Memorandum: Ethical Justification for Controlled Access to Raw Voice Data Samples

Key Takeaways

- The release of raw voice data samples in the B2AI-Voice dataset poses a residual risk of enabling participant re-identification due to the nature of the data (i.e. unprocessed audio recordings accompanied by demographic and clinical information), and recent/emerging advances in AI and speech recognition technology.
- Despite the risks to participant privacy, raw voice data samples combined with demographic and clinical data offer immense value for research and scientific progress.
- To balance risk and value, B2AI-Voice is releasing raw voice data samples through a Controlled Access model.

Introduction

This Memorandum supports the release of Bridge 2 Artificial Intelligence Voice Consortium ("B2AI-Voice") Controlled Access data. It outlines the rationale for releasing raw voice data samples (i.e. unprocessed audio recordings of human speech) under a Controlled Access model, thereby limiting data access to approved researchers who obtain a purpose-specific access approval and sign an institutional data access agreement.

I. Overview of the B2AI-Voice Data Release Strategy

B2AI-Voice is committed to balancing the value of sharing raw voice data samples with the research community with the imperative to safeguard participant privacy. Given the sensitivity of the information collected from consented research participants (including demographic, clinical, and voice data), B2AI-Voice has opted for a three-tiered data release model where data that poses a higher risk of enabling research participant re-identification is released according to more restrictive access controls, and data that poses a lower risk of participant re-identification is released more openly. The B2AI-Voice access tiers, from most restrictive to least restrictive, are Controlled, Registered, and Open. As of April 2025, B2AI-Voice has released data under the Controlled and Registered tiers, with the Open access data available online on the consortium's dashboard.

The tiered data release model (see table below) aligns with global bioethics guidance and to U.S. legislation that regulates scientific research involving humans, chiefly the Federal Policy for the Protection of Human Subjects[1] (popularly known as the "Common Rule"). Further, this model actualizes National Institutes of Health ("NIH") policies[2], which require researchers to implement access controls proportionate to the risks associated with

---

[1] U.S. Dep't of Health & Human Servs., *Federal Policy for the Protection of Human Subjects (45 CFR Part 46)*, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html (last visited Apr. 1, 2025).

[2] National Institutes of Health, *Information to the NIH Policy for Data Management and Sharing: Protecting Privacy When Sharing Human Research Participant Data*, (Jan. 25, 2023), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html; National Institutes of Health, *Designating Scientific Data for Controlled Access*, https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/designating-scientific-data-for-controlled-access#:~:text=Researchers%20should%20consider%20sharing%20participants,%2C%20informed%20consent%2C%20and%20agreements (last visited Apr. 2, 2025).

participant data disclosure. Finally, it guarantees respect for the commitments made to B2AI-Voice research participants when their informed consent was obtained.

Open Access	Registered Access	Controlled Access
Aggregated summary-level data.	Individual-level coded data (i.e., structured clinical and demographic information) and spectrogram data.	Raw voice data, select sensitive clinical and demographic information.
Open access data is made available to the general public through a website that anyone can access.	Researchers must create an account on PhysioNet and enter into a click-wrap agreement confirming their responsible use of research data.	Researchers must apply, receive approval from a data access committee (DAC), and enter into an institutional data access agreement (DAA).

II. Regulatory Frameworks

A Controlled Access data release model can facilitate compliance with provisions of the (a) Common Rule and (b) NIH guidance, which establish requirements and best practices for research involving human participants and safe handling of research derived data.

a. The Common Rule

The Common Rule regulates scientific research involving human participants, including the secondary use and disclosure of identifiable private information,[3] defined as "information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information."[4]

Secondary research with identifiable private information typically requires IRB approval and must respect applicable informed consent.[5] The level of review and protection required scales with the risks established under 46.111(a)(7) including potential harms posed to research participants.

The Common Rule provides exemptions under 46.104(d)(4) for certain secondary research using identifiable private information when specific criteria are met, but these exemptions do not eliminate the responsibility to implement appropriate privacy safeguards.

Identifiability is evaluated relative to the potential identifiers present in the data, the context of data release (i.e., the chosen release model, incentives, and opportunities), and the presently available methods and technologies.[6] For example, the Common Rule provides that federal agencies must re-define what makes health data "identifiable" every four years, to maintain alignment with technological advancements.[7]

b. NIH Guidance on Controlled Access for Data Sharing

[3] 45 CFR § 46.101, 46.104(4).
[4] 45 CFR § 46.102(e)(5).
[5] 45 CFR 46.103.
[6] K. El Emam & L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started* (O'Reilly Media, Inc. 2013).
[7] U.S. Dep't of Health & Human Servs (n 1) 46.102.

The National Institutes of Health (NIH) has established guidelines to protect human participant information and align with federal regulations, including the Common Rule, by emphasizing the importance of controlled access mechanisms. Specifically, the Guidance recommends a controlled access model be implemented when:

(1) The data could be considered substantively sensitive by the inclusion of:
   a. Potentially stigmatizing information which could be used for discriminatory purposes.
   b. Unique attributes related to specific cohorts which could increase the risk of re-identification.
(2) The data cannot be stripped of its identifiability for which the possibility of re-identifying participants cannot be sufficiently reduced, allowing for inferences about the participants identity to still be made.
(3) The risk of unanticipated approaches or technologies that could pose a risk to participant privacy when there is a lack of control over data access, thereby changing the threshold of identifiability.[8]

This Guidance suggests that Controlled Access to raw voice data samples is justified when the data embodies the above characteristics, wherein significant risks to participant privacy are established which must be met with proportionate safeguards.

III. Privacy Risks Associated with Raw Voice Data Samples in the B2AI-Voice Dataset

In following the imperative of the Common Rule to protect identifiable private information, and NIH policy guidelines on safeguarding sensitive data, the following privacy risk analysis has been conducted for raw voice data samples which demonstrates the need for the data to be released under Controlled Access to mitigate the risk of participant re-identification.

a. Nature of the data

The B2AI-Voice raw voice data samples contain a unique variety of acoustic data (respiratory sounds, cough sounds, prolonged vowel, glides, validated speech and free speech data). Few empirical measures exist to quantify the risk of participant re-identification associated with unstructured voice data. Generally, the presence of direct identifiers (e.g., name, civic address, social security number) that could be combined with other public information to enable identification (e.g., profession, partial geographic location, uncommon demographic identifiers) is considered in assessing re-identification risk.[9]

Raw voice data could also contain numerous possible indirect identifiers. Information about an individual's socioeconomic status, their ethnicity, gender identity, and other personal characteristics could be derived from speech, either in listening to it, or through the algorithmic inference of personal characteristics. The content of raw voice data can also be used to attempt re-identification, where it relays information about the individual that could enable identification.[10]

---

[8] National Institutes of Health, *Designating Scientific Data for Controlled Access*, https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/designating-scientific-data-for-controlled-access#:~:text=Researchers%20should%20consider%20sharing%20participants,%2C%20informed%20consent%2C%20and%20agreements (last visited Apr. 2, 2025).
[9] A. Moretón & A. Jaramillo, *Anonymisation and Re-identification Risk for Voice Data*, 7 Eur. Data Prot. L. Rev. 274, 284 (2021), https://doi.org/10.21552/edpl/2021/2/20; M. Phillips & B.M. Knoppers, *The Discombobulation of De-identification*, 34 Nature Biotech. 1102, 1103 (2016), https://doi.org/10.1038/nbt.3655; K. El Emam, *Risk-based De-identification of Health Data*, 8 IEEE Security & Privacy 64, 67 (2010), https://doi.org/10.1109/MSP.2010.103..
[10] S. Ribaric, A. Ariyaeeinia & N. Pavesic, *De-identification for Privacy Protection in Multimedia Content: A Survey*, 47 Signal Processing: Image Comm. 131, 151 (2016), https://doi.org/10.1016/j.image.2016.02.003.

Unlike visual representations of voice data (e.g., spectrograms), raw voice data cannot be reliably anonymized without eliminating information that is useful to science, thereby reducing its utility for research purposes. For example, available methods of transforming raw voice data into synthetic data can eliminate features that are of research interest. These also pose challenges because it is seldom possible to eliminate both the risk of re-identification associated to the content of speech, and that associated to individual characteristics that can be inferred from a person's voice, whilst also maintaining its scientific utility.[11] Thus, because algorithmic re-identification attempts could plausibly be performed using raw voice data, and because it is not possible to perform de-identification without compromising much of the data's scientific value, the risk of re-identification inherent to sharing raw voice data samples for secondary research is considered high.

b. Context of data release

Context risk considers the openness of data release, the incentives to perform re-identification, and the degree to which technological safeguards or organizational mechanisms (e.g., contracts and access controls) mitigate potential re-identification. It also considers how resource-intensive it would be to attempt re-identification, and the degree of specialized knowledge required to do so. Available algorithmic tools could be used to attempt re-identification from the content or voice characteristics of data.[12] The use of these tools is not anticipated to require significant professional expertise. Incentives to attempt re-identification are anticipated to be high for raw voice data, as this data type can be used for numerous fraudulent purposes.[13]

Controlled access release for raw voice data reduces the associated context risk, in that the incentives of authorized research users to attempt re-identification are lowered through the implementation of contractual safeguards.

c. Consideration for current and future technological advancements

Current and emerging technology could reduce the level of technological proficiency required to re-identify participants using the data which additionally contributes to the privacy risk. Progress in speaker recognition and rapid advancements in AI-based voice re-identification pose increasing risks, even when metadata is removed.[14] Additionally, future AI models may develop stronger de-anonymization capabilities, posing unknown risks to participant privacy. Accordingly, available algorithmic tools could be used to attempt re-identification from the content or voice characteristics of data, especially as the use of these tools is not anticipated to require significant professional expertise.

Conclusion

The sharing of raw voice data samples from B2AI-Voice represents the highest-risk stage of data release due to the potential for participant re-identification, driven by both the inherent characteristics of the data and the context of advancing AI and speech recognition technologies. Classification of raw voice data samples under

---

[11] Q. Jin, A.R. Toth, T. Schultz & A.W. Black, *Voice Convergin: Speaker De-identification by Voice Transformation*, in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* 3909, 3912 (2009), https://doi.org/10.1109/ICASSP.2009.4959689.

[12] Yukun Li et al., *Perovskite Nanocrystals for Photodetectors: Engineering the Interfaces and Performance*, 10 *Adv. Sci.* 202309826 (2023), https://doi.org/10.1002/advs.202309826.

[13] Lichao Zhai et al., *Deep Reinforcement Learning-Based Human-Robot Interaction with a Dynamic Task Allocation for Collaborative Assembly Tasks*, 23 *Auton. & Adaptive Syst.* 10688 (2023), https://doi.org/10.1007/s10462-023-10688-w.

[14] Shuo Yang et al., *A Unified Deep Learning Framework for Cross-Domain Image Style Transfer*, in *Proceedings of the 2023 IEEE International Conference on Computer Vision (ICCV)*, IEEE (2023), https://ieeexplore.ieee.org/document/10289030.

Controlled Access facilitates respect for ethical-legal requirements, respect for NIH guidelines, and best balances the open sharing the data as a resource that can be leveraged by researchers, and the protection of research participant privacy.

Bibliography

A. Moretón & A. Jaramillo, *Anonymisation and Re-identification Risk for Voice Data*, 7 Eur. Data Prot. L. Rev. 274, 284 (2021), https://doi.org/10.21552/edpl/2021/2/20.

K. El Emam & L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started* (O'Reilly Media, Inc. 2013).

K. El Emam, E. Jonker, L. Arbuckle & B. Malin, *A Systematic Review of Re-identification Attacks on Health Data*, 6 PLoS ONE e28071 (2011), https://doi.org/10.1371/journal.pone.0028071.

K. El Emam, *Risk-based De-identification of Health Data*, 8 IEEE Security & Privacy 64, 67 (2010), https://doi.org/10.1109/MSP.2010.103.

Lichao Zhai et al., *Deep Reinforcement Learning-Based Human-Robot Interaction with a Dynamic Task Allocation for Collaborative Assembly Tasks*, 23 *Auton. & Adaptive Syst.* 10688 (2023), https://doi.org/10.1007/s10462-023-10688-w.

M. Phillips & B.M. Knoppers, *The Discombobulation of De-identification*, 34 Nature Biotech. 1102, 1103 (2016), https://doi.org/10.1038/nbt.3655.

National Institutes of Health, *Designating Scientific Data for Controlled Access*, https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/designating-scientific-data-for-controlled-access#:~:text=Researchers%20should%20consider%20sharing%20participants,%2C%20informed%20consent%2C%20and%20agreements (last visited Apr. 2, 2025).

National Institutes of Health, *Information to the NIH Policy for Data Management and Sharing: Protecting Privacy When Sharing Human Research Participant Data*, (Jan. 25, 2023), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html.

Q. Jin, A.R. Toth, T. Schultz & A.W. Black, *Voice Convergin: Speaker De-identification by Voice Transformation*, in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* 3909, 3912 (2009), https://doi.org/10.1109/ICASSP.2009.4959689.

S. Ribaric, A. Ariyaeeinia & N. Pavesic, *De-identification for Privacy Protection in Multimedia Content: A Survey*, 47 Signal Processing: Image Comm. 131, 151 (2016), https://doi.org/10.1016/j.image.2016.02.003.

Shuo Yang et al., *A Unified Deep Learning Framework for Cross-Domain Image Style Transfer*, in *Proceedings of the 2023 IEEE International Conference on Computer Vision (ICCV)*, IEEE (2023), https://ieeexplore.ieee.org/document/10289030.

U.S. Dep't of Health & Human Servs., *Federal Policy for the Protection of Human Subjects (45 CFR Part 46)*, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html (last visited Apr. 1, 2025).

Yukun Li et al., *Perovskite Nanocrystals for Photodetectors: Engineering the Interfaces and Performance*, 10 *Adv. Sci.* 202309826 (2023), https://doi.org/10.1002/advs.202309826.