Scoping Review of Voice Datasets Available for Voice AI Research

Ariel Farb, MS[1]; Yaël Bensoussan, MD[2]; Alisha Kodibagkar, MSE[3]; Satrajit S. Ghosh, PhD[3]; Anais Rameau, MD MPhil MS[1]; Maria E. Powell, PhD CCC-SLP[4]

[1]Weill Cornell Medicine, New York, NY; [2]University of South Florida, Tampa, FL; [3]Massachusetts Institute of Technology, Cambridge, MA; [4]Vanderbilt University Medical Center, Nashville, TN

OBJECTIVE

A significant barrier to artificial intelligence-driven biomedical research is lack of access to health-related data. Researchers outside of large, research-oriented healthcare institutions must often rely on public datasets to fuel their research; however, these publicly available resources can be challenging to find and evaluate.

The objective of the NIH ***Bridge2AI Voice*** consortium is to develop a large, ethically-sourced, publicly-available dataset including voice and other multi-modal health data to fuel AI research for biomedical applications. One goal of this organization is to develop a resource of existing voice datasets used for AI applications as a reference for AI researchers. The purpose of this scoping review is to investigate the current state of open-access, acoustic datasets for voice, neurological, and mood disorders by (1) describing their sample sizes, reported demographics, and speech tasks, and (2) identifying similarities in collected variables across disease cohorts. The long-term goal is to develop a more standardized and unbiased data collection protocol and provide access to these open-source speech datasets.

METHODS

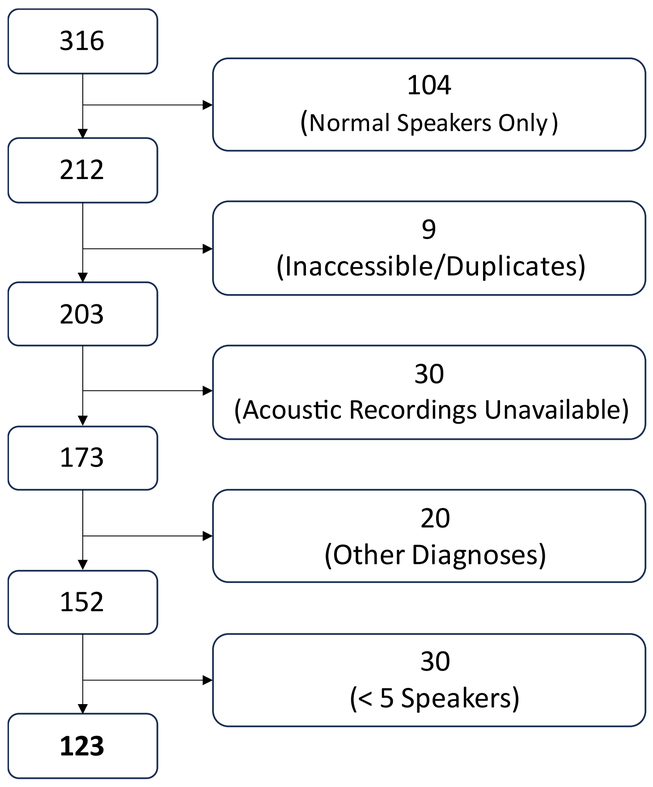
SEARCH STRATEGY

- Queried large AI repositories (e.g., Linguistics Data Consortium, UC Irvine Machine Learning Repository, GitHub, Mendeley Data, etc).
- Queried Google Scholar for disease-specific AI review articles and cross-referenced articles mentioned to identify additional datasets.
- Search terms included various combinations of the following keywords
  - "voice disorders", "mood disorders", "neurological disorders", "speech dataset", "voice dataset", "machine learning", and/or "artificial intelligence"

SCREENING

- Includes health data related to the following diagnostic categories: *voice disorders, mood disorders, neurological disorders*
- Is open access, with or without barriers (Khan et al, 2021)
- Includes accessible voice recordings
- Includes more than 5 speakers

Figure 1: *Diagram of search and screening results*



FINAL DATASET

- 123 datasets were analyzed for sample size, demographics, languages represented, and speech tasks
- Data were grouped by diagnostic category, including mood disorders, voice disorders, neurological disorders, and voice/neurological disorders (i.e. Parkinson's Disease, which represents a significant overlap between these two diagnostic categories.

RESULTS

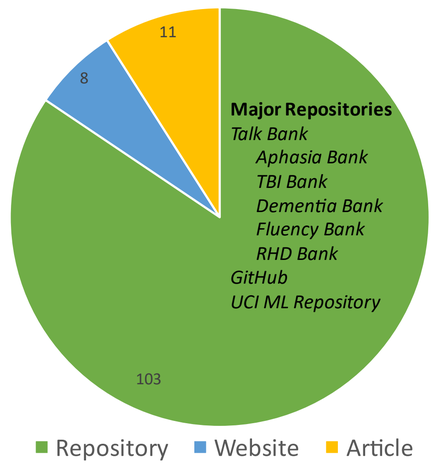
What datasets are available?

Figure 2: *Distribution of datasets across diagnostic category*

Datasets	Voice Disorders	Voice/Neuro Disorders	Neuro Disorders	Mood Disorders	All Disorders
Total Datasets:	11	13	89	9	122
Sample Size: Range	10 - 2,041	14 - 5,826	6 - 1,551	24 - 7,596	6 - 7,596
Median	208	65	22	66.5	29

Where can you find these datasets?

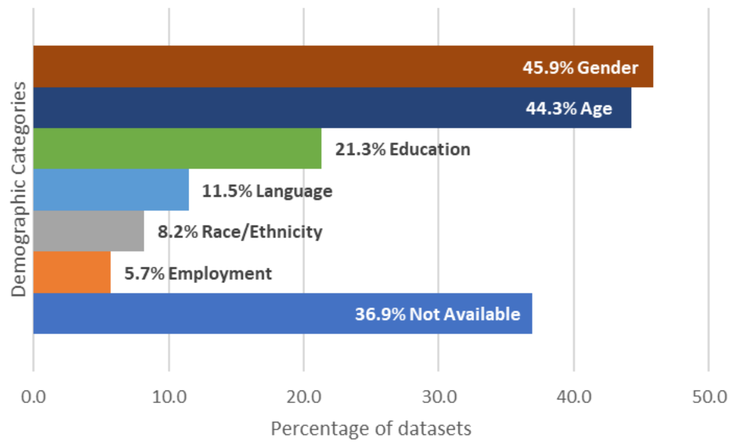
Figure 3: *Dataset sources*



What do these datasets include?

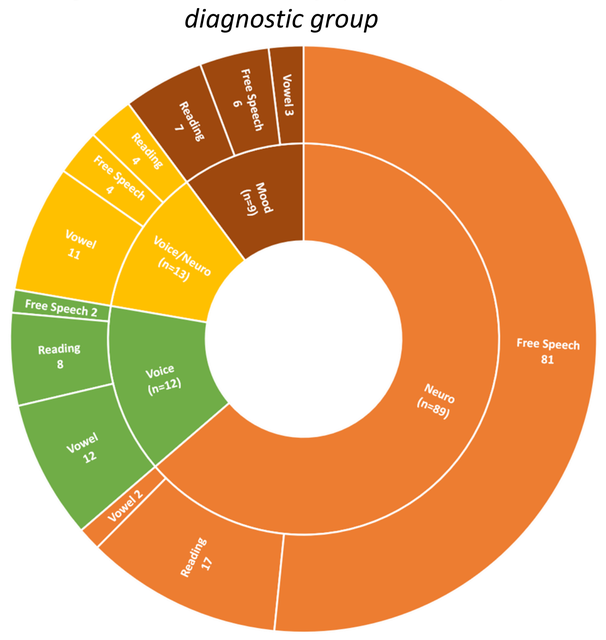
The following results are based on available data provided on the dataset's landing page and/or in the associated article published using the dataset.

Figure 4: *Proportion of datasets reporting demographic data*



- Age and gender are reported in less than 50% of datasets.
- Roughly 37% of datasets do not report demographics collected.

Figure 5: *Distribution of speech tasks by diagnostic group*



- ***Free speech*** tasks include conversations, interview data, group therapy sessions, fluency tasks, etc.
- ***Vowel*** tasks include sustained phonation of /a/, /i/, /o/
- ***Reading*** tasks include words, phrases, or longer reading passages

- Tasks are traditionally hypothesis-driven; however, overlap between diagnostic categories exist:
  - Sustained phonation is utilized across all diagnostic categories, and translates across languages
  - Standardized speech passages are used across diagnostic categories as well (e.g. Rainbow Passage, North Wind and the Sun, etc.)

Figure 6: *Languages represented across diagnostic categories*

LANGUAGE	Voice Disorders		Voice/Neurological Disorders		Neurological Disorders		Mood Disorders		All Disorders	
	n	%	n	%	n	%	n	%	n	%
NA (vowel only)	3	27%	7	54%	1	1%			11	9%
English	3	27%	2	15%	68	76%	4	44%	77	63%
Afrikaans					1	1%			1	1%
Arabic	1	9%							1	1%
Chinese			1	8%	4	4%	3	33%	8	7%
Croatian					1	1%			1	1%
Dutch					2	2%			2	2%
French	1	9%			2	2%			3	2%
German	1	9%			1	1%	1	11%	3	2%
Greek					4	4%			4	3%
Italian	1	9%	1	8%	1	1%			3	2%
Portuguese	1	9%							1	1%
Spanish			1	8%	1	1%			2	2%
Taiwanese					1	1%			1	1%
Tamil					1	1%			1	1%
Turkish			1	8%					1	1%
Not Reported					1	1%	1	11%	2	2%
Total N	***11***		***13***		***89***		***9***		***122***	

CONCLUSIONS

- Many publicly-accessible datasets exist which could be leveraged for AI biomedical research.
- Protocols including sustained phonation, standardized speech passages, and standardized tasks to elicit free speech (e.g. Cookie Thief task) may allow for expanded use of data across diagnostic categories.
- Dataset developers should consider summarizing protocols and demographic data on the dataset landing page for easy evaluation.