Augmented Datasheet for Speech Datasets

Dataset	Paper	Link
WHAM!	[1907.01160] WHAM!: Extending Speech Separation to Noisy Environments (arxiv.org)	WHAM! (whisper.ai)

1. Motivation

What is the speech dataset name, and does the name accurately describe the contents of the dataset?

WSJ0 Hipster Ambient Mixtures (WHAM!); Yes, as the dataset is a mixture of sets of two speakers from the WSJ dataset with a unique noise background scene, recorded from non-stationary ambient environments such as coffee shops. The dataset is created to work on speech separation from audio with real ambient background noise.

Can the dataset be used to draw conclusions on read speech, spontaneous speech, or both?

Read speech (from the original WSJ dataset, which is based on reading English language Wall Street Journal news)

Describe the process used to determine which linguistic subpopulations are the focus of the dataset.

N/A; there were no linguistic subpopulations that were mentioned as being the focus of the datasets. The authors would be able to answer this question.

2. Composition

How many hours of speech were collected in total (of each type, if appropriate), including speech that is not in the dataset? If there was a difference between collected and included, why?

For background noise, the authors collected approximately 80 hours of audio recorded at 44 different locations; approximately 5% of the data was discarded as it contained high SNR intelligible speech. The removal process was done by processing the ambient recordings with the iZotope RX 7 Dialogue Isolate tool to estimate foreground speech, removing clips that had an estimated SNR < -6 dB.

How many hours of speech and number of tokens are in the dataset (by each type, if appropriate)?

81.68 hours (58.03 hours of training, 14.65 hours validation, 9.0 hours test) [Reflects number on website, not paper]

Are there standardized definitions of linguistic subpopulations that are used to categorize the speech data? How are these linguistic subpopulations identified in the dataset and described in the metadata?

N/A; there were no linguistic subpopulations that were mentioned as being the focus of the datasets. The authors would be able to answer this question.

For any linguistic subpopulations identified in the dataset, please provide a description of their respective distributions within the dataset.

N/A; there were no linguistic subpopulations that were mentioned as being the focus of the datasets. The authors would be able to answer this question.

How much of the speech data have corresponding transcriptions in the dataset?

WHAM does not provide any transcripts in the dataset. However, it does provide paths to the WSJ0 files which will have transcripts associated with them.

Does the dataset contain non-speech mediums (e.g. images or video)?

No; the dataset contains only speech.

Do speakers code switch or speak multiple languages, and if so, how is this identified in the data?

No; the dataset is a derivative of wsj0-2mix which contains only English speech.

Does the speech dataset focus on a specific topic or set of topics?

No; the dataset is a derivative of wsj0-2mix, which in turn is derived from the WSJ0 corpus. This contains read speech from a corpus of Wall Street Journal news, which can encompass various topics.

Does the dataset include sensitive content that can induce different emotions (e.g., anger, sadness) that can cause the speakers to produce unusual pitch or tone deviating from plain speech?

Not mentioned in the paper. The dataset is derived from read speech from news articles with no explicit instructions on inducing different emotions, but which could induce different emotions in different readers.

Does the dataset contain content that complies to the users' needs, or does it result in symbolic violence (the imposition of religious values, political values, cultural values, etc.)?

No; the dataset contains read speech from news articles which should not contain value-laden text; however, it’s possible that some sentences could be perceived offensively by individuals from different backgrounds or lingual groups.

3. Collection Process

What mechanisms or procedures were used to collect the speech data, e.g.: is the data a new recording of read speech or an interview? Or is it downloaded speech data from public speeches, lectures, YouTube videos or movies, etc.?

The speech data is from the wsj0-2mix dataset which in turn is derived from the Wall Street Journal (WSJ0) corpus. The authors also collected their own background noise dataset to mix with wsj0-2mix.

Were all the data collected using the same technical methodology or setting, including the recording environment (e.g., lab, microphone) and recording information (e.g., sampling rate, number of channels)?

The WSJ-0 corpus was collected in a standardized method with a close-talking Sennheiser HMD414 microphone and a secondary microphone that could vary. The audio was recorded on two channels.

The ambient sound that the authors collected for WHAM was collected in urban environments in the San Francisco Bay Area. They used an Apogee Sennheister binaural microphone connected to a smartphone where the microphone was mounted to a tripod. The audio is captured at 48 kHz but downsampled to 16 kHz and 8 kHz. Both single-channel and stereo recordings are available.

Is there presence of background noise?

Yes, the background noise dataset was recorded in urban environments and mixed into the original wsj0 dataset for data augmentation and speaker separation tasks. The mixtures are created by applying randomly selected gains in order to achieve relative levels between 0 and 5 dB between the two speech signals prior to mixing in the time domain.

For interviewer/interviewee speech data: during the interview process, did interviewers consistently ask questions that are “fair and neutral”?

N/A; the speech data does not come from interview data.

Have data subjects consented to the disclosure of the metadata in the dataset? Also, does the metadata include sensitive personal information such as disability status?

There is no associated metadata for WHAM. However, for WSJ0, which WHAM is derived from, speakers have gender associated as metadata, and a subset of speakers include age or dialect metadata.

4. Preprocessing/cleaning/labeling

When generating the dataset, was any background noise deleted or adjusted to make all recording qualities similar?

No, the purpose of the dataset is to include ambient or background noise.

Did the data collectors hire human annotators to transcribe the data? If so, how trained were the annotators in speech transcription for this context? How familiar were they with the corpus material, the vocabulary used, and the linguistic characteristics of different dialects and accents?

No transcription process was involved in the WHAM dataset as it derived all speech audio from the wsj0-2mix dataset which in turn comes from the WSJ0 dataset. The WSJ0 dataset relied on text data from the Wall Street Journal.

If multiple transcription methods were used, how consistent were the annotators? How were transcripts validated?

N/A; multiple transcription methods were not used.

If the speech data include transcriptions, what software was used to generate the transcriptions (including, e.g., software used by human transcribers)? Are timestamps included in transcriptions? Are the alignments provided with the transcripts?

N/A; no transcription process was involved in the creation of the WHAM dataset.

Were transcription conventions (such as tagging scheme, treatment of hate speech or swear words, etc.) disclosed along with the corpus?

N/A; no transcription was involved in the creation of the WHAM dataset.

Is additional coding performed, separate to transcriptions and tagging?

No

5. Uses / Distribution / Maintenance

How are redactions performed on the dataset? Are personally identifiable information or sensitive information removed from only transcripts, audio censored from the speech data, or both?

No redactions were mentioned to have been performed. The dataset is a derivative of the wsj0-2mix dataset which should not contain personally identifiable information.

Is there any part of this dataset that is privately held but can be requested for research purposes?

No

Is there a sample dataset distributed? If so, how well does the sample represent the actual dataset? Do they include all forms of speech included in the dataset? How big is the sample?

No; only the full dataset is available for download.

Aside from this datasheet, is other documentation available about the data collection process (e.g., agreements signed with data subjects and research methodology?)

Yes, the dataset has a README on their project page detailing how the dataset was generated as well as scripts that allow the dataset user to generate the mixed data themselves.