

WER we are and WER we think we are

Piotr Szymański^{1,2,*}, Piotr Żelasko^{3,*}, Mikołaj Morzy^{1,4}, Adrian Szymczak¹,
Marzena Żyła-Hoppe¹, Joanna Banaszczak¹, Łukasz Augustyniak^{1,3}
Jan Mizgajski^{1,4}, Yishay Carmiel¹

¹ Avaya Inc., USA

² Wrocław University of Science and Technology, Wrocław, Poland

³ Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

⁴ Poznań University of Technology, Poznań, Poland

* Equal contribution

piotr.szymanski@pwr.edu.pl, pzelasko@jhu.edu

Abstract

Natural language processing of conversational speech requires the availability of high-quality transcripts. In this paper, we express our skepticism towards the recent reports of very low Word Error Rates (WERs) achieved by modern Automatic Speech Recognition (ASR) systems on benchmark datasets. We outline several problems with popular benchmarks and compare three state-of-the-art commercial ASR systems on an internal dataset of real-life spontaneous human conversations and HUB'05 public benchmark. We show that WERs are significantly higher than the best reported results. We formulate a set of guidelines which may aid in the creation of real-life, multi-domain datasets with high quality annotations for training and testing of robust ASR systems.

1 Introduction

The last few years have witnessed unprecedented progress in Automatic Speech Recognition (ASR) systems. They have become ubiquitous in everyday lives, from phones voice assistants through dictation of text messages and e-mails to managing household appliances with home assistants. It is not surprising that major vendors are trying to showcase the quality and accuracy of their products. A comprehensive benchmark of available ASRs (Synnaeve, 2020) cites word error rates (WERs) as low as 2%–3% on standard datasets. These reports may incur a false conviction that automatic speech recognition is mostly a solved problem. Nothing could be further from the truth.

One possible cause for this misconception and gross over-estimation of the accuracy of contemporary ASRs is the confounding of two regimes

of speech recognition. The majority of interactions with ASRs happen in the context of chatbot-like interactions, when a human is fully aware of speaking to a machine. In these circumstances most people simplify their utterances, speaking in short, well-structured phrases which obey the correct grammatical structure of either an interrogative or an imperative sentence. Siegert and Krüger (2018) point out that conversations between people are more lively and dynamic, while the conversations with Alexa remain mostly simple request/response interactions. Significant acoustic differences allowed them to build a classifier which achieves 81% of accuracy for distinguishing between human-human and human-chatbot interactions. Human-chatbot interactions are in stark contrast with typical spontaneous human-human conversations. Such conversations are riddled with various disfluencies (discourse markers, filled pauses, back-channeling), lack clear phrase borders, and human utterances are often not terminated correctly. Oftentimes information is exchanged using non-verbal means, for example, prosody, paralinguistic, and non-verbal vocalizations. Hill et al. (2015) found that human–chatbot communication lacked much of the richness of vocabulary found in conversations among people. All these factors make the correct recognition of spontaneous human conversations a very challenging task.

We do not share the optimistic view of the overall ASR accuracy. In our opinion, ASR systems still have a long way to go toward robust recognition of spontaneous human conversations. Popular ASR benchmarks, such as Librispeech (Panayotov et al., 2015), WSJ (Paul and Baker, 1992), Callhome, Fisher (Cieri et al., 2004), or Switchboard (Godfrey et al., 1992), are – to a different extent – misaligned

with the contemporary domains of ASR applications. Some of these benchmarks are too simple to truly challenge ASRs, while the more challenging ones do not cover the full spectrum of inputs encountered in everyday operations. As such, these benchmarks do not provide reliable estimations of the actual ASR accuracy. On the other hand, on the truly challenging benchmarks, such as the dinner party conversations CHiME5 (Manohar et al., 2019) benchmark, modern ASRs report WERs in the ranges of 46%–73%.

Another problem is the domain adaptation. Both Fisher and Switchboard corpora, although trying to mimic real, spontaneous conversations, are inherently artificial. The protocols for the creation of both corpora involved pairs of voice actors having a conversation on a random subject drawn from a collection of predefined topics. These conversations are very different from real-life spontaneous conversations, where the topics may vary greatly, but at the same time the domain of application imposes strict constraints on the vocabulary and the form of the conversations. Many benchmark datasets represent scripted or semi-scripted conversations (EU Parliament speech transcripts, TED talk transcripts). There are consequential differences between scripted and spontaneous conversations and they affect the results of the ASR evaluation (Shriberg et al., 2001).

Finally, benchmark datasets tend to be quite homogeneous with respect to the demographic features of voice actors. In real-life conversations, factors such as age, gender, ethnicity, and dialect are important determinants of speech particularities. Non-native language speakers are virtually absent from benchmark datasets, and the novelties introduced by these speakers, both acoustic (pronunciation) and linguistic (vocabulary and syntax), are not accounted for in the results of ASR evaluation (Koencke et al., 2020). Even for gender, a variable ostensibly easy to control, we find a huge imbalance in the distribution of the number of utterances in benchmark datasets. Benchmark datasets do not represent the true diversity of real-world conversations, both at input signal characteristics and conversation semantics levels.

In our opinion, what is commonly assumed to constitute the state-of-the-art in the ASR quality is grossly over-estimated. IRL-WERs (in-real-life WERs) are much higher than reported. In this paper, we compile a benchmark of real-life phone

call spontaneous conversations from five different domains and present the unbiased comparison of three major cloud ASR systems. We show that the WERs observed both in the public HUB'05 benchmarks and in real conversations are much higher than usually reported, and this phenomenon is observed across all of the tested domains. Our benchmark consists of call center conversations containing almost 3000 utterances. The conversations were transcribed by two professional human annotators. The conclusions are clear: we are definitely not where we think we are in terms of WERs.

The contributions of our paper are twofold. Firstly, we experimentally compare WERs of three major commercial ASR systems on publicly available benchmarks and we contrast these results with our internal benchmark of real-life human spontaneous conversations. Secondly, we raise community awareness regarding the problems caused by the optimistic bias towards the ASR accuracy. We issue a call to action with the aim of addressing this bias by researchers and funding institutions.

2 The State of the Error Rate

We evaluate several ASR systems on a multi-domain dataset of 50 call center conversations recorded at 8 kHz, using standard, modern telephony quality. The dataset comprises 8.5 hours of audio, including 2.2 hours of speech. These calls consist of 1595 agent and 1361 customer utterances. An average utterance consists of 10 words, thus resembling a sentence-length phrase. For speech recognition we use three different state-of-the-art commercial ASR solutions which we find representative in terms of what is available on the market. When a given ASR vendor offered such an option, we used a telephonic speech model. We report the WER of evaluated ASRs on our dataset and on the HUB5'00 Switchboard and CallHome evaluation subset to allow comparison to the state of the art on publicly available data.

We collected per channel audio data as RIFF (little-endian) data, WAVE audio, Microsoft PCM, 16 bit, mono 8000 Hz. These audio data files were transcribed by each of the ASR systems and the transcription results were exported as CTM files. We used the NIST Scoring Toolkit (sctk) scripts to obtain WER scores from the reference STM and predicted CTM files, using the established hubscr.pl script. We collected these sctk-based scores into data frame and organized the results into tables

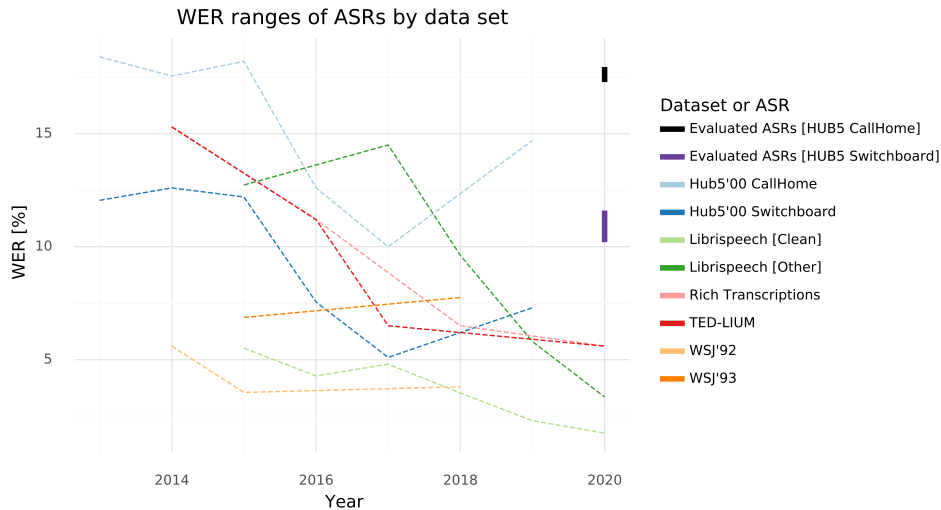


Figure 1: WER ranges in ASR systems. Reference values taken from the *Wer are we* report (Synnaeve, 2020) and the *Papers with Code* website (Stojnic et al., 2020) for ASR solutions published in the last 5 years. Outliers were removed for the sake of figure readability.

per domain and the user of a given channel.

We would wholeheartedly like to provide a more reproducible contribution, however, due to legal constraints, we are required to make the experimental procedure more obscure than we would like. While we understand the scientific expectation of complete experimental transparency and reproducibility, we are not able to provide the community with neither the benchmark data nor the detailed information about evaluated systems.

To set our benchmark in context, we present the median ASR WER per year reported in scientific papers as aggregated by the *wer_we_are* report (Synnaeve, 2020). They provide a basis for an optimistic view on both the state of the art in ASRs published by scientific teams and a visible trend of improvement. Depending on the dataset, the error rates, as of 2019, do not exceed 15%. Depending on the data origin and quality, WER drops as low as 2%–4% for the Librispeech (Panayotov et al., 2015) corpus of approximately 1000 hours of 16kHz recordings of English audiobooks. The WSJ’92 and ’93 corpora (Consortium et al., 1994) contain 73 hours of clean speech dictation and clean conversational speech of journalists; and ASRs have admitted a WER of 3%–7% on this dataset. The TED-LIUM ASR (Rousseau et al., 2012) task consists of 118 hours of speech recorded in high quality from TED Talks and its WER is reported at 5%. The currently best WER for the HUB’05¹ evaluation is at 5% for the Switchboard

¹LDC2002T43

part and 9% for the CallHome part – both of which are phone conversational datasets.

Unfortunately, as we see in Table 1, the commercial ASR systems in our evaluation achieve nearly double the error rates on both HUB’05 evaluation subsets. This result may be explained as follows. Firstly, the results are typically reported using the oracle segmentation of the evaluation data, running the speech recognizer on each segment of speech separately. Instead, we evaluated each ASR by providing the whole 5 minutes chunks of audio, so that each system had to perform speech activity detection (SAD) first. This resembles the real-life application usage. Secondly, for each public dataset, the ASR system usually uses a language model (LM) estimated on the training set, making it representative of the domain data. In contrast, commercial ASR systems have to use more general LMs which are likely to perform worse on any specific benchmark.

ASR	CCC	SWBD	CallHome
ASR 1	17.9	11.62	17.69
ASR 2	19.2	11.45	18.6
ASR 3	16.5	10.2	15.85

Table 1: WER [%] comparison on benchmarks

The gap between the state-of-the-art results reported on public benchmarks and real-life is even more visible in the case of multi-domain datasets. In Table 2 we present WERs obtained on our internal benchmark. The five domains evaluated in

the benchmark include booking touristic reservations, finance, two types of insurance domains, and telecommunication conversations. We can see that insurance #2 was the easiest of domains for the ASRs, followed by finance. The highest error rates are obtained for booking and wireless telecommunication calls, which can be related to the fact that more of their utterances contain entities related to date and time, money, places, and product or company names. We also include the division of WERs per channel. Per channel differences are not large enough to assume that one of the channels is in a significantly better position for performing downstream NLP tasks.

	ASR 1	ASR 2	ASR 3
Booking	21.19	22.16	20.95
Finance	16.82	18.46	15.83
Insurance 1	18.01	20.20	17.84
Insurance 2	15.25	17.11	13.73
Telecomm.	19.75	23.31	17.62
Agent	16.97	17.83	16.49
Customer	17.87	20.99	16.48

Table 2: Internal benchmark WER [%]

3 Call to Action

The main reason for publishing the results presented in Section 2 is to support our call to action. We want to encourage the ASR and NLP communities to pursue research and collaboration to address the shortcomings of current ASR benchmarks. We need to collect and annotate audio datasets that are much better aligned with contemporary application domains of ASR systems, work on extended and more inclusive acoustic models representing a much broader spectrum of dialects, account for technological advances which influence physical properties of processed audio signals, and develop language models for multi-domain conversations.

The situation where most available spontaneous conversation datasets are over 20 years old is both easy to overlook and hard to believe. What is worse, judging by the standards one expects from modern NLP corpora, these datasets lack rich annotations. In many cases, part-of-speech tags and dependency structure are missing, and few datasets have complete dialog act annotations. Hardly any datasets contain named entity annotations beyond very basic NER schemes.

High quality annotations enable building better

language models and can improve ASR quality metrics even in the simplest of scenarios when the ASR was trained with a language model aware of part of speech (POS) annotations. With just such basic annotations Stiefel and Vu (2017) reported a small WER decrease even on data sets without proper human annotation, with POS tag obtained via automatic tagging.

The LDC catalog, the most comprehensive repository of language-related corpora, lists only four phone conversation datasets² collected recently (i.e., datasets which are not based on Fisher, Callhome, or Switchboard). None of these new datasets contain rich annotations, and the datasets are not related to real-world, multi-domain conversations – precisely the type of application the industry is trying to deploy. For more advanced applications, such as machine translation or machine comprehension, such resources are scarce.

Recently published datasets for spoken question answering are synthetically produced by text-to-speech systems (Lee et al., 2018; Li et al., 2018). This artificial generative procedure strips training data of many subtle characteristics of human speech, such as prosodic features and non-verbal acoustic cues. Many aspects of spontaneous human speech are difficult to model in a text-to-speech synthesizer, rendering the resulting recordings less aligned with data on which ASRs will be applied.

These problems are not insurmountable. A thoughtful collaboration between academia and industry partners can lead to the creation of high-quality training and testing datasets. Importantly, these datasets should – preferably – be published under an open-access license, or at least should be made available through openly accessible data platforms like Mozilla’s Common Voice.

It is of important to address the legal issues of open-sourcing the recordings of the human voice, which currently hold back the industry from releasing conversational speech datasets such as ours. This is an opportunity for funding institutions to spark interdisciplinary research and economic growth through R&D breakthroughs in the area of automatic speech recognition.

There are multiple factors which may contribute to the progress in ASR for human conversational speech. These include funding schemes, grant programs, data donations, and student internships, to name a few.

²LDC2015S08,LDC2019S06,LDC2013S05,LDC2010S02

In our opinion, a comprehensive program for advancing the field should encompass the following:

- preparing new audio and transcript datasets with rich annotations including: POS tags, dependency structure, entity spans, sentiment annotations, question and answer pairs, dialogue (Bunt, 2011) and discourse annotation (Louwerse and Mitchell, 2003), and augmenting existing corpora with NLP annotations;
- developing methods and tools for improving ASR acoustic and language model training and adapting NLP models and pipelines to conversational applications;
- developing tools that allow collecting conversational speech and recording real, spontaneous conversations that could be crowd-sourced and released openly like Librispeech;
- designing open solutions that can serve as common benchmarks for joint ASR+NLP tasks to monitor the progress of the field;
- organizing crowd-sourcing collection efforts similar to Mozilla CommonVoice where users could donate their phone calls and/or transcriptions;
- constructing new ASR quality measures, based on more richly annotated data, to better evaluate various aspects of transcription quality.

4 Conclusions

We argue that contemporary ASR systems cannot cope with spontaneous human conversations satisfactorily, contrary to many beliefs in the NLP community. To substantiate our claim we compile a benchmark consisting of real-world, multi-domain phone call spontaneous conversations on which we observe higher WERs than on traditional datasets. We believe that the overly optimistic perception of ASR accuracy is detrimental to the development of conversational NLP downstream applications. In order to overcome this problem, we ask the community to engage in interdisciplinary research between the academia and industry partners in both ASR and NLP domains and we discuss actions that can be taken. We hope that our call to action will provide some guidelines toward the improvement of ASR systems in the upcoming decade.

References

- Harry Bunt. 2011. The semantics of dialogue acts. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 1–13. Association for Computational Linguistics.
- Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.
- Linguistic Data Consortium et al. 1994. Csr-ii (wsj1) complete. *Linguistic Data Consortium, Philadelphia*, vol. LDC94S13A.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.
- Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.
- Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. Odsqa: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956. IEEE.
- Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.
- Max M Louwerse and Heather Hite Mitchell. 2003. Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse processes*, 35(3):199–239.
- Vimal Manohar, Szu-Jui Chen, Zhiqi Wang, Yusuke Fujita, Shinji Watanabe, and Sanjeev Khudanpur. 2019. Acoustic modeling for overlapping speech recognition: Jhu chime-5 challenge system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6665–6669. IEEE.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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.

- Douglas B Paul and Janet M Baker. 1992. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics.
- Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.
- Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology*, pages 1359–1362.
- Ingo Siegert and Julia Krüger. 2018. How do we speak with alexa: Subjective and objective assessments of changes in speaking style between hc and hh conversations. *Kognitive Systeme*, 2018(1).
- Moritz Stiefel and Ngoc Thang Vu. 2017. Enriching ASR lattices with POS tags for dependency parsing. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 37–47, Copenhagen, Denmark. Association for Computational Linguistics.
- Robert Stojnic, Ross Taylor, Marcin Kardas, Viktor Kerkez, and Ludovic Viaud. 2020. Papers with code - speech recognition. <https://paperswithcode.com/task/speech-recognition>. [Online; accessed 5-October-2020].
- Gabriel Synnaeve. 2020. Wer are we? https://github.com/syhw/wer_are_we. [Online; accessed 5-October-2020].