PERSONALIZED AUTOMATIC SPEECH RECOGNITION TRAINED ON SMALL DISORDERED SPEECH DATASETS

Jimmy Tobin and Katrin Tomanek

Google Research, USA
{jtobin,katrintomanek}@google.com

ABSTRACT

This study investigates the performance of personalized automatic speech recognition (ASR) for recognizing disordered speech using small amounts of per-speaker adaptation data. We trained personalized models for 195 individuals with different types and severities of speech impairment with training sets ranging in size from <1 minute to 18-20 minutes of speech data. Word error rate (WER) thresholds were selected to determine Success Percentage (the percentage of personalized models reaching the target WER) in different application scenarios. For the home automation scenario, 79% of speakers reached the target WER with 18-20 minutes of speech; but even with only 3-4 minutes of speech, 63% of speakers reached the target WER. Further evaluation found similar improvement on test sets with conversational and out-of-domain, unprompted phrases. Our results demonstrate that with only a few minutes of recordings, individuals with disordered speech could benefit from personalized ASR.

Index Terms— automatic speech recognition, speech disorders, personalized models

1. INTRODUCTION

Voice controlled home automation technology offers convenience for many; however, often the automatic speech recognition (ASR) systems that power these technologies do not work for the millions of individuals with speech impairments [1]. This population could arguably benefit the most from home automation and other voice controlled assistive technology. People who have speech impairments often also have mobility impairments, due to their conditions. While ASR accuracy for typical speech may be as high as 95% [2, 3], accuracy drops off significantly for disordered speech with varying levels of impairment severity and intelligibility [4].

Adaptation of speaker independent ASR systems with dysarthric speech has shown promising results, but they cite the dearth of speech data as a major hurdle [5]. Recent work has shown the potential of personalizing ASR models for recognizing disordered speech [6, 7, 8, 9, 10]. However, often-times these promising results are based on relatively large amounts of speech recordings per speaker (often in the range of hours). For people with speech impairments, recording so many speech samples can be impractically difficult.

For example, [6] reports average WER improvements of 75% through model personalization on the Euphonia corpus [11], a large corpus with speech recordings from people with speech impairments. While these WER improvements are very promising, it should be noted that this comes with significant recording times per speaker: The median number of utterances per speaker in that study is 1529 (an average of about 2 hours of speech recordings). Recording speech data is a significant investment of time and effort, especially for individuals with motor or cognitive impairments, due to the need to trigger recording start/stops and flipping to the next prompt). It has been reported that 4-7 hours of recording time were required to contribute 1500 phrases [11]. Two recent studies [7, 12] suggest that large WER improvements may be achieved with less data. However, both studies don't analyze in detail how much data is actually needed, how different levels of severity of speech impairment may impact the amount of data required or implications for practical applications.

This paper aims to close this gap. We analyze how much speech data from people with disordered speech is needed for successful model personalization, i.e. models that are usable for voice technology applications and human-in-the-loop conversations. We aim to make technology usable with minimal effort demanded of the speaker. Moreover, understanding how much speech data is required helps accurately set expectations for ASR performance and better utilize the limited time and energy of people with speech impairments.

2. METHODS

2.1. Dataset

We use a subset of the Euphonia corpus [11]. This corpus consists of over 1 million samples (over 1300 hours) of more than 1000 anonymized speakers with different types and severity levels of speech impairments. All our experiments are performed on a subset of 195 speakers who have each recorded more than 1000 utterances.[1] The resulting subset

[1] Speakers with more recordings worked with Euphonia for longer. ALS-TDI was an early partner, explaining the bias towards speakers with ALS.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

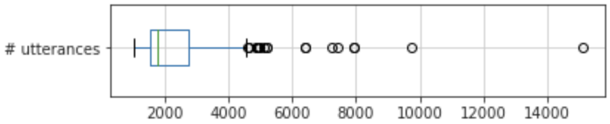

Fig. 1. Distribution of training utterances across all 195 speakers. Median: 1788 utterances (2.1 hours).

Number of utterances	Average duration in minutes (std dev)
250	18.1 (6.6)
100	7.3 (2.7)
50	3.6 (1.3)
30	2.2 (0.8)
10	0.7 (0.3)

Table 1. Number of utterances and average duration of recordings averaged across speakers.

is diverse in terms of severity of speech impairment (7.7% typical, 35.4% mild, 31.8% moderate, 25.1% severe) and covered etiologies (39.2% with amyotrophic lateral sclerosis, 13.4% cerebral palsy, 10.8% Parkinson's Disease, 5.2% hearing impairment, 3.1% Ataxia etc). We use the predefined per-speaker train, dev, and test splits (80%/10%/10%) of the Euphonia corpus, which ensures that there is no phrase overlap between these splits. The median number of utterances per speaker amounts to 1788 for these 195 speakers, but variance is high (see Figure 1). For each speaker, subsets of the training data of various sizes (10, 30, 50, 100, 250 utterances) were created by randomly sampling from all of the speaker's training utterances. For each size, 10 subsets were created by repeated random sampling. Subsets can overlap due to sampling with replacement between sets. Table 1 shows the resulting subsample sizes along with the average duration.

2.2. ASR model personalization

For all our experiments, we employ the Recurrent Neural Network Transducers (RNN-T) architecture [13, 14, 15] which allows deployment on mobile devices, supports streaming, and has demonstrated high performance. We use 8 LSTM encoder layers and 2 LSTM layers for the language model component. Input features are 128-dimensional log Mel features computed every 10 milliseconds. 4 consecutive features are stacked with a stride of 3 frames to yield a 512-dimensional input to the encoder every 30 milliseconds. There are 4096 output units that correspond to the word piece tokens.

We follow the same fine-tuning recipe as described in [6]: We start from a speaker-independent base model pre-trained on 162k hours of typical speech. This base model has been optimized to (a) be robust across various application domains and acoustic conditions, and (b) generalize well to unseen conditions [16]. We only update the first 5 encoder layers. SpecAugment [17] is used for data augmentation. Models were trained to a maximum number of steps based on training set size: 5000 steps for sets with <100 utterances, 10000 steps for 100 utterances, and 15000 steps for 250 utterances. The best checkpoint was picked using each speaker's dev set.

		target WER			
domain	speakers	5	10	15	20
home automation	195	59%	81%	90%	94%
conversational	93	18%	59%	74%	82%

Table 2. Percentage of speakers reaching target WERs of 5, 10, 15 and 20 when personalizing with all data per speaker.

2.3. Evaluation

We use Euphonia's test set splits, consisting of >= 100 utterances for the selected speakers. We report WER on two domains of the Euphonia corpus: phrases associated with a) home automation[2] and b) human conversation[3]. 93 of the 195 speakers recorded a conversational test set, while all speakers recorded a home automation test set. Speaker WERs reported at each subsample size are averages across the 10 models personalized with their associated random subsample.

We chose different WER thresholds as success criteria for our two application scenarios: Home automation and conversational domains. Voice assistants – Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's Assistant being the most popular ones – typically utilize Natural Language Understanding capabilities to infer user intent allowing them to function even without perfect transcriptions from the ASR system. We found that on Google Assistant, a WER of 15 corresponds to a success rate of around 80% for home automation commands. The conversational domain consists of less structured, longer and more complex phrases. Accordingly, we selected a higher WER target of 20, assuming that this renders transcriptions useful in a human conversation.

3. RESULTS

We report **Success Percentage**, the percentage of speakers to reach a target WER, as we aim to understand the expected portion of speakers for which a personalized model trained on a specific amount of data will work in an adequate way.

Table 2 shows the Success Percentage for four WER thresholds (5, 10, 15 and 20) when using all the data available per speaker (as per Figure 1). The conversational domain is more challenging, percentages are accordingly lower than for home automation. When using all data per speaker we reach a target WER of 15 for 90% of the speakers on home automation, and a target WER of 20 for 82% of the speakers on the conversational domain. These numbers serve as baselines.

[2] These are short phrases of 3.2 words on average, eg, "Turn on the radio"

[3] Longer phrases of 7.4 words on average with open vocabulary.

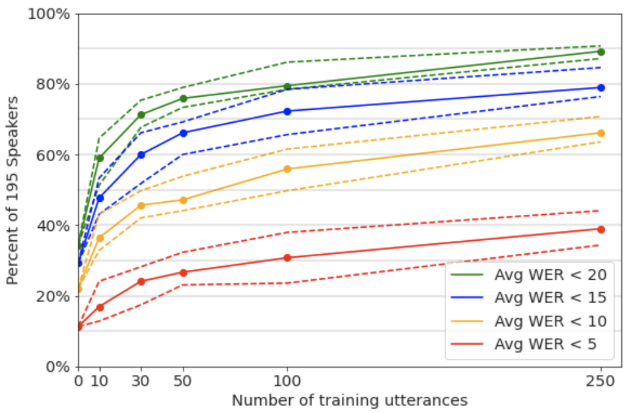

Fig. 2. Success percentage on home automation domain, different target WERs (dashed lines: 95% confidence interval).

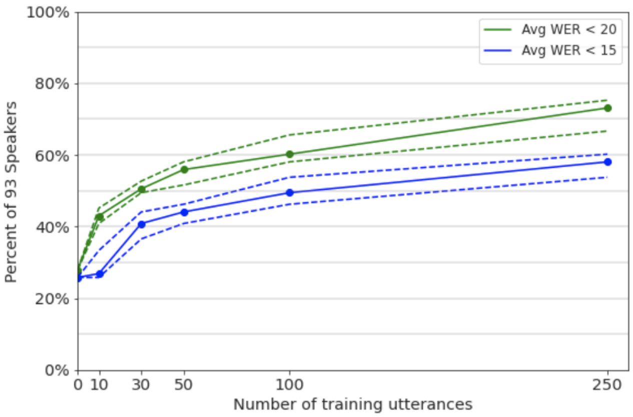

Fig. 3. Success Percentage on conversational domain, different target WERs (dashed lines: 95% confidence interval).

3.1. Percentage of speakers to reach a target WER

Figures 2 and 3 show Success Percentages for both domains for several target WERs and training set sizes. On the home automation domain with a target WER of 15, without personalization the out-of-the-box model reaches this WER for only 35% of the speakers with speech impairment. However, when models are personalized with 250 training utterances per speaker, Success Percentage increases to 79%, and with as little as 50 utterances, we still see a Success Percentage of 63%. On the conversational domain, Success Percentages are generally lower and we restrict our analysis to higher target WERs. For a target WER of 20, the out-of-the-box model yields a Success Percentage of only 29%. Models personalized with 250 utterances see a Success Percentage of 73%.

Overall, this analysis shows that for many speakers only a small amount of data is needed to achieve satisfactory model performance. Many speakers will not need to record more than 250 utterances, with the biggest improvements coming from the first 50. Some users may still want to record more than 250 utterances in order to get sufficient improvement, albeit with marginal gains at a relatively high recording price.[4]

3.2. Break-down by severity of speech impairment

	num	number of utterances				
severity	spkrs	0	10	50	250	All
Typical	15	100%	100%	100%	100%	100%
Mild	69	45%	78%	93%	96%	99%
Mod	62	10%	32%	60%	81%	94%
Severe	49	10%	8%	27%	47%	69%

Table 3. Percentage of speakers reaching target WER of 15 on home automation domain, split by severity.

	num	number of utterances				
severity	spkrs	0	10	50	250	All
Typical	4	100%	100%	100%	100%	100%
Mild	29	41%	62%	86%	93%	97%
Mod	37	16%	35%	46%	70%	86%
Severe	23	17%	22%	26%	48%	52%

Table 4. Percentage of speakers reaching target WER of 20 on conversational domain, split by severity.

To further understand how much personalization data is really necessary for a given speaker, we here report Success Percentages for the different levels of severity of speech impairment. For home automation, we chose a target WER of 15 (Table 3), for conversational we chose a target WER of 20 (Table 4). For all speakers rated as non-typical, we clearly see Success Percentages increase uniformly as the number of training utterances increases.[5] As severity increases, we observe that the maximum possible Success Percentage using all data decreases: only 52% of speakers with a severe speech impairment on conversational and 69% of those speakers on home automation. With even fewer training data (especially <100 utterances), Success Percentages here sharply decline.

On the other hand, for speakers with a mild speech impairment, small amounts of training data yield big Success Percentages, getting close to the rates achievable with all data: With only 50 utterances (i.e. about 3-4 minutes of recorded speech, Table 1) we can expect 90% of these speakers to yield the target WER of 15. This is not far off of the 99% Success Percentage of the ”all data” scenario which clocks in with a

[4] By recording an order of magnitude more data than 250 utterances, the Success Percentage increases by 11% more speakers (home automation domain) and 9% more speakers (conversational domain).

[5] There is one exception: on home automation and severely impaired speakers, Success Percentage drops from 10% to 8% (going from 0 to 10 utterances). This is due to variance coming from repeated sampling of small training sets. The 95% confidence interval for severe speakers at 10 utterances is around 8-15% and the average is at the bottom of that interval.

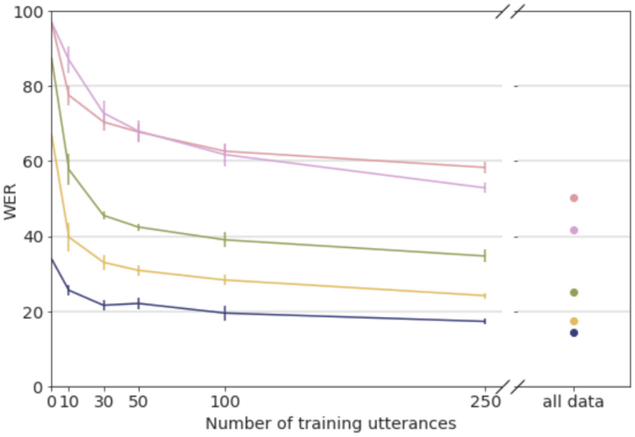

Fig. 4. Average WER and 95% confidence intervals on organic test set of 5 selected speakers. Last point is WER when training with all data of given speaker.

median duration of 2.1 hours of speech recordings (1788 utterances) per speaker. These are large gains when contrasted with the 45% Success Percentage of the out-of-the-box model (results on conversational domain analoguous).

This analysis shows that the answer to the question of how much data is needed to achieve well working, personalized models for speakers with speech impairment clearly depends on the severity of impairment. Personalized models for speakers with moderate and severe impairment benefit significantly from more training data beyond the maximum subsampling size of 250 utterance of this study. For mildly impaired speakers, however, the effort of recording more samples for personalization has only marginal benefit.

3.3. Implications for spontaneous free-form speech

In the previous sections, both the training and the test data were taken from the Euphonia corpus where all recordings are based on prompted speech [11]. Our experiments clearly show that useful personalized models can be obtained with relatively small amounts of training data per speaker. An important, yet to answer question is whether such models also work in a scenario of spontaneous speech of an unknown domain. In other words, how well do models personalized on small amounts of data generalize to other domains, acoustic conditions, and scripted vs unscripted speech.

For 15 of the speakers included in our study, we recorded a so-called ”organic test set”. In direct interactions with these individuals, their spontaneous speech on a topic of their choice was recorded, broken into utterances and then transcribed by speech professionals. Speakers recorded an average of 140 test utterances. Figure 4 shows the WER curves for five selected speakers.[6] Unsurprisingly, WERs on the organic test set are higher because spontaneous speech is more unlike the training set. However, just as with the prompted home automation and conversational domains, we can see a knee in the WER curves within the first 100 utterances – the biggest improvements are even gained within the first 50 utterances. Table 5 shows the relative WER improvement for different training set sizes averaged over all speakers. Note, that there is significant variance between the 15 speakers and we can see ongoing relevant gains beyond 250 utterances. Overall we observe the same trends as on prompted data, showing that training with small amounts of out-of-domain data does indeed generalize and leads to large model quality improvements on spontaneous speech.

utterances	avg WER impr (std dev)
avg = 3239 (std dev = 1717)	56% (15%)
250	45% (16%)
50	34% (12%)
10	17% (12%)

Table 5. Average WER improvement over out-of-the-box model on organic test set across 15 speakers.

4. CONCLUSIONS

Our study shows that small amounts of training data can indeed be sufficient to provide useful personalized ASR models for a large percentage of speakers with speech impairments, especially for use cases like home automation with limited vocabulary and less complex phrases. With as little as 18.1 minutes of recorded speech per speaker on average, we can personalize models with a pre-defined target WER for 96% of speakers with mild, and 81% with moderate speech impairment across a variety of etiologies and types of disorders.

This study also emphasizes how severity of speech impairment and amount of training data needed for useful models are correlated. For speakers with more severe speech impairment, we see larger gradual model improvements as the amount of training data is increased. We also show that even with small amounts of training data from narrow domains and prompted speech, personalized speech models show big improvements in recognition quality on out-of-domain phrases and spontaneous speech, suggesting good generalization properties of personalized models.

In future work, we aim to study how we can further reduce the amount of data needed for speech model personalization based on task-specific utterance selection.

5. ACKNOWLEDGEMENTS

We’d like to thank all participants of Project Euphonia, as well as Rus Heywood and Julie Cattiau for proof-reading.

[6] Due to the small amount of speakers and the fact that the organic test sets of all speakers have different characteristics in domain and phrase complexity we refrain from showing Success Percentage for a target WER.

6. REFERENCES

[1] Neil Bhattacharyya, "The prevalence of voice problems among adults in the united states," *The Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, 2014.

[2] Robert Stonjic, Ross Taylor, Marcin Kardas, Viktor Kerkez, Ludovic Viaud, Elvis Saravia, and Guillem Cucurull, "Papers with code - speech recognition," paperswithcode.com/task/speech-recognition, 2021, Accessed: 2021-09-25.

[3] Gabriel Synnaeve, "Wer are we?," github.com/syhw/wer_are_we, 2021, Accessed: 2021-09-25.

[4] Meredith Moore, Hemanth Venkateswara, and Sethuraman Panchanathan, "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems," in *Proc. Interspeech 2018*, 2018, pp. 466–470.

[5] Mumtaz Begum Mustafa, Siti Salwah Salim, Noraini Mohamed, Bassam Al-Qatab, and Chng Eng Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PLOS ONE*, vol. 9, no. 1, pp. 1–11, 01 2014.

[6] Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek, "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases," in *Proc. Interspeech 2021*, 2021, pp. 4778–4782.

[7] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Proc. Interspeech 2019*, 2019, pp. 784–788.

[8] Han Zhu, Li Wang, Pengyuan Zhang, and Yonghong Yan, "Multi-Accent Adaptation Based on Gate Mechanism," in *Proc. Interspeech 2019*, 2019, pp. 744–748.

[9] Robert Gale, Liu Chen, Jill Dolata, Jan van Santen, and Meysam Asgari, "Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques," in *Proc. Interspeech 2019*, 2019, pp. 11–15.

[10] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech 2019*, 2019, pp. 4115–4119.

[11] Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia," in *Proc. Interspeech 2021*, 2021, pp. 4833–4837.

[12] Katrin Tomanek, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim, "On-device personalization of automatic speech recognition models for disordered speech," *arXiv e-prints, abs/2106.10259*, 2021.

[13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP 2013*, 2013, pp. 6645–6649.

[14] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al., "Streaming end-to-end speech recognition for mobile devices," in *Proc. ICASSP 2019*, 2019, pp. 6381–6385.

[15] Tara N Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, Shuo-yiin Chang, Wei Li, Raziel Alvarez, Zhifeng Chen, et al., "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *Proc. ICASSP 2020*, 2020, pp. 6059–6063.

[16] Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N. Sainath, and Trevor Strohman, "Recognizing long-form speech using streaming end-to-end models," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 920–927.

[17] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.