25, no. 1, pp. 76–82.

[13] S. Frota, M. Cruz, R. Cardoso, I. Guimarães, J. J. Ferreira, S. Pinto, and M. Vigário, "(Dys)Prosody in Parkinson's Disease: Effects of Medication and Disease Duration on Intonation and Prosodic Phrasing," vol. 11, no. 8, p. 1100.

[14] S. M. Luciano, F. Panico, R. De Biase, L. Catalano, L. Sagliano, and L. Trojano, "Neural correlates of emotional prosody in Parkinson's disease: A systematic review." [Online]. Available: https://doi.org/10.3758/s13415-025-01379-w

[15] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," pp. 2765–2769. [Online]. Available: https://www.isca-archive.org/interspeech_2024/gengembre24_interspeech.html

[16] N. Tomashenko, E. Vincent, and M. Tommasi, "Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems," pp. 5128–5132. [Online]. Available: https://www.isca-archive.org/interspeech_2025/tomashenko25_interspeech.html

[17] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, pp. 1180–1189. [Online]. Available: https://proceedings.mlr.press/v37/ganin15.html

[18] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Interspeech 2019*, pp. 3700–3704. [Online]. Available: http://arxiv.org/abs/1911.04913

[19] F. Teixeira, A. Abad, B. Raj, and I. Trancoso, "Privacy-Oriented Manipulation of Speaker Representations," vol. 12, pp. 82 949–82 971. [Online]. Available: https://ieeexplore.ieee.org/document/10547045

[20] R. Aloufi, H. Haddadi, and D. Boyle. Emotionless: Privacy-Preserving Speech Analysis for Voice Assistants. [Online]. Available: http://arxiv.org/abs/1908.03632

[21] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion. [Online]. Available: http://arxiv.org/abs/2106.10132

[22] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco. The VoicePrivacy 2024 Challenge Evaluation Plan. [Online]. Available: http://arxiv.org/abs/2404.02677

[23] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, pp. 15 747–15 760. [Online]. Available: https://aclanthology.org/2024.findings-acl.931/

[24] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive Image Generation using Residual Quantization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11 513–11 522. [Online]. Available: https://ieeexplore.ieee.org/document/9879532/

[25] J. Yao, H. Liu, E. S. Chng, and L. Xie, "EASY: Emotion-aware Speaker Anonymization via Factorized Distillation," in *Interspeech 2025*. ISCA, pp. 3219–3223. [Online]. Available: https://www.isca-archive.org/interspeech_2025/yao25_interspeech.html

[26] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. A. Nguyen, M. Riviere, A. Mohamed, E. Dupoux, and W.-N. Hsu, "Text-Free Prosody-Aware Generative Spoken Language Modeling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, pp. 8666–8681. [Online]. Available: https://aclanthology.org/2022.acl-long.593/

[27] T. Lertpetchpun, T. Trachu, J. Lee, T. Feng, D. Byrd, and S. Narayanan. Accent Vector: Controllable Accent Manipulation for Multilingual TTS Without Accented Data. [Online]. Available: http://arxiv.org/abs/2603.07534

[28] K. Choi, E. Yeo, C. J. Cho, D. Harwath, and D. R. Mortensen. [b]=[d]-[t]+[p]: Self-supervised Speech Models Discover Phonological Vector Arithmetic. [Online]. Available: http://arxiv.org/abs/2602.18899

[29] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling." [Online]. Available: https://openreview.net/forum?id=PqvMRDCJT9t

[30] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Curran Associates Inc., pp. 14 005–14 034.

[31] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Association for Computational Linguistics, pp. 6255–6271. [Online]. Available: https://aclanthology.org/2025.acl-long.313/

[32] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance." [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI

[33] Resemble AI, "Chatterbox-TTS." [Online]. Available: https://github.com/resemble-ai/chatterbox

[34] J. Cui, Z. Yang, N. Li, J. Tian, X. Ma, Y. Zhang, G. Chen, R. Yang, Y. Cheng, Y. Zhou, G. Yu, X. Gu, and J. Tang. GLM-TTS Technical Report. [Online]. Available: http://arxiv.org/abs/2512.14291