INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

CLAC: A Speech Corpus Of Healthy English Speakers

R'mani Haulcy, James Glass

Computer Science and Artificial Intelligence Laboratory
Massachussetts Institute of Technology
Cambridge, MA 02139, USA
{rhaulcy,glass}@mit.edu

Abstract

This paper introduces the Crowdsourced Language Assessment Corpus (CLAC), a speech corpus consisting of audio recordings and automatically-generated transcripts for several speech and language tasks, as well as metadata for each of the speakers. The CLAC was created to provide the community with a collection of audio samples from various speakers that could be used to learn a general representation for speech from healthy subjects, as well as complement other health-related speech datasets, which tend to be limited. In this paper, we describe the data collection protocol and summarize the contents of the dataset. We also extract timing metrics from the recordings of each task to explore what those metrics look like for a large, English-speaking population. Lastly, we provide an example of how the dataset can be used by comparing the metrics to those extracted from a small sample of Frontotemporal Dementia subjects. We hope that this dataset will help advance the state of the art in the health and speech domain.

Index Terms: speech recognition, health, amazon mechanical turk, language comprehension, speech corpus

1. Introduction

Speech has been shown to be a useful modality for diagnosing subjects with various forms of cognitive impairment, including Parkinson's disease [1, 2, 3], Alzheimer's disease [4, 5, 6, 7, 8], Frontotemporal Dementia (FTD) [9, 10, 11], Huntington's disease [12, 13, 14, 15, 16], and more. For this reason, datasets consisting of speech from healthy subjects and subjects diagnosed with various neurocognitive disorders have been collected and used to distinguish healthy subjects from cognitively impaired subjects. However, these datasets tend to be limited in size and often are not publicly available [17, 18]. More speech data is needed for subjects with cognitive impairment for researchers to be able to generalize their findings.

While speech from impaired subjects is needed, speech from healthy subjects is also necessary and can be useful for learning what the speech profile of a healthy population looks like. In a recent review paper, Voleti et al. [19] acknowledged that the characterization of the variability of the speech in healthy populations is a critical research area for advancing the state of the art. We hope to contribute to this research area by providing a dataset of healthy speakers, primarily from the United States, that can be used to gain a more complete understanding of what variability looks like in healthy populations. In this paper, we present a speech dataset consisting of audio recordings and automatically-generated transcripts from speakers that were presumed healthy. They completed several simple language tasks that are present in other health-related speech datasets [9, 20, 21], including common picture description tasks like the cookie theft task [22, 23, 24], which has been used to classify numerous cognitive disorders [24, 25, 26, 27, 28, 29, 30, 31]. In addition to exploring what speech looks like in a healthy population, the dataset presented in this paper can be used to compare the speech of healthy, English-speaking populations in different countries, and/or supplement the data in other health-related datasets.

As far as we know, this is the largest collection of healthy English speakers completing language comprehension tasks and we believe that the scientific community can benefit from the public release of this dataset.

2. Data collection

The audio recordings in the dataset were collected through Amazon Mechanical Turk (AMT), a crowdsourcing website that allows workers to complete tasks created by businesses and researchers (Requesters) for a set cost. The tasks that the workers complete are called Human Intelligence Tasks (HITs). In order to qualify to complete the HIT we created, workers were required to be in the United States (a small subset of the workers were located in different countries) and the percentage of assignments that were submitted by the workers and approved by previous Requesters had to be 90% or higher. Each worker had a unique worker ID. The worker IDs for approved submissions were used to ensure that each worker was only allowed to complete the HIT one time.

2.1. Task selection

The HIT used to collect the data described in this paper consisted of several tasks. Each worker was first asked to select their gender ("Male", "Female", or "Other") and age (a number between 18 and 90, or "Over 90"). Some workers were also asked to select the number of years of education they completed, with 12 years being equivalent to completing high school. There is no education information for 250 workers because the education question was added after data collection began. Each worker was also asked to tell us whether they had a cold, allergy, or other health-related symptoms that might affect their speech the day they completed the HIT ("Yes" or "No"). After that, the workers were asked to complete several simple tasks, all of which can be seen in Table 1, along with the corresponding prompts and the number of audio files in the dataset for each task. These tasks were selected because they have been used to assess and diagnose subjects with impaired speech in previous research [9] and they could be easily implemented and completed by the workers without a proctor present.

The cookie theft and picnic pictures used for the picture description tasks can be seen in Figure 1. For the "repeat 5 times" task, some workers were initially asked to record themselves saying each of the 3 words 5 times in one recording. Subsequent workers were asked to submit separate recordings for

http://dx.doi.org/10.21437/Interspeech.2021-1810

Table 1: *The tasks workers were asked to complete, the corresponding prompts, and the number of audio files in the dataset for each task.*

Task	Prompt	Audio Files
Counting From 1 To 20	Record yourself counting from 1 to 20.	1,816
Days Of The Week	Record yourself saying the days of the week, starting with Monday.	1,829
Cookie Theft	Record yourself describing everything that you see in the picture below using complete sentences.	1,832
Picnic	Record yourself describing everything that you see in the picture below using complete sentences.	808
Grandfather	Record yourself reading the following passage: "You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language."	1,832
Rainbow	Record yourself reading the following passage: "The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon."	1,832
Repeat 5 Times	Record yourself repeating the following words 5 times each in the same recording: artillery, catastrophe, impossibility.	250
Repeat 5 Times Artillery	Record yourself repeating the word "artillery" 5 times.	1,582
Repeat 5 Times Catastrophe	Record yourself repeating the word "catastrophe" 5 times.	1,582
Repeat 5 Times Impossibility	Record yourself repeating the word "impossibility" 5 times.	1,582
SMR	Record yourself repeating /pataka/ (pah tah kah) as fast as you can for 10 seconds.	1,832
Max Phonation	Please take a deep breath and then record yourself sustaining voicing of the vowel /a/ (ah) at a comfortable pitch and loudness level for as long as you can.	1,832

each word. As a result, there are 250 workers with one repetition recording and 1,582 workers with 3 separate repetition recordings. The picnic picture description task was also added after data collection began. As a result, 1,024 workers did not complete the picnic picture description task.

2.2. Validation

Transcripts were automatically generated for each of the submitted audio files using the Google Speech Recognition API [32]. The transcripts were then used to validate the submitted audio files by checking the number of words in the transcript and the length of the audio file. If the number of words and length of the audio file were satisfactory (different threshold values were used for different tasks), then the worker was allowed to move on to the next task. Otherwise, they were asked to complete the task again. These validation checks were added to ensure that workers did not submit incomplete assignments.

2.3. Summary statistics

916 speakers selected "Female" for their gender, 903 speakers selected "Male", and 13 speakers selected "Other". The average age of the workers was 35.7 years and the average years of education was 15.4. Histograms showing the age and education distributions can be seen in Figure 2. Workers were located in 962 unique cities, all 50 US states, and 12 unique countries. The majority of the workers (1,815) were located in the United States.

3. Data analysis

An audio activity detection tool called auditok [33] was applied to each AMT recording to determine the start and end times of the speech. The tool used a log energy threshold value to detect the sections of audio that contained speech by ignoring sounds below a certain threshold. A 65dB log energy threshold value was used. The detected start and end times were used to extract several timing metrics from the recordings for each task: the total duration of each audio file (in seconds), the number of speech segments, the speech rate (speech segments per second), the number of pauses, the total duration of pauses, and the proportion of pause time (total duration of pauses divided by the total duration of the audio file). Those metrics were used to explore what timing looks like for a general population that is presumed healthy. The average value for each of the metrics mentioned above can be seen in Table 2 for each of the tasks completed by the workers.

3.1. FTD comparison

One way that we anticipate the AMT data being used is to compare the speech of cognitively impaired individuals with that of healthy speakers. The data can also be used to explore how the speech of healthy speakers differs in different regions/countries. To illustrate this, we also extracted the timing metrics mentioned above from the cookie theft audio files of 58 healthy Australian subjects, and Australian subjects with different types of Frontotemporal Dementia (FTD). The FTD data used is a subset of a larger FTD dataset, part of which has been used in previous

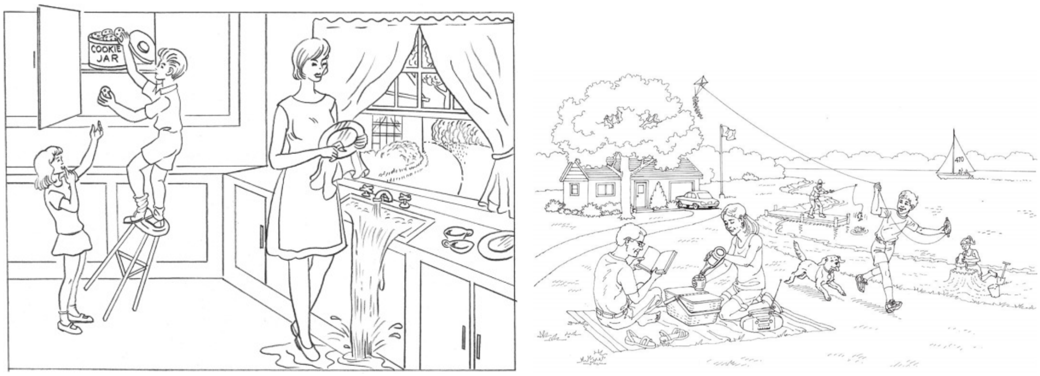

Figure 1: *The images workers were asked to describe for the cookie theft (left) and picnic (right) picture description tasks.*

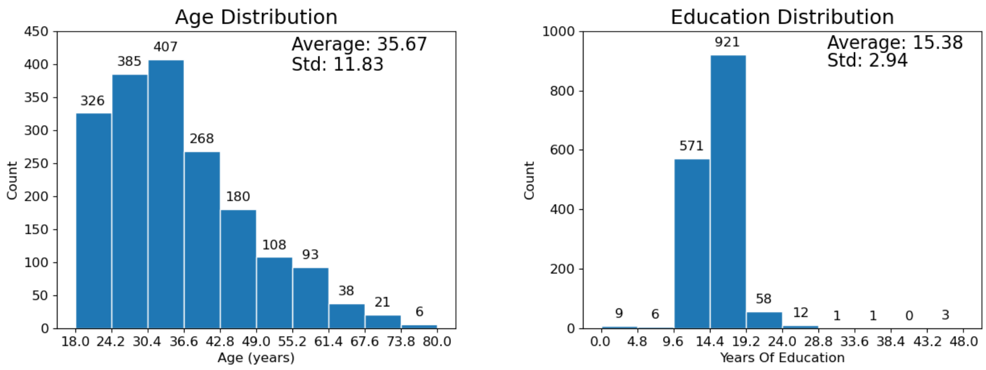

Figure 2: *The age and education distributions of the data.*

research to explore which speech characteristics are most salient for the detection of the behavioral variant of FTD (bvFTD) [9]. More information about the language assessment of the subjects in the dataset can be found in [9]. Timing metrics were extracted from 11 subjects with bvFTD, 6 subjects with the semantic variant of Primary Progressive Aphasia (svPPA), and 7 subjects with the logopenic variant of PPA (lvPPA).

The averaged metrics for each of the FTD variants and the healthy subjects can be seen in Table 3. The results show that each of the timing metrics are lower for the healthy Australian speakers compared to the healthy speakers in the CLAC, which consists primarily of American speakers. However, due to the large difference in sample size for the two groups, we can not draw any strong conclusions from this observation. The results also show that the timing metrics are the same or higher for each of the FTD groups compared to the healthy groups. Previous research has shown that the speech of lvPPA and bvFTD subjects is characterized by a greater proportion of pause time and an increased number of pauses, the speech of lvPPA subjects is also characterized by an increase in pause duration, and the speech of svPPA subjects is characterized by a decreased speech rate [10]. All of the results in Table 3 are consistent with the findings in previous research, except for the increase in speech rate for svPPA subjects compared to the healthy subjects. This discrepancy may be due to the limited sample size of the svPPA subjects.

The comparison of healthy speakers with FTD subjects is just one example of how the CLAC can be used. Similar experiments can easily be conducted with a different kind of dataset consisting of healthy speech, impaired speech, or both.

4. Limitations

While we hope that the dataset can aide researchers in understanding what the speech of the general population looks like, we acknowledge that there are some limitations associated with the dataset:

- Self-reported metadata: Each worker was allowed to report their age, gender, and years of education. The information submitted by the workers could not be verified. Therefore, some of the information may be incorrect.
- Recording environment: Since the workers were allowed to complete the tasks from wherever they were, there was a lot of variety in the type of microphones that were used and the environments that the workers were in. While the difference in recording quality may make analysis more challenging, the variety will also lead to greater generalizability.
- Health assumption: We made the assumption that all workers were healthy and did not ask them about their previous medical history. For this reason, we cannot know for sure that each speaker is healthy and it is possible that some speakers in the dataset may have conditions that can impair their speech.

Table 2: *The average values for the timing metrics extracted for each task.*

Task	Speech Duration	Num. Speech Segments	Speech Rate	Num. Pauses	Pause Duration	Proportion Pause Duration
Cookie Theft	30.50	13.42	0.45	12.43	6.47	0.21
Picnic	43.04	19.21	0.45	18.22	10.18	0.24
Counting 1 To 20	19.37	17.21	0.92	16.23	6.87	0.33
Days Of The Week	7.30	6.01	0.85	5.04	2.04	0.26
Grandfather	48.45	21.00	0.44	20.00	8.56	0.17
Rainbow	12.49	5.41	0.45	4.42	1.61	0.13
Repeat 5 Times	20.35	12.59	0.63	11.61	6.02	0.28
Repeat 5 Times Artillery	6.38	4.57	0.74	3.59	1.86	0.26
Repeat 5 Times Catastrophe	6.65	4.62	0.73	3.65	2.00	0.28
Repeat 5 Times Impossibility	7.45	4.59	0.64	3.61	1.81	0.22
Smr	9.94	6.07	0.64	5.10	1.19	0.12
Max Phonation	10.79	4.22	0.48	3.27	2.95	0.22

Table 3: *The average values for the timing metrics extracted for each FTD variant and healthy category on the cookie theft task.*

Group	Speech Duration	Num. Speech Segments	Speech Rate	Num. Pauses	Pause Duration	Proportion Pause Duration
CLAC (n = 1,832)	30.50	13.42	0.45	12.43	6.47	0.21
Healthy (n = 58)	26.97	11.29	0.41	10.29	4.79	0.16
bvFTD (n = 11)	60.88	25.63	0.45	24.63	26.18	0.41
svPPA (n = 6)	55.82	22.56	0.60	21.56	34.06	0.51
lvPPA (n = 7)	116.35	51.38	0.433	50.38	67.04	0.58

- Different accents: The majority of the workers were located in the United States. However, there is still a variety of different accents and dialects due to differences in the locations and backgrounds of the workers. While this makes the dataset less "clean", it can also be good for generalizability.
- Duplicate worker submissions: Each worker has a unique worker ID and that information was used to ensure that a worker with a particular worker ID was not allowed to complete our HIT more than once. However, there was no way to check whether someone had multiple AMT accounts. Therefore, we can not rule out the possibility that the same speaker completed the HIT multiple times from different AMT accounts.
- Transcript quality: The quality of the automatically-generated transcripts varies significantly depending on accent and recording quality. Therefore, some transcripts have high accuracy while others may have incorrect words or may be missing some words completely. Future releases of the dataset will include a corrected version of the ASR transcripts.
- Age range: The majority of the participants are not within the age range of subjects that are typically diagnosed with cognitive disorders. However, it can still be useful to have speech from younger speakers that can be used to possibly examine how speech differs between speakers of different ages in our general population.

While there are some limitations, there are also some benefits, including the fact that (1) diarization is not needed for the recordings in this dataset because only the voice of the worker is present in the recording and (2) we are not aware of any other datasets of this magnitude with speech from healthy subjects completing cognitive tasks that can complement other health-related speech data, making this dataset a significant contribution to the field.

5. Conclusions

In this paper, we presented CLAC, a speech dataset consisting of audio recordings and automatically-generated transcripts from 1,832 speakers located in the United States, as well as 11 other countries. We demonstrated how the dataset can be used to characterize the speech of a healthy, English-speaking population and distinguish between healthy subjects and subjects with some form of cognitive impairment. We discussed the limitations of the dataset and believe that the dataset is a valuable contribution to the scientific community, despite those limitations. In the future, we plan to expand the data collection to include other English dialects, such as British English. We also plan to use this dataset to evaluate the utility of the data for augmenting experimental data for patients with conditions such as FTD. The dataset can be downloaded from https://groups.csail.mit.edu/sls/downloads/clac/downloads.cgi.

6. Acknowledgements

The authors would like to thank Millennium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals) and Mathworks for funding this research. The authors would also like to thank Adam Vogel for sharing the FTD dataset and providing guidance for how to design the data collection tool.

7. References

[1] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[2] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[3] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[4] R. Haulcy and J. Glass, "Classifying Alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 3833, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2020.624137

[5] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, no. Preprint, pp. 1–27, 2020.

[6] J. V. E. López, L. Tóth, I. Hoffmann, J. Kálmán, M. Pákáski, and G. Gosztolya, "Assessing alzheimer's disease from speech using the i-vector approach," in *International Conference on Speech and Computer*. Springer, 2019, pp. 289–298.

[7] A. Pompili, T. Rolland, and A. Abad, "The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2202–2206. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2833

[8] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection," in *Proc. Interspeech 2020*, 2020, pp. 2167–2171. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2557

[9] A. P. Vogel, M. L. Poole, H. Pemberton, M. W. Caverlé, F. M. Boonstra, E. Low, D. Darby, and A. Brodtmann, "Motor speech signature of behavioral variant frontotemporal dementia: Refining the phenotype," *Neurology*, vol. 89, no. 8, pp. 837–844, 2017.

[10] M. L. Poole, A. Brodtmann, D. Darby, and A. P. Vogel, "Motor speech phenotypes of frontotemporal dementia, primary progressive aphasia, and progressive apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 4, pp. 897–911, 2017.

[11] V. C. Zimmerer, C. J. Hardy, J. Eastman, S. Dutta, L. Varnet, R. L. Bond, L. Russell, J. D. Rohrer, J. D. Warren, and R. A. Varley, "Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: an approach based on usage-frequency," *Cortex*, vol. 133, pp. 103–119, 2020.

[12] T. N. Grimstvedt, J. U. Miller, M. R. van Walsem, and K. J. B. Feragen, "Speech and language difficulties in huntington's disease: A qualitative study of patients' and professional caregivers' experiences," *International Journal of Language & Communication Disorders*, 2021.

[13] A. P. Vogel, C. Shirbin, A. J. Churchyard, and J. C. Stout, "Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset?" *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, 2012.

[14] S. Skodda, U. Schlegel, R. Hoffmann, and C. Saft, "Impaired motor speech performance in huntington's disease," *Journal of Neural Transmission*, vol. 121, no. 4, pp. 399–407, 2014.

[15] W. Hinzen, J. Rosselló, C. Morey, E. Camara, C. Garcia-Gorro, R. Salvador, and R. de Diego-Balaguer, "A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease," *Cortex*, vol. 100, pp. 71–83, 2018.

[16] J. C. Chan, J. C. Stout, and A. P. Vogel, "Speech in prodromal and symptomatic huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases," *Neuroscience & Biobehavioral Reviews*, vol. 107, pp. 450–460, 2019.

[17] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[18] J. Novikova and A. Balagopalan, "On speech datasets in machine learning for healthcare."

[19] R. Voleti, J. M. Liss, and V. Berisha, "A review of automated speech and language features for assessment of cognitive and thought disorders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 282–298, 2019.

[20] M. L. Henry and S. M. Grasso, "Assessment of individuals with primary progressive aphasia," in *Seminars in speech and language*, vol. 39, no. 3. NIH Public Access, 2018, p. 231.

[21] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[22] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[23] D. Kokkinakis, K. L. Fors, K. C. Fraser, and A. Nordlund, "A swedish cookie-theft corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[24] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks," *Journal of clinical and experimental neuropsychology*, vol. 40, no. 9, pp. 917–939, 2018.

[25] E. Giles, K. Patterson, and J. R. Hodges, "Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information," *Aphasiology*, vol. 10, no. 4, pp. 395–408, 1996.

[26] P. V. Cooper, "Discourse production and normal aging: Performance on oral picture description tasks," *Journal of Gerontology*, vol. 45, no. 5, pp. P210–P214, 1990.

[27] H. Choi, "Performances in a picture description task in japanese patients with alzheimer's disease and with mild cognitive impairment," *Communication Sciences & Disorders*, vol. 14, no. 3, pp. 326–337, 2009.

[28] C. Mackenzie, M. Brady, J. Norrie, and N. Poedjianto, "Picture description in neurologically normal adults: Concepts and topic coherence," *Aphasiology*, vol. 21, no. 3-4, pp. 340–354, 2007.

[29] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 260–268, 2018.

[30] M. F. Mendez and M. Ashla-Mendez, "Differences between multi-infarct dementia and alzheimer's disease on unstructured neuropsychological tasks," *Journal of Clinical and Experimental Neuropsychology*, vol. 13, no. 6, pp. 923–932, 1991.

[31] T. Bschor, K.-P. Kühl, and F. M. Reischies, "Spontaneous speech of patients with dementia of the alzheimer type and mild cognitive impairment," *International psychogeriatrics*, vol. 13, no. 3, pp. 289–298, 2001.

[32] A. Zhang, "Speech Recognition (Version 3.8) [Software]," Available from https://github.com/Uberi/speech_recognition#readme, 2017.

[33] "auditok (Version 0.1.8) [Software]," Available from https://pypi.org/project/auditok/, 2020.