Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7253–7258
Marseille, 20-25 June 2022

Speak: A Toolkit Using Amazon Mechanical Turk to Collect and Validate Speech Audio Recordings

Christopher Song[1], David Harwath[2], Tuka Alhanai[3], James Glass[4]
[1]Johns Hopkins University
Baltimore, MD 21218, USA
[2]University of Texas at Austin
Austin, TX 78712, USA
[3]New York University Abu Dhabi
Abu Dhabi, UAE
[4]Massachusetts Institute of Technology
Cambridge, MA 02139, USA
[1]csong23@jhu.edu, [2]harwath@cs.utexas.edu, [3]tuka.alhanai@nyu.edu, [4]glass@mit.edu

Abstract

We present Speak, a toolkit that allows researchers to crowdsource speech audio recordings using Amazon Mechanical Turk (MTurk). Speak allows MTurk workers to submit speech recordings in response to a task prompt and stimulus (e.g. image, text excerpt, audio file) defined by researchers, a functionality that is not natively offered by MTurk at the time of writing this paper. Importantly, the toolkit employs multiple measures to ensure that speech recordings collected are of adequate quality, in order to avoid accepting unusable data and prevent abuse/fraud. Speak has demonstrated utility, having collected over 600,000 recordings to date. The toolkit is open-source and available for download.

Keywords: data collection, speech recognition, human-computer interaction, crowdsourcing

1. Introduction

Speech audio recordings remain an important source of data used in academic research. They are used in a variety of academic disciplines, including speech recognition, human computer interaction (Clark et al., 2019), neuroscience (Anumanchipalli et al., 2019; Akbari et al., 2019), psychology (Pouw et al., 2020), and medicine (Anfinrud et al., 2020; Karan et al., 2020; Alhanai et al., 2017). Given the continued relevance of speech data to academia, attempts to scale its collection are not new; researchers have collected speech recordings in large-scale efforts since the early 1990's (Hirschman et al., 1993), but the challenge of collecting this data in a way that is both time- and resource-efficient continues to be a challenge.

Technological development has both helped and sustained the issue. It has contributed to the evolution of collection efforts: from manual collection across collaborating institutions (Hirschman et al., 1993), to telephone-based collection (Zue et al., 1997), and eventually to web-based collection (e.g. through open-source data, crowdsourcing). Crowdsourcing has particularly shown promise as an efficient way to collect these recordings, with tools having previously been built (Saylor, 2015; McGraw et al., 2010) to source recordings from mainstream crowdsourcing platforms such as Amazon Mechanical Turk (MTurk).

However, as some web technologies continue to develop, others stagnate. The dependencies underpinning these previous crowdsourcing tools have lost support to the point of deprecation, thereby rendering the tools difficult to use for data collection. Researchers have continued to adopt other previously-established methods for speech data collection, such as soliciting recordings manually, or extracting them from open-source data; such solutions indeed have their own benefits, but they come with their own unique drawbacks as well (e.g. significant additional secondary efforts required, lack of task-specificity, etc.)

Crowdsourcing as a collection schema continues to provide potential solutions to these challenges; so long as it is relatively easy to implement, its economic and time-saving benefits cannot be ignored (McGraw et al., 2010). Speak was developed using contemporary web-based tools in order to bring these benefits back into the hands of researchers – and capitalize on modern developments in the process, to gain additional benefits (e.g. expanded browser compatibility, automated validation tasks, and improved user experience for researchers and workers alike). And it has realized these benefits in practice; Speak has been used to collect over 600,000 speech recordings (Hsu et al., 2021) in support of academic research efforts.

2. Related Works

The works most related to the Speak tool are studies on crowdsourcing tools developed for speech data collection. They can be split into two broad categories: tools which collect recordings using mainstream crowdsourcing platforms, and tools which collect recordings through independently designed crowdsourcing platforms.

Studies involving similar tools do exist that are designed to use mainstream crowdsourcing platforms,

usually also MTurk. The WAMI toolkit (Gruenstein et al., 2008), developed in 2008, is a speech crowdsourcing tool that worked in conjunction with MTurk. Unfortunately, the AudioController, WAMI's component responsible for client-side audio recording/transmission, was built as an embedded Java applet, later rebuilt in Flash, both of which have been phased out - thereby rendering the tool obsolete. Patricia Saylor's Spoke (Saylor, 2015), developed in 2015, was built to replace WAMI using modern Web Speech APIs. Unfortunately, its dependencies ultimately experienced a similar fate.

Other crowdsourcing tools have been built to collect speech recordings using independently designed platforms (Ardila et al., 2020; Warden, 2017). As they do not use mainstream, dedicated crowdsourcing platforms to collect their recordings, they require substantial additional efforts coordinating outreach and compensation (when/if applicable).

Other studies have collected speech data through a variety of methods other than crowdsourcing, including manual collection, and extraction through open-source data.

Manual collection involves researchers use a dedicated device/environment to record participants completing the desired speech task (Sakar et al., 2013; Burkhardt et al., 2005; Leonard and Doddington, 1991). It is inherently limited by the location and time/resources of the research team, and it requires significant amounts of secondary efforts, including participant outreach, variance of recording environments (Barker et al., 2018; Avila et al., 2019), and/or data post-processing (Mysore, 2014). These factors may pose logistical challenges when collecting speech datasets at scale. In addition, manual collection faces the challenge of capturing recordings with devices/environments representative of those used by users in the real world. Previous studies have addressed this in a variety of ways, such as collecting the data at participants' homes (Barker et al., 2018), adding "real-world noise" when post-processing the data (Mysore, 2014), or otherwise deliberately varying their recording environments (Avila et al., 2019) – all of which further require more time and/or resources to accomplish.

Extraction through open-source data, such as TED Talks (Afouras et al., 2018) and public domain audiobooks (Panayotov et al., 2015; Snyder et al., 2015), can benefit from the vast amount of data available online for collection. However, similar to manual collection, this data may not be representative of real-world recordings; in addition, collecting data from these sources sacrifices task-specificity, as the subject matter of the recordings is generally not targeted to any specific task.

By using the crowdsourcing model, the Speak tool allows researchers to collect task-specific speech data in real-world environments with limited secondary effort. Built on modern libraries, Speak overcomes the deprecation of previous tools, and uses a mainstream crowdsourcing platform to leverage its established user base and incentive structure.

3. Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) is a common tool used for crowdsourcing operations, with over 250,000 workers worldwide, as of 2019 (Robinson et al., 2019). Workers complete crowdsourcing tasks (called Human Intelligence Tasks, or HITs) for financial compensation. MTurk allows researchers to limit which workers can complete their HITs by defining worker eligibility requirements such as geographic location, demographic attributes, and previous experience (Services, 2017). At the time of writing this publication, MTurk does not natively offer researchers the ability to collect speech data from its participants. The Speak tool allows for this collection of speech data from MTurk workers, by embedding a web application as an HTML element in a custom HIT.

4. Tool Overview

The Speak tool allows researchers to present participants (MTurk workers) with a user-friendly portal that shows a stimulus/prompt (a photo, video, or text) provided by the researcher, as well as an in-browser tool used to submit speech audio recordings. The tool consists of the following components:

- A suite of Python (Oliphant, 2007) scripts using the Amazon Web Services Boto3 API (Garnaat, 2018) to deploy/delete custom HITs, view/save HIT attempt logs, and accept/reject HIT attempts.
- A Flask web application to handle all stimulus display, speech recording, validation, and data storage functions. The web application is deployed using uWSGI and nginx (Relan, 2019) in order to handle higher traffic loads, and is compatible with all versions of Google Chrome, Firefox, Edge, Safari, and Opera.

The Speak tool automates tasks within this collection process, in order to improve the researcher and user experience. Figure 1 outlines the process flow underpinning toolkit operations, and maps the participant's and researcher's experiences throughout usage of the tool. At a high level, the data collection process using Speak can be broken into three phases:

1. **Deployment**: a researcher configures the Speak tool to display media from their stimulus dataset and use their desired web server. The researcher also configures the MTurk scripts to communicate with their instance of the Speak tool. They deploy the web application on their server, then embed its URL into custom HITs, which they deploy onto MTurk. At the end of this phase, the server is on, the HITs are deployed, and the researcher is ready to collect speech recordings from MTurk workers.

2. **Collection**: participants log onto MTurk, complete the HITs, and submit them for approval. The Speak tool automatically ensures that all collected speech recordings pass a series of quality checks before the participants can submit their HIT attempts for approval. The researcher views logs of all HIT attempts, and accepts them to compensate participants. This phase continues until the researcher has collected their desired number of speech recordings.

3. **Conclusion**: The researcher removes HITs from MTurk, and shuts down the Speak server. At the end of this phase, the HITs are no longer visible to MTurk workers.

5. Highlighted Features

5.1. Stimuli

The Speak web application can display a series of stimuli for the MTurk worker, if a stimulus is required as part of the data collection prompt. Within a series, each stimulus is presented individually, so that the worker can focus on one at a time when recording their responses. The number of stimuli in each series is set by the researcher. Stimuli can be in the form of photos, videos, audio, or text.

5.2. Data Storage

All data collected through the Speak tool is stored such that the file structure of the speech recordings directory mimics that of the stimuli dataset. For each stimulus file, a folder named after the stimulus is created, and each folder contains files for all recordings associated with that stimulus, as well as an automatically generated transcript for each recording. The naming conventions for the files are as follows:

- Recording: `(workerID)_(attemptID).wav`
- Transcript: `(workerID)_(attemptID)_transcript.txt`

5.3. Audio Recorder

During the HIT, Speak displays a customizable set of instructions at the top of the screen, and a simple audio recorder (Diamond, 2016) on the bottom of the screen, consisting of a sound meter (Wilson, 2017) and a "Record" button. The instructions can be customized according to the researcher's specific task. The sound meter is an orange bar above the "Record" button that moves in response to the amplitude of the worker's voice, so the worker can adjust their mic positioning before and during recording. The meter turns blue during recording, and red if the worker is speaking too loudly. Figure 2 shows an illustrative example of what an MTurk worker would see while completing a Speak HIT.

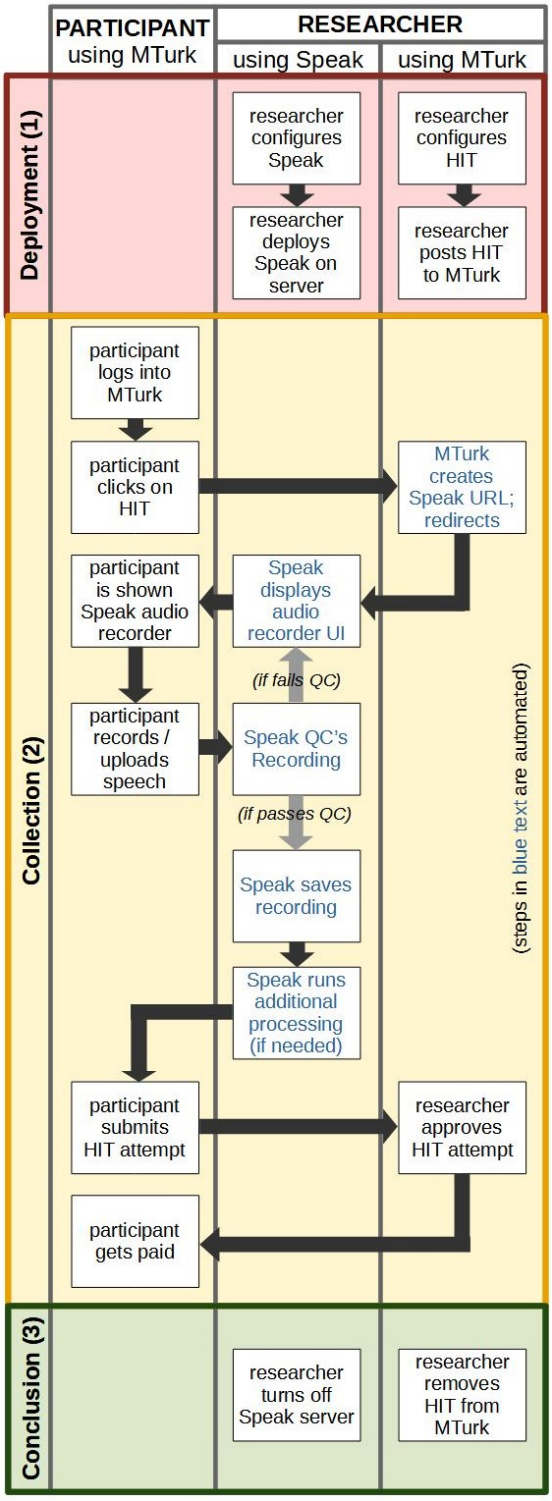

Figure 1: Swimlane process diagram showing the roles of all users and platforms during the speech data collection process through Speak and MTurk. Sections are color-coded by phase (i.e. deployment, collection, conclusion), and individual lanes are separated by user and platform (i.e. Speak or MTurk). All steps with blue text denote fully automated processes.

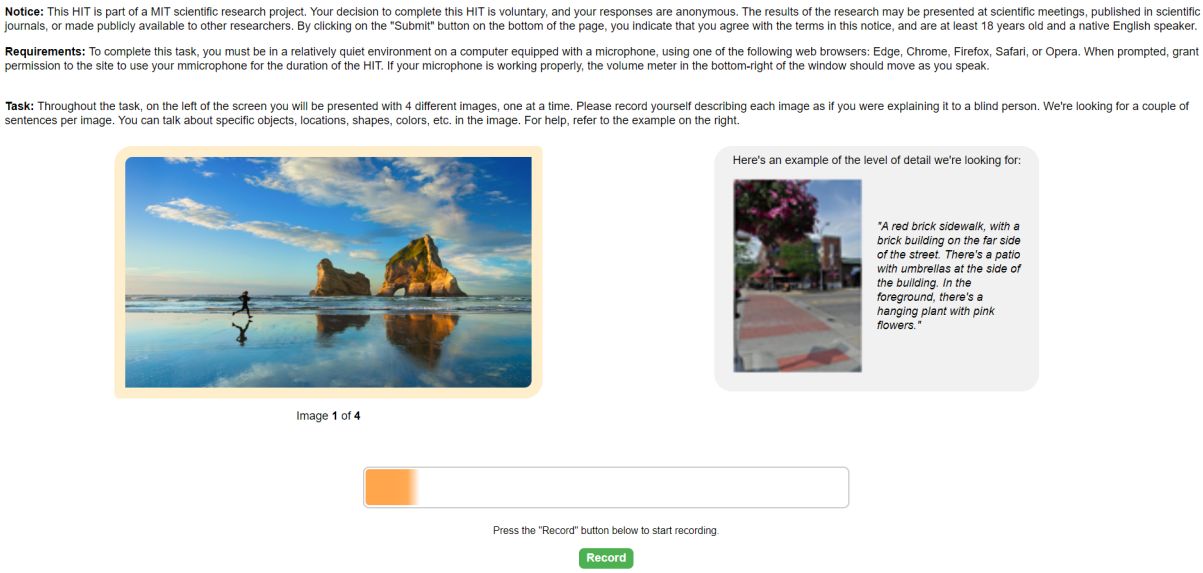

Figure 2: A screenshot displaying the Speak tool audio recorder. In this example, the stimulus (bottom left) is a photo, and there are four stimuli in the HIT. The gray box containing the example stimulus and transcript (bottom right) disappears as soon as the worker presses the "Record" button.

5.4. Example Transcript Display

In order to ensure the MTurk worker fully understands the task, Speak displays an example stimulus, as well as a transcript of what an acceptable recording would contain. The examples are customizable. When the user presses the "Record" button, the example stimulus and transcript disappear, in order to discourage the worker from simply reading the example transcript for their recording submission.

5.5. Quality Control/Validation

Speak has checks in place to mitigate MTurk worker attempts to submit poor quality recordings. Upon failure of these quality checks, the MTurk worker is politely asked to check their mic and try submitting another recording. The exact quality tests are conducted completely in the backend, and are not shown to workers, in order to prevent adversarial attempts to thwart them. Quality checks are at the individual recording level, as well as at the overall HIT attempt level.

- **Recording-level:** Speak transcribes audio using Google Speech Recognition API, checks that the recording contains above a certain number of words (e.g. 5 words), and that the recording is longer than a certain length (e.g. 3 seconds).
- **HIT attempt-level:** The Speak app checks that the worker spent above a certain amount of time overall on the HIT attempt (e.g. 20 seconds), and that the worker passed recording-level validation checks for each final recording.

5.6. Logs

HIT activity is stored in MTurk logs and displayed real-time in the Speak server console. Logs saved on MTurk servers are accessible using the Boto3 API. Using Speak tool scripts, log data is also dumped and into JSON files and saved upon approval of HITs, if desired. Table 1 contains all of the raw log information collected for each HIT attempt.

6. Conclusion

This paper presents the Speak tool, a web application which allows researchers to collect speech audio recordings and solicit/compensate volunteers through Amazon Mechanical Turk. Previous similar tools have greatly impacted the speed and scale of research efforts requiring task-specific speech audio data, and we believe Speak has the potential to extend this impact; to this end, we are releasing this tool open-source on GitHub[1]. The tool has demonstrated utility, having been used to collect over 600,000 speech recordings. It is our hope that by further reducing the administrative and economic burdens of collecting speech recordings, the Speak tool can lower the barriers to speech data collection, and give researchers more freedom to better focus on developing novel ideas.

[1]Available at `https://github.com/soupdtag/speak-tool`.

Field name	Description
hit_id	HIT attempt ID
worker_id	MTurk worker ID
datetime_completed	Date and time attempt was submitted
elapsed_time	Amount of time spent completing the HIT attempt
probably_not_fraud	Quality check: whether 'elapsed time' was above a threshold set by the researcher ('True' or 'False')
worker_ip	IP address of MTurk worker
worker_country	Country of MTurk worker, found from IP address
worker_region	Region of MTurk worker, found from IP address
worker_city	City of MTurk worker, found from IP address
test_idx	An index number used to identify the series of stimuli presented to the MTurk worker
test_passed	Quality check: whether all recordings submitted by the MTurk worker worker all quality checks ('True' or 'False')
questions_passed	The above quality check, but with the results listed for each individual stimulus
question_0_img	File location of the stimulus used for Question 0
question_0_rec	File location of the recording used for Question 0
question_0_transcript_loc	File location of the recording transcript, as a .txt file, submitted by this MTurk worker for Question 0
question_0_transcript	The above transcript, as a string
...	*(continues for N questions)*
question_N_img	File location of the stimulus used for Question N
question_N_rec	File location of the recording used for Question N
question_N_transcript_loc	File location of the recording transcript, as a .txt file, submitted by this MTurk worker for Question N
question_N_transcript	The above transcript, as a string

Table 1: A list of all fields within a log entry for a single HIT attempt.

7. Bibliographical References

Afouras, T., Chung, J. S., and Zisserman, A. (2018). Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1):874, Jan.

Alhanai, T., Au, R., and Glass, J. (2017). Spoken language biomarkers for detecting cognitive impairment. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 409–416.

Anfinrud, P., Stadnytskyi, V., Bax, C. E., and Bax, A. (2020). Visualizing speech-generated oral fluid droplets with laser light scattering. *New England Journal of Medicine*, 382(21):2061–2063.

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, Apr.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*.

Avila, A. R., Gamper, H., Reddy, C., Cutler, R., Tashev, I., and Gehrke, J. (2019). Non-intrusive speech quality assessment using neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635. IEEE.

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-September:1561–1565.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., and R Cowan, B. (2019). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4):349–371, 09.

Diamond, M. (2016). Recorder.js: A plugin for

recording/exporting the output of web audio api nodes. https://github.com/mattdiamond/Recorderjs.

Garnaat, M. (2018). boto documentation. https://buildmedia.readthedocs.org/media/pdf/boto/latest/boto.pdf.

Gruenstein, A., McGraw, I., and Badr, I. (2008). The wami toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 141–148.

Hirschman, L., Bates, M., Dahl, D., Fisher, W., Garofolo, J., Pallett, D., Hunicke-Smith, K., Price, P., Rudnicky, A., and Tzoukermann, E. (1993). Multi-site data collection and evaluation in spoken language understanding. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Hsu, W.-N., Harwath, D., Miller, T., Song, C., and Glass, J. (2021). Text-free image-to-speech synthesis using learned segmental units. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5284–5300, Online, August. Association for Computational Linguistics.

Karan, B., Sahu, S. S., and Mahto, K. (2020). Parkinson disease prediction using intrinsic mode function based features from speech signal. *Biocybernetics and Biomedical Engineering*, 40(1):249 – 264.

Leonard, R. G. and Doddington, G. R. (1991). A speaker-independent connected-digit database. *Instruments Incorporated, Central Research Laboratories, Dallas, TX*, 75266.

McGraw, I., Lee, C.-y., Hetherington, I. L., Seneff, S., and Glass, J. R. (2010). Collecting voices from the cloud. In *LREC*, pages 1576–1583.

Mysore, G. J. (2014). Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Pouw, W., Harrison, S. J., and Dixon, J. A. (2020). Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, 149(2):391.

Relan, K. (2019). Deploying flask applications. In *Building REST APIs with Flask*, pages 159–182. Springer.

Robinson, J., Rosenzweig, C., Moss, A. J., and Litman, L. (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PloS one*, 14(12):e0226394.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., and Kursun, O. (2013). Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834.

Saylor, P. (2015). Spoke: A framework for building speech-enabled websites. Master's thesis, Massachusetts Institute of Technology.

Services, A. W. (2017). Amazon mechanical turk developer guide. https://docs.aws.amazon.com/AWSMechTurk/latest/\AWSMechanicalTurkRequester/Concepts_Qualifications\Article.html.

Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Warden, P. (2017). Speech commands: A public dataset for single-word speech recognition. *Dataset available from http://download. tensorflow. org/data/speech_commands_v0*, 1.

Wilson, C. (2017). Simple volume meter. https://github.com/cwilso/volume-meter.

Zue, V. W., Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R., and Schmid, P. (1997). From interface to content: Translingual access and delivery of on-line information. In *Fifth European Conference on Speech Communication and Technology*.