A List of Audio Datasets for Deep learning

Individual Study Report
Spring 2021

Aditya Dutt
Department of Computer Science
University of Florida
Gainesville, FL

Submitted To: Dr. Paul Gader

I. Introduction

All sound-based machine learning models rely on large amounts of audio data. There are several datasets focused on speech, environmental sounds, music, etc. These datasets are not easily accessible on a single website. So, this report aims to provide a list of all audio datasets along with a short description and their URL. It contains datasets that can be used for gender detection, automatic speech recognition, music instrument detection, bird sound detection, speaker recognition, music genre recognition, emotion classification, etc. It would be useful for choosing the appropriate dataset for a given task.

II. List of datasets

- **Arabic Speech Corpus** - The Arabic Speech Corpus is a speech corpus developed for speech synthesis. The synthesized speech as an output using this corpus has produced a high quality, natural voice. The corpus contains phonetic transcriptions of more than 3.7 hours. It contains 1813 wav files containing spoken utterances and the phoneme labels with time stamps of the boundaries where these occur in the .wav files.
- **AudioMNIST** - The dataset consists of 30000 audio samples of spoken digits (0-9) of 60 different speakers. Metadata also contains age and gender of each speaker.
- **Common Voice** - Common Voice is used for speech recognition. You can contribute your voice by reading a sentence. It is 12GB in size and contains spoken text based on text from a number of public domain sources like user-submitted blog posts, old books, movies, and other public speech corpora.
- **CHIME** - It is a noisy speech recognition dataset which is around 4GB in size. It contains recordings of 4 speakers. There are 9000 recordings over 4 noisy locations.
- **CREMA-D** - CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages 20 and 74 coming from a variety of races and ethnicities. The sentences spoken were presented using one of six different emotions: Anger, Disgust, Fear, Happy, Neutral and Sad. Emotion level is also mentioned (Low, Medium, High and Unspecified).
- **DAPS Dataset** - DAPS dataset consists of 20 speakers (10 female and 10 male) reading 5 excerpts each from public domain books (which provides about 14 minutes of data per speaker).
- **Deep Clustering Dataset** – It is created to training deep discriminative embeddings to solve the cocktail party problem.
- **DIPCO** (Dinner Party Corpus by Amazon) - The dataset simulates a dinner party scenario taking place in everyday homes. The recordings are of 4 amazon employees having a natural conversation in English. The dataset contains the audio recordings and human labeled transcripts of a total of 10 sessions with a duration between 15 and 45 minutes.
- **EmoV DB** – It contains emotions of male and female speakers with different emotions. Its purpose was to synthesize emotional voice.
- **Emotional Voice dataset** – It contains 2,519 speech samples by 100 actors from 5 different cultures.
- **Free Spoken Digit Dataset** – It contains 6 speakers, 3,000 recordings (50 of each digit per speaker) with English pronunciations.

- **Flickr Audio Caption** – It contains 40,000 spoken captions of 8,000 natural images. It is 4.2 GB in size. It was collected in 2015 to investigate multimodal learning schemes for unsupervised speech pattern discovery.
- **ISOLET Data Set** - It helps to predict which letter was spoken. It can be a simple classification problem. Its size is around 38.7 GB.
- **Libriadapt** - It is designed to facilitate domain adaptation research for ASR models and contains the domain shifts in the data due to microphones, speaker accents, and acoustic environment.
- **Libri-CSS** - Continuous speech separation (CSS) is an approach to handling overlapped speech in conversational audio signals. It derived from *LibriSpeech* by concatenating the corpus utterances to simulate a conversation and capturing the audio replays with far-field microphones.
- **LibriMix** - LibriMix is an open-source dataset for source separation in a noisy environment. Based on *LibriSpeech* signals (pure subset) and WHAM audio. (It will also enable cross-dataset testing).
- **Librispeech** - LibriSpeech is an organization of about 1000 hours of 16Khz reading English speech from textbooks from the *LibriVox* project.
- **LJ Speech** - It contains 13,100 short audio clips of a single speaker. Passages are read from 7 non-fiction books. A transcription also exists for each clip. Clips are of 1 to 10 seconds length.
- **Microsoft Scalable Noisy Speech Dataset** - This dataset consists of large collections of clean speech files and variety of environmental noise files. It is sampled at 16 kHz. A great application of this database is to train in deep learning models to compress background sound.
- **Multimodal EmotionLines Dataset (MELD)** - The Multimodal EmotionLines Dataset (MELD) was created by developing and extending the EmotionLines database. MELD contains the same dialogue conditions found in EmotionLines, but also includes audio and visual and text captions. MELD has more than 1400 dialogues and 13000 voices from the Friends TV series. Each sentence in a conversation is labeled: Disgust, Sadness, Anger, Surprise, Neutrality, Joy and Fear.
- **NISQA Speech Quality Corpus** – It contains 14000 speech samples with simulated (codecs, packet-loss, background noise) and live (mobile phone, Zoom, Skype, WhatsApp) voice call degradation conditions. All files are labelled with ratings of the overall quality in terms of Noise, Coloration, Discontinuity, and Loudness.
- **Noisy Dataset**: This database was designed to train and test speech enhancement methods that operate at 48kHz. Speech samples are from VCTK dataset.
- **Parkinson's speech dataset** - The training data is from 20 Parkinson’s Disease (PD) patients and 20 healthy people.
- **Persian Consonant Vowel Combination (PCVC) Speech Dataset** - The Persian Consonant Vowel Combination (PCVC) Speech Dataset is a Modern Persian speech corpus for speech recognition and speaker recognition. It contains 6 vowels and 23 Persian consonants. The sound samples contain all possible combinations of vowels and consonants (138 samples for each speaker).
- **The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** - The Ryerson Audio-Visual Database of Emotional Speech and Song (*RAVDESS*) contains 7356 files (total size: 24.8 GB). This database contains samples from 24 professional actors (12 female, 12 male) vocalizing two lexically matched statements in a neutral North American accent. The speech contains angry, happy, sad, calm, surprise, fearful, and disgust expressions. Similar emotions are for songs.
- **SAVEE Dataset** – It contains samples from 4 male actors in 7 different emotions. It consists of 480 British English utterances.
- **SparseLibriMix** – It is a dataset developed for source separation in noisy environments and with variable overlap-ratio.

- **Speech Accent Archive** – It can be used for accent detection problems.
- **Speech Commands Dataset** - This dataset (1.4 GB) has 65,000 one-second-long utterances of 30 words, by thousands of different people.
- **Spoken Commands dataset** – It contains free audio samples which is around 10M words. It can be used for voice activity detection and recognition of syllables (single-word commands). It contains data from 3 speakers, 1,500 recordings (50 of each digit per speaker) with English pronunciations.
- **Spoken Wikipedia Corpora** – It is a 38 GB dataset available in both audio and without audio format.
- **Ted-LIUM** - The TED-LIUM corpus was made from talks recordings and their transcriptions are available on the TED website (noncommercial).
- **Thorsten dataset** – It is a German language dataset. It contains 22,668 recorded phrases, 23 hours of audio, and phrase length of 52 characters on average.
- **TIMIT dataset** - TIMIT contains recordings of 630 speakers of eight major dialects of American English. It includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance.
- **VCTK dataset** – It contains 110 English speakers with various accents. Each speaker reads out about 400 sentences. Samples are mostly 2-6 seconds long, at 48 kHz 16 bits, for a total dataset size of around 10 GB.
- **VCTK-2Mix** - VCTK-2Mix is an open-source dataset for source separation in noisy environments. It is derived from *VCTK* signals and *WHAM* noise. It will also enable cross-dataset experiments.
- **VoxCeleb** - VoxCeleb is a large-scale speaker identification dataset. It contains around 100,000 utterances by 1,251 celebrities, extracted from You Tube videos. It's a useful dataset for isolating and identifying which speaker the voice belongs to.
- **VoxForge** - VoxForge was set up to collect transcribed speech for use with Free and Open-Source Speech Recognition Engines.
- **VoxPopuli** - VoxPopuli dataset contains 100K hours of unlabelled speech data in 23 languages. It has 1.8K hours of transcribed speech data in 16 languages.
- **WHAM! and WHAMR!** - The WSJ0 Hipster Ambient Mixtures (WHAM!) dataset pairs each two-speaker mixture in the wsj0-2mix dataset with a unique noise background scene. WHAMR! is an extension to WHAM! that adds artificial reverberation to the speech signals in addition to the background noise. The noise audio was collected at various urban locations throughout the San Francisco Bay Area in late 2018. The environments primarily consist of restaurants, cafes, bars, and parks. Size of WHAM! dataset: 17.65 GB unzipping to 35 GB.
- **Zero Resource Speech Challenge** - The goal of the Zero Resource Speech Challenge is to construct a system that learns an end-to-end Spoken Dialog (SD) system, in an unknown language, from scratch, using only information available to a language learning infant. "Zero resource" refers to zero linguistic expertise (e.g., orthographic/linguistic transcriptions), not zero information besides audio (visual, limited human feedback, etc.). The fact that 4-year-olds spontaneously learn a language without supervision from language experts show that this goal is theoretically reachable.
- **AudioSet** – It consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. (The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds).

- **Bird audio detection challenge** - This challenge contained new datasets (5.4 GB) collected in real live bioacoustics monitoring projects, and an objective, standardized evaluation framework. Detecting bird noises is an important task for automatic wildlife monitoring.
- **Free Music Archive** - It is a dataset for music analysis. Its size is 1000 GB. It can used for MIR tasks.
- **Karoldvl-ESC** - The ESC-50 dataset is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification.
- **Million Song Dataset** - The Million Song Dataset is a freely available collection of audio features and meta-data for a million contemporary popular music tracks. 280 GB in size.
- **MUSDB18** – It contains Multi-track music dataset for music source separation. 150 tracks (22 Gb).
- **Public domain sounds** - Good for wake word detection; a wide array of sounds that can be used for object detection research (524 MB - 635 SOUNDS - Open for public use).
- **RSC Sounds** - RSC sounds from RuneScape Classic (8-bit, u-law encoded, 8000 Hz pcm samples).
- **Urban Sound Dataset** – It contains two datasets *UrbanSound* and *UrbanSound8K*. Urban Dataset contains 1302 labeled sound recordings. Each recording is labeled with the start and end times of sound events from 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music.

Additional GitHub repositories and websites:

- **Awesome_Diarization** - A list of Speaker Diarization papers, libraries, datasets, and other resources. Diarization is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity. It is github repository.
- **ASR datasets** - A list of publicly available datasets for ASR purposes and other speech activities.
- **Environmental audio dataset** - Audio data collection and manual data annotation both are tedious processes, and lack of proper development dataset limits fast development in the environmental audio research. This page tries to maintain a list of datasets suitable for environmental audio research.
- Kaggle contains several datasets. Many audio datasets can be found here.

References

- Neural Arabic Text Diacritization: State of the Art Results and a Novel Approach for Machine Translation, Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh and Mahmoud Al-Ayyoub, EMNLP-IJCNLP 2019.
- *J.* Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.
- A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild Computer Science and Language, 2019
- J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018.
- A. Nagrani, J. S. Chung, A. Zisserman VoxCeleb: a large-scale speaker identification dataset INTERSPEECH, 2017.
- Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.
- S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. ACL 2019
- Chen, S.Y., Hsu, C.C., Kuo, C.C. and Ku, L.W. EmotionLines: An Emotion Corpus of Multi-Party Conversations. arXiv preprint arXiv:1802.08379 (2018)
- S. Zahiri and J. D. Choi. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In The AAAI Workshop on Affective Content Analysis, AFFCON'18, 2018.
- S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation. ACL 2019.
- Zhang, Yazhou, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. "Quantum-Inspired Interactive Networks for Conversational Sentiment Analysis." IJCAI 2019
- Zhang, Dong, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. "Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations." IJCAI 2019
- Ghosal, Deepanway, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation." EMNLP 2019.
- Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Serra. Audio tagging with noisy labels and minimal supervision. In Proceedings of DCASE2019 Workshop, NYC, US (2019).
- Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), pp 486-493. Suzhou, China, 2017.
- https://lionbridge.ai/datasets/12-best-audio-datasets-for-machine-learning/
- https://lionbridge.ai/datasets/voice-and-sound-data-for-machine-learning/
- https://github.com/jim-schwoebel/voice_datasets
- http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/
- https://homepages.tuni.fi/toni.heittola/datasets.html
- https://urbansounddataset.weebly.com/urbansound.html
- https://commonvoice.mozilla.org/en/datasets
- https://sigsep.github.io/datasets/dsd100.html
- https://github.com/robmsmt/ASR_Audio_Data_Links