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

Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia

Robert L. MacDonald[1], *Pan-Pan Jiang*[1], *Julie Cattiau*[1] , *Rus Heywood*[1], *Richard Cave*[2], *Katie Seaver*[3], *Marilyn Ladewig*[4], *Jimmy Tobin*[1], *Michael P. Brenner*[1,5], *Philip C. Nelson*[1], *Jordan R. Green*[3,5], *Katrin Tomanek*[1]

[1]Google Research, USA
[2]MND Association, UK
[3]MGH Institute of Health Professions, USA
[4]Cerebral Palsy Associations of New York State, USA
[5]Harvard University, USA
bmacdonald@google.com, katrintomanek@google.com

Abstract

Speech samples from over 1000 individuals with impaired speech have been submitted for Project Euphonia, aimed at improving automated speech recognition systems for disordered speech. We provide an overview of the corpus, which recently passed 1 million utterances (>1300 hours), and review key lessons learned from this project. The reasoning behind decisions such as phrase set composition, prompted vs extemporaneous speech, metadata and data quality efforts are explained based on findings from both technical and user-facing research.

Index Terms: automated speech recognition, disordered speech data, dysarthria, speech data collection

1. Introduction

Current automated speech recognition (ASR) systems achieve high accuracy on typical speech but suffer from significant performance degradation on disordered speech. Recent ASR performance improvements, driven by end-to-end deep neural network (DNN) based methods require large training datasets. However, available datasets of disordered speech are smaller than typical speech datasets, and are harder to acquire. In this sparse data regime, close collaboration between machine learning (ML) researchers and speech domain specialists proves essential for multiple elements of data collection. Such collaboration both ensures that the dataset meets the technical requirements for building accurate ASR models, while also ensuring effective recruitment of and fair compensation for participants. This paper presents details related to data collection from Project Euphonia (g.co/euphonia), where an interdisciplinary team of ML researchers, speech professionals, data collection and infrastructure specialists collaborate in the construction of a large data set (>1M utterances, >1000 speakers) of disordered speech.

Many prior datasets of disordered speech, while useful for applications like speech diagnostics and training traditional ASR systems consisting of separate acoustic and language models [1], are insufficient in size for training modern end-to-end DNN-based ASR systems. Improving ASR performance on disordered speech using these methods therefore requires creation of significantly larger, high quality speech data.

1.1. Considerations for building ASR models

ASR systems process input audio signals and produce text transcripts, so the base requirement for building and testing models for this task are paired audio and text samples. Additional metadata such as a (de-identified) user tag, recording device type, and timestamp can be useful for activities such as error analysis and quality control.

The quantity and quality of data required to achieve a given level of performance, generally unknown *a priori*, will depend strongly on factors such as use case (e.g., vocabulary). For example, personalizing an ASR model to work for a specific individual may require more data from that speaker than a speaker-independent model, which requires data from a large number of speakers. We also might seek a model that works in a limited domain, for example queries related to home automation.

Acoustic factors can also impact ASR performance. High quality speech audio data for research purposes is commonly collected in controlled lab environments with professional recording equipment. This reduces noise and variability stemming from devices (e.g., microphone type and position, audio pre-processing) and environments (e.g., echoes and ambient noise). Project Euphonia's research objectives target ASR systems integrated into consumer devices under conditions of practical use. Performance tests of ASR models require datasets that capture typical noise and other variations. Our dataset is designed with this in mind.

1.2. Objectives of Project Euphonia

Project Euphonia aims to enhance independence and connectedness of people with speech impairments through improved communication tools. Primary research efforts focus on ASR, leveraging rapid and ongoing improvements of end-to-end models [2]. State of the art models achieve low WER on typical speech when trained on large data sets (e.g., 6-8% WER from ~15 million utterances [3]). Such data sets are orders of magnitude larger than any reported corpus of disordered speech.

How much data from people with impaired speech would be required to achieve useful WER? Given the diverse range of speech impairments, can a speaker-independent model for disordered speech be developed, or are personalized models more practical? To answer foundational questions like these,

http://dx.doi.org/10.21437/Interspeech.2021-697

the team constrained the research scope to one language (English) and adults with speech disorders such as dysarthria, deaf speech, severe hypernasality due to cleft lip and palate, and stuttering. We hope that lessons learned will allow more efficient expansion to other languages and disorders.

In what follows we outline the major elements and findings from Euphonia's data collection program. The present paper focuses on the design and construction of the dataset, whereas the corresponding research on training ASR models with this corpus is described in more detail in a companion paper [4]. The dataset described below includes over 1 million utterances and has enabled promising results such as median WER of 5% for personalized models across 432 speakers on the home automation task.

2. Related work

2.1. Existing data sets of disordered speech

Canonical datasets of disordered speech such as the Whitaker database (6 speakers with cerebral palsy, 46 phrases/speaker) [5], the Nemours database (11 speakers with cerebral palsy or head trauma, 740 phrases/speaker) [6], the Universal Access Speech database (19 speakers with cerebral palsy, 765 words/speaker) [7] and the Torgo database (8 speakers with cerebral palsy or ALS, 700 phrases/speaker) [8], are limited either by the number of speakers or number of utterances per speaker or diversity of speech disorder etiologies. Furthermore, these speech data collections have been conducted in centralized, controlled environments, lacking diversity of speaking contexts, recording device types, and ambient environment, such as, variation related to microphone location, room reverberation and background noise.

2.2. Performance gap for ASR on atypical speech

While benchmarks for ASR accuracy on typical speech continue to advance, with reported WER around 5% for selected domains, performance on disordered speech lags significantly behind, with studies showing a nearly 2x difference in WER [9].

Given the large datasets required to train modern ASR models, it was an open question about whether enough data could be collected from atypical speakers to close this gap. Early research from Project Euphonia demonstrated a substantial improvement in WER for ASR models fine tuned on individual speakers with disordered speech [10], where an initial dataset containing 6.7 hours of audio from 67 people with ALS enabled a 62% improvement in WER of a base model adapted using approximately 30 minutes of data from a single speaker. This study validated the idea that fine tuning a state of the art ASR model with even a relatively modest training set could lead to results that are potentially useful. This led us to collect a larger dataset to ascertain the generalizability of the approach to other etiologies, different severities of speech impairment, and to study how performance varies across different speech domains.

3. Key collection elements

3.1. Privacy and consent

User consent and privacy have been factored in from the outset. All of our participants were 18 years or older. Data are collected only after explicit consent has been granted by individuals to provide speech samples for the purpose of research and improving speech recognition products and services. We provide users with clear information on the purpose of the data collection and scope of research. Participants submitting voice recordings are assigned non-identifying user ID's to protect privacy. They can request a copy of their recordings at any time.

3.2. Outreach to diversify the dataset

Partnerships with ALS-TDI, Team Gleason, MND Association, ALS Association, Canadian Down Syndrome Society, and LSVT Global, among many others, have significantly helped raise awareness of Euphonia in speech-affected communities. These advocacy groups also provided thoughtful input on unique needs of their communities, which may include mobility or cognitive impairments.

Technical support has also proved vital to maintaining high throughput of data collections, as participants come with a variety of device types and technology backgrounds. Our customer support team fields email requests, providing known solutions or escalating to the engineering team as necessary.

3.3. Prompted vs spontaneous speech

Once a participant grants permission, they are invited to log into Google's browser-based data collection tool, ChitChat. Participants record individual utterances and have a chance to review and re-record before saving. ChitChat is designed to record audio data where a participant reads a sequence of phrases verbatim. We prioritized scripted speech to yield high transcript conformity and reduce transcription cost. Note that post-facto transcriptions of disordered speech are particularly challenging and error-prone. Scripted speech also allows for studies across multiple speakers, where phrases are identical. A downside is that it reduces the overall linguistic diversity of the aggregated phrase set.

3.4. Lab vs home recordings

Participants record remotely, usually without special equipment or intervention from the research team. This allows us to rapidly grow the speaker population, and also provides recordings in the setting where ASR models would be most useful to the participants. This is a significant difference compared to corpora where people record in a lab setting. Because of the home setup, the Euphonia dataset exhibits a wider, but also more realistic, spectrum of acoustic diversity. For example, approximately a third of utterances to date were recorded on a phone or tablet, with the rest recorded on laptop or desktop computers. Audio data streams from these devices are typically mono, float 32, sampled at 16kHz or above. None of the recordings were captured on smart speakers or related voice-only devices.

3.5. Noise and artifact types

Deciding what types of noise to accept or reject is an important design decision. We label dozens of utterances for every speaker with different noise types, including:

- Technical: segmentation problems (cut-offs at beginning or ending), signal dropouts, undersampling
- Acoustic: environmental noise, secondary speaker cross-talk, low/high recording amplitude, mouse clicks, low signal to noise ratio, reverberation
- Transcripts: misread utterances, non-normalized transcripts

- Speaker: coughs, involuntary vocalization, slow or irregular reading rate.

The sensitivity of ASR model accuracy to noise varies significantly across these different elements and can be difficult to predict. For example, we found signal to noise ratio to correlate with WER particularly strongly for speakers with mild impairments [4].

3.6. Phrase quantity and repetitions

We had to weigh important trade-offs based on feedback from ML researchers and programmatic resource constraints. For example, should a longer list of phrases be recorded from a smaller number of people or vice versa? Should outreach focus on a large number of individuals from a few etiologies or fewer people spread across more etiologies?

An important decision for data collection is the targeted quantity of utterances from a given individual. This is a critical decision, as the conditions that lead to disordered speech might limit speaking endurance, recording rate or the ability to follow simple directions and read prompts. It is important to set requirements that include enough phrases so that the data can be used to train or test a useful recognition system, but not so many phrases to overburden the participant.

Project Euphonia set a target of 1500 utterances per individual, attempting to strike a balance between data richness and participant burden. Recording guidelines anticipate that recordings are done across multiple sessions and encourage participants to limit session length to avoid discomfort and risk of injury. Feedback from participants indicated that the total time required to complete 1500 phrases was typically 4-7 hours, and we provided compensation for individuals based on the number of phrases recorded.

We generated several phrase lists of lengths ranging from a few dozen to 3000 for specific aspects of research, but primarily assigned a standard list of 1500 utterances from home automation and caregiver domains. This allowed us to study the impact of ASR accuracy as a function of number of utterances for both personalized and speaker-independent models. Our experiments showed that personalized models always improved with more data, but the degree of improvement varies across individuals and language/application domains. We also found that speaker-independent model performance improved more when doubling the number of speakers vs doubling the number of phrases per speaker.

As our research objectives evolved, so too did the standard phrase lists. These changes, combined with occasionally skipped phrases by participants, eroded the cross-speaker consistency of the dataset. To overcome this, we built "dense sets" by performing a search of our corpus to select groups of speaker/phrase combinations that were as large as possible while preserving cross-speaker uniformity. For example, we compiled a set of 300 phrases common to 143 speakers.

Another important issue is intra-speaker variability: how much does a phrase change when a participant repeats it? To probe this, we repeated roughly 10% of phrases in several lists. The identical phrases were spaced apart in the prompt list to capture effects such as diurnal variability.

Finally, we asked a small number of participants with degenerative diseases like ALS to re-record phrase lists after several months. These utterances can be used for experiments examining robustness of personalized ASR models against progression of speech impairments.

3.7. Use-cases and domains

Our initial data collection tasks started with phrases selected from conventional speech/voice assessment frameworks, such as the Harvard Sentences [11], the Bamboo Passage [12] and the Cornell Movie-Dialogs Corpus [13]. While these phrases proved useful for assessment of speech disorder severities and classifications, they did not provide the ML research team useful test sets for targeted applications such as conversations, home automation or voice assistant queries.

To probe ASR performance across these use cases, we constructed phrase lists spanning multiple, targeted speech domains. These lists included multiple non-orthogonal dimensions such as use case, sentence length/complexity, and scripted vs extemporaneous speech. Domains include home automation ("Turn on the lights"), caregiver ("I am hungry"), and conversational ("Hey, why don't you come over for dinner tomorrow"). We created specialized phrase lists tailored for a specific research goal, such as evaluating performance on long sentences or enriching the corpus for short phrases with high-frequency words. We also collected and transcribed 8,000 unscripted utterances from 30 speakers for testing.

3.8. Quality control

Building ASR systems with low WER requires high quality data, which can be costly to create. Investments in data quality improvement often need to be prioritized. For example, inaccurate transcripts in test datasets will lead to inaccurate measures of WER, making test set quality a high priority.

Assessments of collected audio data revealed that quality problems generally appeared randomly across speakers but systematically within a speaker's utterances. So a quality control process that is exhaustive across speakers but sampled across utterances is a good balance between economy and effectiveness. We had speech professionals listen to dozens of initial utterances from each speaker to identify systemic recording issues like a bad microphone or background noise. We notified speakers of these issues before assigning longer phrase lists. Reviewers also added audio quality tags such as excessive background noise or reverberation, amplitude clipping or cut-offs. We also manually reviewed utterances used in the test set to correct transcript mismatches.

3.9. Metadata

We assigned each utterance a set of metadata that includes (de-identified) speaker ID, language, transcript(s) and associated domain, timestamp and other information from manual quality evaluation. Multiple transcripts may be associated with an utterance. For example, the exact prompt is retained, even when the speaker said something different and a corrected transcript is created.

ASR research also benefits from speaker-level metadata. For each participant, speech professionals conduct a comprehensive auditory-perceptual speech assessment while listening to a subset of utterances, which were selected for their suprasegmental and segmental diversity. Assessors classify speech disorder type (e.g., dysarthria, apraxia, phonological) and presence/severity of 24 abnormal speech features (e.g., hypernasality, articulatory imprecision, dysprosody). Euphonia's research team uses these labels to identify characteristics of atypical speech that have the greatest impact on recognition accuracy [4].

4. Data Collection Status Quo

As of February 2021 we have recorded over 1 million utterances comprising over 1300 hours of disordered speech by more than 1000 speakers. 579 speakers have recorded at least 300 phrases, our chosen threshold for training personalized models based on studies exploring WER vs training data size (Table 1). The median number of utterances per speaker is 1529; 13 speakers have recorded over 5000 utterances (with one speaker having recorded as many as 15116 utterances).

Table 1: *Composition of Euphonia dataset for participants who recorded at least 300 phrases. Other information such as age, race/ethnicity, education level were not collected.*

	Overall (N=579)
Sex	
FEMALE	217 (37.5%)
MALE	349 (60.3%)
Missing	13 (2.2%)
Etiology	
AMYOTROPHIC LATERAL SCLEROSIS	173 (29.9%)
DOWN SYNDROME	105 (18.1%)
PARKINSONS DISEASE	66 (11.4%)
CEREBRAL PALSY	54 (9.3%)
ATAXIA	21 (3.6%)
STUTTERING	20 (3.5%)
OTHER	18 (3.1%)
HEARING IMPAIRMENT	17 (2.9%)
MUSCULAR DYSTROPHY	14 (2.4%)
UNKNOWN	10 (1.7%)
MULTIPLE SCLEROSIS	9 (1.6%)
STROKE	9 (1.6%)
TRAUMATIC BRAIN INJURY	5 (0.9%)
VOCAL FOLD PARALYSIS	4 (0.7%)
CLEFT LIP and PALATE	3 (0.5%)
LARYNGECTOMY	3 (0.5%)
PRIMARLY LATERAL SCLEROSIS	3 (0.5%)
SPINAL MUSCULAR ATROPHY	3 (0.5%)
BRAIN TUMOR	1 (0.2%)
MULTIPLE SYSTEMS ATROPHY	1 (0.2%)
Missing	40 (6.9%)
Speech Disorder	
DYSARTHRIA	305 (52.7%)
SPEECH SOUND DISORDER	104 (18.0%)
WNL	47 (8.1%)
DYSPHONIA	24 (4.1%)
STUTTERING	20 (3.5%)
DEAF	13 (2.2%)
APRAXIA	9 (1.6%)
APRAXIA & DYSARTHRIA	3 (0.5%)
CLEFT LIP and PALATE	3 (0.5%)
UNKOWN	3 (0.5%)
HYPERNASALITY	2 (0.3%)
OTHER	2 (0.3%)
APHASIA	1 (0.2%)
Missing	43 (7.4%)
Speech Severity	
NORMAL	85 (14.7%)
MILD	233 (40.2%)
MODERATE	136 (23.5%)
SEVERE	105 (18.1%)
PROFOUND	8 (1.4%)
Missing	12 (2.1%)

While participant consent terms preclude Google from making the entire dataset public, we are pursuing ways to allow participants to easily share their data to a central repository, opening access to other research teams.

5. Personalized ASR models

The major use of this corpus to date is to adapt ASR models on a per speaker level. A detailed discussion is given in [4]. As an example, Figure 1 shows the WERs of both unadapted and personalized ASR models of speakers who recorded at least 300 utterances. Overall we find that adaptation is extremely effective, with average WERs improving by about 75% and yielding WERs similar to those of typical speakers. Over 80% of these personalized models have WER below 15%.

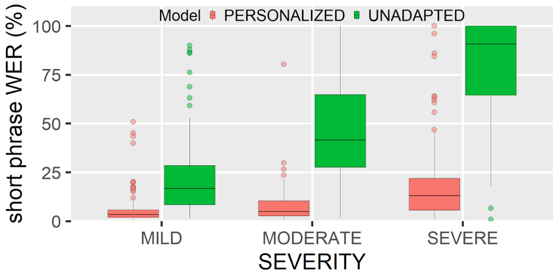

Figure 1: *WER for both unadapted and personalized ASR models using the Euphonia dataset.*

6. Conclusions

We have provided insights into how Project Euphonia compiled a large dataset of disordered speech, covering a wide variety of etiologies and severities. Key decisions on data collection methods, tools and processes are described, many of them informed by essential feedback from researchers using the data to optimize ASR models as well as from speech and language professionals. The corpus to date is the largest known dataset of diverse disordered speech, comprising over 1300 hours of recordings, 1 million utterances and associated transcripts from over 1000 speakers. This corpus has been successfully used to create personalized ASR models with significantly improved accuracy and to research optimization strategies for ASR model personalization.

In the future we aim to develop speaker-independent ASR models capable of accurately recognizing disordered speech of unseen speakers. For that, we are extending our data collection by expanding outreach to enhance our dataset diversity and reducing our phrase list length to drive up overall speaker count.

7. Acknowledgements

We'd like to thank the 1000+ participants who have shared their voice samples. We also thank Dimitri Kanevsky and Aubrie Lee for substantial data contributions, guidance on collection processes. We thank Mungkol Sarin, Knot Pipatsrisawat, Alena Butryna, and Joshua Advincula for critical tooling support. Jordan Green's work was partially supported by NIH-NIDCD under Grant K24DC016312.

We thank groups who connected participants with Project Euphonia, including: ALS-TDI, Team Gleason, CDSS, CureDuchenne, ALSA, LSVT Global, MND Association, and Cerebral Palsy Association of New York State.

8. References

[1] N. M. Joy, S. Umesh, “Improving Acoustic Models in TORGO Dysarthric Speech Database,” *IEEE Trans Neural Syst Rehabil Eng.* 26(3), pp. 637-645, 2018

[2] J. Schalkwyk, “An All-Neural On-Device Speech Recognizer,” Google AI Blog. https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html (accessed March 25, 2021).

[3] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C. Chiu, A. Kannan, “Minimum word error rate training for attention-based sequence-to-sequence models,” 2017, arXiv:1712.01818.

[4] J. R. Green, R. L. MacDonald, P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, K. Tomanek, “Automatic speech recognition of disordered speech: Personalized models now outperforming human listeners on short phrases,” in *Proc. Interspeech*, 2021.

[5] J. R. Deller., M. S. Liu, L. J. Ferrier, and P. Robichaud, “The Whitaker database of dysarthric (cerebral palsy) speech,” *The Journal of the Acoustical Society of America* 93, pp. 3516-3518, 1993

[6] Menndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., and Bunnell, H. T. (1996). “The Nemours database of dysarthric speech,” *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 1962–1966, 1996.

[7] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., and Frame, S., “Dysarthric speech database for universal access research,” in Proc. Interspeech, *2008, pp* 1741–1744.

[8] F. Rudzicz, A. K. Namasivayam, T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” Lang Resources & Evaluation, 46, pp. 523–541, 2012

[9] M. Moore, H. Venkateswara, and S. Panchanath, “Whistleblowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems,” in *Proc. Interspeech, 2018*, pp 466–470.

[10] J. Shor, D. Emanuel, O. Lang, O.. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, Y. Matias, Yossi, “Personalizing ASR for Dysarthric and Accented Speech with Limited Data,” in Proc. Interspeech, 2019.

[11] https://en.wikipedia.org/wiki/Harvard_sentences

[12] Bamboo passage citationYunusova, Y., Graham, N. L., Shellikeri, S., Phuong, K., Kulkarni, M., Rochon, E., ... & Green, J. R. (2016). Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). *PloS one*, 11(1), e0147573.

[13] C. Danescu-Niculescu-Mizil and L. Lee, “Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs,” *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics*, 2011, pp. 76–87