How Might We Create Better Benchmarks for Speech Recognition?

Alëna Aksënova
Google
New York City, USA
alenaks@google.com

Daan van Esch
Google
Amsterdam, NL
dvanesch@google.com

James Flynn
Google
New York City, USA
jpflynn@google.com

Pavel Golik
Google
New York City, USA
golik@google.com

Abstract

The applications of automatic speech recognition (ASR) systems are proliferating, in part due to recent significant quality improvements. However, as recent work indicates, even state-of-the-art speech recognition systems – some which deliver impressive benchmark results, struggle to generalize across use cases. We review relevant work, and, hoping to inform future benchmark development, outline a taxonomy of speech recognition use cases, proposed for the next generation of ASR benchmarks. We also survey work on metrics, in addition to the de facto standard Word Error Rate (WER) metric, and we introduce a versatile framework designed to describe interactions between linguistic variation and ASR performance metrics.

1 Introduction

The applications of ASR systems are many and varied; conversational virtual assistants on smartphones and smart-home devices, automatic captioning for videos, text dictation, and phone chat bots for customer support, to name a few. This proliferation has been enabled by significant gains in ASR quality. ASR quality is typically measured by *word error rate* (WER), or, informally, the Levenshtein distance between the target transcript and the machine-generated transcript (Levenshtein, 1966; Wang et al., 2003)—see Section 3.

Current state-of-the-art accuracy is now in low-single-digits for the widely used Librispeech benchmark set (Panayotov et al., 2015), with e.g. Zhang et al. (2020) achieving a WER of 1.4%. However, as Szymański et al. (2020) have pointed out, overall, our current ASR benchmarks leave much to be desired when it comes to evaluating performance across multiple real-world applications. Typical benchmark sets beyond Librispeech include TIMIT (Garofolo et al., 1993), Switchboard (Godfrey et al., 1992), WSJ (Paul and Baker, 1992), CALLHOME (Canavan et al., 1997), and Fisher (Cieri et al., 2004).[1] These benchmark sets cover a range of speech use cases, including read speech (e.g. Librispeech), and spontaneous speech (e.g. Switchboard).

However, with many ASR systems benchmarking in the low single digits, small improvements have become increasingly difficult to interpret, and any remaining errors may be concentrated. For example, for Switchboard, a considerable portion of the remaining errors involve filler words, hesitations and non-verbal backchannel cues (Xiong et al., 2017; Saon et al., 2017).

Furthermore, achieving state-of-the-art results on one of these sets does not necessarily mean that an ASR system will generalize successfully when faced with input from a wide range of domains at inference time: as Likhomanenko et al. (2020) show, "no single validation or test set from public datasets is sufficient to measure transfer to other public datasets or to real-world audio data". In one extreme example, Keung et al. (2020) show that modern ASR architectures may even start emitting repetitive, nonsensical transcriptions when faced with audio from a domain that was not covered at training time—even in cases where it would have achieved perfectly acceptable Librispeech evaluation numbers. Inspired by Goodhart's law, which states that any measure that becomes a target ceases to be a good measure, we argue that as a field, it behooves us to think more about better benchmarks in order to gain a well-rounded view of the performance of ASR systems across domains.

In this paper, we make three contributions. First, we provide a taxonomy of relevant domains, based on our experience developing ASR systems for use in many different products, with the goal of helping make next-generation benchmarks as representative as possible (Biber, 1993). Second, we argue that optimizing only for WER, as most current benchmarks imply, does not reflect considerations that are ubiquitous in real-world deployments of ASR technology: for example, pro-

[1] For an overview of such datasets and benchmarks, see https://github.com/syhw/wer_are_we. Additionally, FAIR recently released the Casual Conversations dataset intended for AI fairness measurements (Hazirbas et al., 2021).

Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 22–34

duction considerations such as latency and compute resources can imply additional interrelated optimization objectives. We survey relevant work on additional metrics that can be used to measure ASR systems. Third, we describe what metadata would be useful in next-generation benchmark data sets in order to help analyze the interaction between linguistic variation and performance of ASR systems—for example, to measure how well an ASR system holds up in the face of sociolinguistic variation within the target language, or second-language accents, as in e.g. Feng et al. (2021).

2 ASR Use Cases

With ASR use cases spanning many applications and tasks, ideally ASR systems would be robust to various classes of variation in speech input. For example, an ASR system which provides automatic captions for video meetings would recognize words from many different semantic fields, adaptable to the topic of the meeting. Speech characteristics may also vary across domains: for example, the speech style used when dictating text messages differs from the style of a group conversation, where speakers may occasionally talk over each other.

An ideal benchmark set would include what we will call 'horizontal' and 'vertical' variation. Horizontal challenges refer to a wide variety of scenarios where ASR may be used, while vertical challenges involve e.g. diversity in topics, encoding formats, and others.

2.1 Horizontals: ASR applications

ASR application domains can be roughly subdivided based on the number of speakers, the mode of speech (spontaneous vs. prepared speech) and the intended recipient (human or device). An ideal benchmark set would cover as many of these horizontals as possible—e.g. through merging existing benchmark sets, as does Likhomanenko et al. (2020), and adding additional data to cover any gaps.

Dictation *Text dictation* is a popular use case of ASR systems — one of the first successful commercial applications with broad appeal. This feature serves both convenience and accessibility, allowing users to enter text without manually typing. Dictation tends to involve relatively slow speech, typically that of a single speaker, who is aware they are interacting with a device, and who may consciously modify their speech patterns to facilitate device understanding (Cohn et al., 2020). Dictation may have applications in many fields. One with many idiosyncratic challenges is *medical dictation*, where ASR systems are used to help medical personnel take notes and generate medical records (Miner et al., 2020; Mani et al., 2020). This poses challenges in the support of domain-specific jargon, which we will discuss in subsection 2.2. In a related application, *dictation practice* is sometimes used by language learners, often in combination with a pronunciation feedback system (McCrocklin, 2019). In other contexts, transcription of dictated audio may be part of a composite pipeline, such as in *automatic translation*, where the initial transcript feeds a subsequent system for translation to another language.

Voice Search and Control *Voice search* and other *conversational assistant* products enable users to access information or invoke actions via spoken input. Similar to dictation, audio in such settings is typically single-speaker, with human-to-device characteristics. Compared to dictation, queries may be somewhat shorter, and may contain proper nouns (e.g. place names or business names). Semiotic-class tokens such as times (Sproat et al., 2001) are also more common in this setting. A related type of human-to-device speech is *interactive voice response (IVR)*, where callers to customer support may first interact with a voice chatbot, which can help gather information prior to redirecting the call, or potentially resolve issues itself. (Inam et al., 2017).

Voicemails, Oration, and Audiobooks While dictation users may modify their speech based on the knowledge that they are dictating directly to a device, ASR systems may also be used to help provide transcriptions for voicemail messages (Padmanabhan et al., 2002; Liao et al., 2010), parliamentary speeches (Gollan et al., 2005; Steingrímsson et al., 2020), and so on. Such settings, while still typically single-speaker, include artifacts of spontaneity—e.g. fillers or hesitations like 'uh', backchannel speech, as well as disfluencies, false starts, and corrections (Jamshid Lou and Johnson, 2020; Mendelev et al., 2021; Knudsen et al., 2020). Transcribing *audiobooks* includes elements of dictation and oration: due to their read-speech nature, audiobooks typically contain less spontaneity than typical human-to-human speech (Igras-Cybulska et al.), but they are usually more natural than human-to-device speech.[2]

Conversations and Meetings In settings such as *human-to-human conversations*, the task of the ASR system typically involves transcribing spontaneous speech among several participants within a single audio recording. For example, *meeting transcription*

[2]Transcription of audiobooks is a primary goal of Librispeech (Panayotov et al., 2015), one of the most common benchmarks for ASR today, even though practically speaking, transcribing audiobook audio is not a common task for most real-world ASR systems—given that audiobooks are typically produced based on an existing 'transcription', namely the ground-truth written text of the book.

can help to improve accessibility of video meetings, or may serve to document conversations (Kanda et al., 2021); see e.g. Janin et al. (2004); Carletta et al. (2005) for relevant data sets. Another use case for transcriptions of human-to-human conversations is *customer-agent conversations*, as well as other types of *telephony*, which can help monitor the quality of phone-based customer service.

Podcasts, Movies and TV *Podcast transcription* forms a related, and fast-growing, application area, with recent data sets including Clifton et al. (2020). Podcast transcription is in some ways similar to the long-standing task of automatically transcribing *interviews*, e.g. to help make them more accessible, as in various oral-history projects (Byrne et al., 2004). Finally, another similar use case is the transcription of motion pictures, including documentaries, which may require increased robustness to non-speech audio, such as music and special effects. Spontaneous speech is common to these human-to-human, multi-speaker settings, with fillers such as 'uh', overlap, and interruption between speakers. We draw a distinction between movie subtitling and TV closed captioning. Subtitling is an 'offline' task in that the entire audio is available to the ASR system at recognition time, and the setting allows for multiple passes, including human post-editors. Compare to closed captioning, where streaming ASR processes a live broadcast with tight latency constraints. Additionally, these two modes have different transcription conventions and formatting requirements. Subtitles often contain non-verbal cues that support comprehension for hearing impaired, and are optimized for readability. Conversely, closed captions are often projected in upper case with fewer constraints, such as line breaks, to denote speaker turns.

2.2 Verticals: Technical challenges

ASR applications do not just differ in the style of speech. Other dimensions include: the semantic content of the input speech (a lecture about nuclear physics involves very different terminology than a phone conversation to set up a car maintenance appointment), the audio encoding format, and sample rate, among others. Again, the *ideal* benchmark should cover as many of these factors as possible.

Terminology and Phrases ASR systems applied to a wide range of domains need to recognize hundreds of thousands, if not millions, of distinct words. Such systems typically involve a language model trained on large volumes of text from multiple sources. To benchmark an ASR system's capability across a wide range of topics, test sets could include terms and phrases from many different fields: consider medical terminology (e.g. 'ribonucleotides'), historical phrases (e.g. 'Yotvingians'), and many more. ASR systems should also be savvy to neologisms (e.g. 'doomscrolling'), although, admittedly, the fast-changing nature of neologisms and trending phrases makes this particularly challenging. Another area that deserves special attention in measurements is loanwords, which may have pronunciations that involve unusual grapheme-to-phoneme correspondences; such words may even necessitate personalized pronunciation learning (Bruguier et al., 2016).

Speed Recordings where speech is significantly faster or slower than average may pose additional recognition challenges (Siegler and Stern, 1995; Fosler-Lussier and Morgan, 1999), so the ideal benchmark should also cover samples with various speech rates. This is particularly important for paid services, where users sometimes artificially speed up the recordings or cut out easily detectable portions of silence in order to reduce costs. Such processing can introduce unnatural shifts in pitch and add confusion to the punctuation at speaker turn, and sentence boundaries.

Acoustic Environment The setting in which the input audio was recorded (real-life or phone conversation, video call, dictation) can also materially impact ASR performance, and settings with high amounts of background noise can be particularly challenging. Ideally, test sets should be available to measure how robust an ASR system is in the face of background noise and other environmental factors (Park et al., 2019; Kinoshita et al., 2020). The entertainment domain contains a large amount of scenes with background music, which often have lyrics that are usually not meant to be transcribed. Even call center conversations sometimes contain hold music which is not part of the payload of the call.

Encoding Formats Lastly, different audio encodings (linear PCM, A-law, μ-law), codecs (FLAC, OPUS, MP3) and non-standard sample rates such as 17 kHz may affect recognition quality, and should be represented (Sanderson and Paliwal, 1997; Hokking et al., 2016). The same holds for audio that has been up- or down-sampled, e.g. between 8 kHz typical for telephony and 16 kHz or above, for broadcast media.

2.3 Practical Issues

We argue that the more horizontal and vertical areas are covered by a benchmark, the more representative it will be, and hence the more appropriate for measuring ASR progress. There are some practical matters that are also important to consider when creating the ideal benchmark.

Transcription Conventions Creating transcriptions of human speech in a consistent manner can be unexpectedly challenging: for example, should hesitations like 'uh' be transcribed? How should transcribers handle unusual cases like the artist 'dead mouse', which is written as 'deadmau5' by convention? And if a speaker says 'wanna', should the transcription reflect that as such, or should the transcriber transcribe that as 'want to'? The answer to such questions will depend on the downstream use context (e.g. a dialog system, where hesitations may be useful, or an email message, where they may need to be omitted instead). For example, while in closed captioning or podcast transcriptions omitting repetitions, disfluencies, and filler words (e.g. "like", "kind of") is considered desirable, this might not be appropriate for some other ASR domains such as subtitling. Defining and applying a comprehensive set of transcription conventions, as e.g. Switchboard (Godfrey et al., 1992) and CORAAL (Kendall and Farrington, 2020), is critical in building high-quality data sets. It is also important to detect and correct transcription errors in annotated corpora (Rosenberg, 2012).

Perhaps the most important choice in such transcription conventions is whether to adopt 'spoken-domain' transcriptions, where numbers are spelled out in words (e.g. 'three thirty'), or 'written-domain' transcriptions, where they are rendered in the typical written form ('3:30'). Many data sets use spoken-domain transcriptions only, but often in real-world ASR deployments it is valuable for readability and downstream usage (e.g. by a natural-language understanding system), to have fully-formatted, written-domain transcripts, as described by O'Neill et al. (2021)—who also provide a written-domain benchmark data set.

Representativeness For any ASR test set, at least two considerations come into play: first, how closely does the test set approximate reality; and second, is the test set sufficiently large to be representative? For example, test sets that are intended to measure how well an ASR system deals with speech with background noise should have a realistic amount of background noise: not too little, but also not too much—e.g. to the point that even human listeners stand no chance of transcribing the audio correctly. Adding noise artificially, as established e.g. by the Aurora corpora (Pearce and Hirsch, 2000; Parihar and Picone, 2002), does not take into account the Lombard effect. In terms of size, analyses akin to Guyon et al. (1998) are helpful to ensure that any change is statistically significant; we are not aware of much work along these lines for ASR systems specifically, but it seems like it would be worthwhile to explore this area more. The ultimate goal should be to increase the predictive power of error metrics.

3 Metrics: WER and Beyond

Assume, for the sake of argument, that an impressive selection of test sets has been collected in order to create our imagined ideal next-generation benchmark for ASR, covering many use cases, technical challenges, and so on. The performance of an ASR system could now be measured simply by computing a single, overall WER across all the utterances in this collection of test sets—and a system that yields lower WER on this benchmark could be said to be 'better' than a system with higher WER.

However, in a real-world deployment setting, the question of which system is 'best' typically relies on an analysis of many metrics. For example, imagine a system with a WER of 1.5% but an average transcription latency of 2500 milliseconds, and another system that achieves 1.6% WER but a latency of only 1250 milliseconds: in many settings, the second system could still be more suitable for deployment, despite achieving worse WER results. Of course, 'latency' itself is not a well-defined term: sometimes the measurement is reported as the average delay between the end of each spoken word and the time it is emitted by the ASR system, while in other cases the measure is based only on the first or the last word in an utterance. Neither is well-defined in presence of recognition errors. Yet another kind of latency is end-to-end latency, involving everything between the microphone activity and the final projection of results, including network overhead and optional post-processing like capitalization, punctuation etc. A "pure" ASR latency metric ignores those and focuses on the processing time of the recognizer, while latency in the context of voice assistant commands may consider the delay before successful recognition of a command, which might sometimes precede the actual end of utterance. In this section, we describe how, much like latency, even WER itself has many nuances, and we point to other metrics, beyond WER and latency, that can be considered account when measuring ASR systems.

3.1 WER

The workhorse metric of ASR is the Word Error Rate, or WER. Calculating WER is relatively easy on spoken-domain transcriptions with no formatting (e.g. 'set an alarm for seven thirty') but quickly becomes a nuanced matter when processing written-domain transcriptions—for example, if the ground truth is provided as 'Set an alarm for 7:30.' with capitalization and punctuation, is it an error in WER terms if the system emits lowercase 'set' instead of uppercase 'Set', as given in the ground truth? Typically, for

standard WER calculations in such scenarios, capitalization and word-final punctuation is not considered to be a factor, and other metrics are calculated for fully-formatted WER—e.g. case-sensitive WER, where 'set' vs 'Set' would be considered an error.

WER can also be calculated on only a subset of relevant words or phrases: for example, it may be helpful to compute separate error rates for different kinds of semiotic classes, such as spoken punctuation, times, or phone numbers—as well as for different semantic areas, such as relevant domain terminology vs. generic English words. The assessment of ASR quality on rare phrases is yet another issue—average WER does not always adequately reflect how well an ASR system picks up rare yet important words, suggesting it may be valuable to know WER for common and less common words. A related approach is to use precision-recall, e.g. as Chiu et al. (2018) do for medical terminology. Such 'sliced' approaches can help provide insight into the recognition quality of words or phrases that are particularly salient in a given setting. For example, if a system that is intended for use in a voicemail transcription setting achieves 3% overall WER, but it mistranscribes every phone number, that system would almost certainly not be preferred over a system that achieves 3.5% overall WER, but that makes virtually no mistakes on phone numbers. As Peyser et al. (2019) show, such examples are far from theoretical; fortunately, as they show, it is also possible to create synthetic test sets using text-to-speech systems to get a sense of WER in a specific context. Standard tools like NIST SCLITE[3] can be used to calculate WER and various additional statistics.

Importantly, it is possible to calculate the local WER on any level of granularity: utterance, speaker turn, file, entire recording etc. The *average* WER alone, weighted by the number of words, is not sufficient to describe the shape of the distribution over the individual local measurements. Given two ASR systems with identical WERs, we almost always prefer the one with the lower standard deviation, as it reduces the uncertainty w.r.t. the worst case. A more accurate metric that samples the shape of the distribution consists of percentiles (e.g. 90, 95 or 99) that are more suitable to provide an upper bound. Additionally, reporting the standard deviation allows researchers to judge whether an improvement in WER is significant or just a statistical fluctuation. The same argument holds true for latency.

Finally, WER can also be calculated on not just the top machine hypothesis, but also on the full *n-best* list, as in e.g. Biadsy et al. (2017).

[3] https://www.nist.gov/itl/iad/mig/tools

3.2 Metadata about Words

Correctly transcribing speech into text is the most critical part of an ASR system, but downstream use cases may require more than just a word-by-word textual transcription of the input audio. For example, having *per-word confidence scores* can be helpful in dialog systems (Yu et al., 2011); having accurate timestamps at the word level is essential in many application of the long form domain, such as closed captioning, subtitling and keyword search; having *phonemic transcriptions* for every word enables downstream disambiguation (e.g. when the transcription gives 'live', did the user say the adjective [lɪv] or the verb [laɪv]); and emitting *word timings* to indicate where each word appeared in the audio can be important for search applications, especially for longer recordings. The ideal ASR benchmark would also make it possible to verify this metadata: for example, if it is possible to use forced alignment to infer where in the audio words appear, and to check how accurately an ASR system is emitting word timings (Sainath et al., 2020a). *speaker diarization* is yet another type of of metadata that can be emitted at a per-word or per-phrase level, for which independent benchmarks already exist (Ryant et al., 2021).

3.3 Real-Time Factor

A general metric for the processing speed is the real-time factor (RTF), commonly defined as the ratio between the processing wall-clock time and the raw audio duration (Liu, 2000). Streaming ASR systems are required to operate at an RTF below one, but in applications that do not require immediate processing an RTF over one might be acceptable. As with WER and latency, RTF samples form a distribution, whose shape is important in understanding the behavior in the worst case. The process of finding the most likely hypothesis in ASR (often referred to as "decoding" for historical reasons) requires an efficient exploration of the search space: a subset of all possible hypotheses. The larger the search space, the slower the search, but the more likely is the recognizer to find the correct hypothesis. A small search space allows for quick decoding, but often comes at the cost of higher WER. It is common to report an RTF vs WER curve which shows all possible operating points, allowing for mutual trade off. Note this definition operates with the wall-clock time, thus ignoring the hardware requirements. It is common to normalize the RTF by the number of CPU cores and hardware accelerators.

3.4 Streaming ASR

For ASR systems that stream output to the user while recognition is ongoing, as in many voice assistant

and dictation applications, additional metrics will be useful, e.g. measuring the *stability of partial results*, which reflects the number of times the recognizer changes previously emitted words while recognizing a query (Shangguan et al., 2020). A related dimension is *quality of the intermediate hypotheses*: a streaming system that emits highly inaccurate intermediate hypotheses can yield a jarring user experience, even if the final hypothesis achieves an acceptable WER. This is particularly important in combination with a downstream application like machine translation that can be very sensitive to corrections in partial hypotheses (Ansari et al., 2020).

Yet another factor is streaming latency, e.g. how quickly partials are emitted (Shangguan et al., 2021), and more generally, the delay between the end of the user's input and the finalized transcription (Sainath et al., 2020b; Yu et al., 2021). The accuracy of the *endpointer* module can significantly affect this latency: endpointers need to strike the right balance between keeping the microphone open while the user may still continue speaking (e.g. if the user pauses briefly to collect their thoughts), while closing it as soon as the user is likely to be done speaking, and a number of relevant endpointer metrics can be calculated, as in e.g. Li et al. (2020).

3.5 Inference and Training

Latency is influenced by many factors beyond the quality of the endpointer: for example, the number of parameters in the ASR model, the surrounding software stack, and the computational resources available will impact the duration of the recognition process for an audio recording, in both streaming and non-streaming - batch recognition settings. Compressing models can help them run faster, and in more settings (Peng et al., 2021), although the impact of shrinking models should be measured carefully (Hooker et al., 2020a,b).

Beyond inference, training may also be worth benchmarking in more detail: factors such as the number of parameters in the model, the model architecture, the amount of data used, the training software, and the hardware available will influence how long it takes to train an ASR model using a given algorithm. Benchmarks such as MLPerf (Mattson et al., 2020) do not yet incorporate speech recognition, but this may be worth exploring in the future.

3.6 Contextual Biasing

Certain phrases or words are sometimes expected in dialogue contexts (e.g. 'yes' or 'no'), along with particular types of words (e.g. brand names in the context of shopping). In such cases, ASR systems may allow for *contextual biasing* to increase the language model probability of relevant words or phrases (Aleksic et al., 2015). Measuring contextual biasing typically involves evaluating a relevant test set twice: once with, and once without the contextual biasing enabled (the default behavior). Even when contextual biasing is enabled, it will typically be desirable for the system to continue to recognize other words and phrases without too much of an accuracy impact, so that recognition results remain reasonable in the event that the input does not contain the words or phrases that were expected—typically anti-sets will be used, as described by Aleksic et al. (2015). Contextual biasing plays a key role in classical dialogue systems like IVR.

3.7 Hallucination

In some cases, ASR models can *hallucinate* transcriptions: e.g. providing transcriptions for audio even where no speech is present, or simply misbehaving on out-of-domain utterances (Liao et al., 2015; Keung et al., 2020). Intuitively, this type of errors should be reported explicitly as the "insertion rate", which is calculated as part of the WER anyway. However, insertion errors are rather rare and do not stand out strongly in presence of speech and natural recognition errors.

Measuring whether an ASR system is prone to such hallucinations can be done by running it on test sets from domains that were unseen at training time. In addition, it is possible to employ *reject sets* which contain various kinds of audio that should *not* result in a transcription: for example, such reject sets may cover various noises (e.g. AudioSet Gemmeke et al. (2017)), silence, speech in other languages, and so on.

A related topic is *adversarial attacks*, when a particular message is 'hidden' in audio in a way that humans cannot hear, but which may deceive ASR systems into transcribing in an unexpected way; measuring robustness to such issues would be desirable, but it remains an active area of research—much like the creation of such attacks more broadly (Carlini and Wagner, 2018).

3.8 Debuggability and Fixability

Finally, one aspect of ASR systems that tends to be important for real-world deployments, but which is hard to quantify in a numeric metric, is how easy it is to debug and fix any misrecognitions that may arise. For example, if a new word such as 'COVID-19' comes up which is not yet recognized by the system, it would be preferable if adding such a new word could be done without necessitating a full retrain of the system. While quantifying this property of ASR systems is hard, we believe that the degree to which it is easy to debug and fix any ASR system is worth mentioning.

4 Demographically Informed Quality

As previously discussed, the ideal benchmark for ASR systems would cover as many horizontals and verticals as possible, and would involve various kinds of metrics beyond just WER. Another important dimension, however, would be the availability of demographic characteristics, and analyzing the metrics based on such characteristics. Such demographic characteristics may correlate with linguistic variation—for example, non-native speakers of English may have an accent showing traces of their native language—which may in turn impact ASR performance. Having demographic characteristics can help produce analyses like the one reported by Feng et al. (2021), who analyzed differences in recognition performance for different accents, age ranges, and gender within an ASR system.

The ideal benchmark set, then, should include sufficient metadata to run similar analyses, enabling developers to understand how their system behaves when processing various accents or dialects; to see whether factors like gender and age influence recognition performance in their system. Linguistic variation may take many different shapes, including:

- phonetic differences, e.g. vowel realizations that are specific to a given accent
- phonological differences, e.g. various number of phonemes in different dialects of a language
- lexical differences, e.g. region-specific terms
- syntactical differences, e.g. double-negatives
- voice quality differences, e.g. pitch differences, which are correlated with parameters such as gender and age (Liao et al., 2015)

Fortunately, several data sets already exist with relevant demographic tags for many utterances, e.g. Mozilla Common Voice (Ardila et al., 2020) which offers public data sets across many languages with dialect and accent tags. There are also academic data sets produced by sociolinguists, such as CORAAL for AAVE (Kendall and Farrington, 2020), ESLORA for Galician Spanish (Barcala et al., 2018), the Corpus Gesproken Nederlands for Dutch (van Eerten, 2007), and others. Such corpora provide a useful blueprint for providing such metadata, and we believe that it would be valuable for similar tags to be available for as many other data set as possible. As Andrus et al. (2021) show, at times it will likely be difficult to get the demographic metadata that is needed, but still, getting such data wherever possible is important—as they put it, "what we can't measure, we can't understand".

Even where demographic information is already present in ASR evaluation sets, it can be a valuable to conduct an analysis of the target user base for a deployed ASR system in order to ensure that all relevant tags are available. For example, if a data set has labels for four distinct accents, but the target user base is known from sociolinguistic research to use six distinct accents, this gap will not necessarily be evident when running an analysis of any possible differences among the four accents for which tags are available. It is important to understand the sociolinguistic characteristics of the target user base, and to cover as many of these properties as possible. Given that language has almost infinite variation as you zoom in—in the extreme, everyone has a slightly different voice—this is a task that requires careful sociolinguistic judgement and analysis, calling for interdisciplinary collaboration between linguists and developers of ASR systems.

Even when a rich set of tags is available, it can be difficult to interpret the results. We describe a simple, metric-independent population-weighted visualization framework designed to evaluate ASR systems based on such demographic metadata. Our approach supports the different language variations outlined above, and we propose this analyses as a valuable addition to future benchmarks.

4.1 Population-Weighted Slicing Framework

Factors like accents (native or non-native), dialects, gender, and others can result in linguistic variation, and this may in turn impact ASR performance. Thus it can be valuable to calculate WER, latency, and other metrics not just on a data set as a whole, but

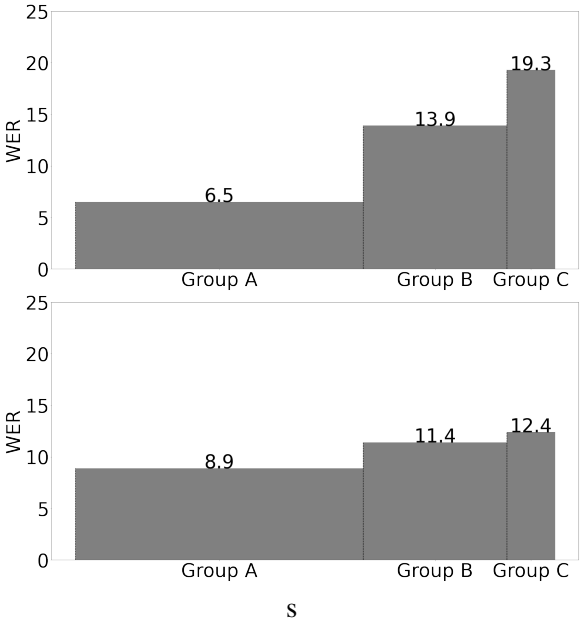

Figure 1: Examples of WER sliced into groups A, B, and C, with the width of the bars reflecting relative sizes of those groups.

also to slice metrics based on such meta-linguistic parameters.

Such sliced metrics can be used to determine any performance gap between groups, and if so, what efforts may need to be undertaken to shrink such gaps. The ideal test set should be representative of the target user base, but as this may be hard to achieve at data collection time, it can make sense to re-weight any metrics based on real-world population statistics: for example, imagine a scenario where 98% of the recordings in a data set come from native speakers, with the remaining 2% coming from non-native speakers. If the target deployment setting involves more like 15% non-native speech, the metrics obtained over the 2% slice of the data set coming from non-native speakers should carry 15% of the weight.

To make such analyses easier, we propose subdividing all speakers into mutually exclusive groups based on relevant linguistic or demographic criteria. For example, consider a scenario where the real-world population is subdivided into 3 mutually exclusive groups: group A (60% of the population), group B (30%), and group C (10%). The two subplots of Figure 1 visualize examples of evaluations of two ASR models for slices corresponding to these groups, with the WER scores represented by the height of the bars, and the width of the bars reflecting the size of the groups.

Even in the actual test data set, group A covers 80% of the test data, with groups B and C accounting for 10% each (i.e. under-representing group B and over-representing group A), this population-weighted framework provides an intuitive way to address this imbalance, and understand how ASR systems perform in the face of linguistic diversity. The average WER of the system can be calculated as an average of all WER scores across population groups, weighted according to the size of those groups—which may differ from the WER obtained by simply calculating the WER on the actual data set, as we have re-weighted based on the real-world distribution.

Importantly, while the average weighted WER is a useful metric, the full distribution should still be understood: continuing the example depicted on Figure 1, the average WER for both scenarios in this case would be 10[4], but the disparity between the various groups in the plot where group C achieves a WER of 19.3% is clearly much bigger in one scenario than another.

Given WER measurements for several groups of speakers, we should also measure the *disparity* of the ASR performance across various groups. In a simplified way, one could calculate the difference between the best-performing and the worst-performing groups, but see Mitchell et al. (2020) for a general discussion of ML fairness metrics. While the WER gap in the best-group and the worst-performing group for the scenario depicted on the second subplot of Figure 1 is 3.5 absolute points, the gap is 12.8 absolute points for the distribution on the first subfigure—despite these two systems having the same average WER, one system is clearly more consistent than another.

Slicing can be based on just a single parameter, such as accent, gender, or age, but in reality, speakers are likely to fall into several categories at once. Therefore, it may make sense to look at *intersectional* groups: for example, ASR performance of 20-30 years old female speakers of Chicano English from Miami. Obtaining such rich metadata, however, may be challenging. Also, the more groups we intersect, the stronger the effect of data sparsity becomes: it may be challenging to fill every bucket with enough samples to obtain solid statistics and to control for all other variables not considered. At any rate, as long as mutually exclusive groups can be defined—whether based on a single parameter or in an intersectional way—this framework can help provide a more thorough understanding of various ASR metrics. Weighting by population also allows re-balancing potentially unbalanced test sets, and gives insight into what kinds of ASR performance would be encountered by different groups.

The goal of this approach is to generate new insights into the ASR accuracy for each slice without making assumptions about the causal interaction between the underlying latent variables. The analytical methods we discuss here are much more detailed than what is commonly employed for ASR system evaluation nowadays, but this level of detail is more usual in the field of variationist sociolinguistics, suggesting potential for future collaborations (Labov, 1990; Grama et al., 2019).

4.2 Defining slices

To evaluate the ASR systems in a framework that we are proposing, it is crucial to define representative and mutually exclusive slices. While the classification we suggest in this section is by no means exhaustive, it can be used as a starting point.

Regional language variation Many languages have regional language variation. For example, in the United States alone, there are 3 main regional groups of dialects: the Inland North, the South, and the West (Labov, 1991), with multiple cities developing their own regional language variants. Such regional variants may involve regional phonology ('get' rhymes with 'vet' in the North, and with 'fit' in the South), and even significant lexical and syntactic

[4]Top subplot: 6.5*0.6 + 13.9*0.3 + 19.3*0.1 = 10; bottom subplot: 8.9*0.6 + 11.4*0.3 + 12.4*0.1 = 10;

differences ('going/planning to' can be expressed as 'fixin' to' in the South). Aksënova et al. (2020) has shown how such regional variation can be explored, and how it can impact ASR performance. Ideally, then, as many regional variants as possible should be covered by the ideal benchmark for a given language.

Sociolects Along with regional differences, there may also also linguistic diversity introduced by speakers of various *sociolects*: in American English, one might think of AAVE, Chicano (Mexican-American) English, and others. For example, AAVE—covered by the CORAAL data set (Kendall and Farrington, 2020)—has distinctive syntactic constructions such as *habitual be* ('She be working') and *perfective done* ('He done run'), along with systematic phonological differences (Wolfram, 2004). And even within a single sociolect such as AAVE there might be linguistic diversity (Farrington et al., 2020). Sociolects may impact ASR quality (Koenecke et al., 2020), and it would therefore be desirable for benchmarks to cover as many sociolects as possible.

L2 background Speech produced by non-native (L2) may reflect some characteristics of their native (L1) language (Bloem et al., 2016), making it important to measure the impact of L2 accents on ASR accuracy. One relevant data set for English is the GMU Speech Accent Archive Weinberger (2015), which collects such data for L2 speakers of English.

Gender, age, and pitch Recognition performance may vary depending on the gender or age of the speaker (Liao et al., 2015; Tatman, 2017; Tatman and Kasten, 2017; Feng et al., 2021). In some cases, as in Common Voice (Ardila et al., 2020; Hazirbas et al., 2021), self-reported metadata is available. Where such information is not available, it may make sense to fall back to a proxy analysis based on pitch—which is known to be correlated with factors such as age and gender—in order to understand whether there are recognition accuracy differences for various pitch buckets, as in Liao et al. (2015).

Speech impairments Accuracy rates of standard ASR systems may also degrade for speech produced by people with speech impairments. Recent work has investigated ways to collect relevant data (Grill and Tučková, 2016; Park et al., 2021), enabling analyses of ASR systems in this area. However, given the high degree of variability in this space, a more robust path at least for the near-term future may be designing personalized ASR systems for people with non-standard speech (Shor et al., 2019). Beyond speech impairments, voice technologies could bring benefits to people with various types of diseases and impairments such as Alzheimer's, Parkinson's, and hearing loss.

5 Conclusion

The ultimate goal of benchmarking should be the ability to predict how well an ASR system is going to generalize to new and unseen data. In the previous sections we have argued that a single aggregate statistic like the average WER can be too coarse-grained for describing the accuracy in a real-world deployment that targets multiple sociolinguistic slices of the population. Ideally, the insights generated by the proposed analysis would be actionable, from the composition of the training data to fine-grained twiddling with a clear objective function.

Before we conclude, we should point out that any benchmark that implemented even a fraction of the metrics outlined above would yield rich amounts of information—which will likely pose challenges in terms of organizing, presenting, and understanding all this material. Model report cards, as outlined by Mitchell et al. (2019), may be a natural way to capture this information for an ASR system—although we would suggest calling them *system* report cards instead, given that most ASR systems do not consist solely of a single monolithic model. Given the sheer amount of variation in the ways in which people speak, and a large number of technical factors, measuring ASR systems is a complicated task. Today's benchmarks clearly leave room for improvement, whether it is through covering more horizontal domains (different kinds of speech), measuring the impact of cross-cutting vertical issues (e.g. factors like background noise), using more metrics than just WER (e.g. latency), and including demographic characteristics. We hope that our survey of these areas, and the simple population-weighted visualization framework we introduced, can help improve future benchmarks—not just for English, but also for the thousands of other languages spoken in our world today. This will clearly be a long-term journey, but it will be very important for the field as a whole to find ways to measure ASR systems better as speech recognition research continues to advance.

6 Acknowledgements

We thank our colleagues on the Google Speech team for many thoughtful discussions on this topic, especially Petar Aleksic, Geoff Fischer, Jonas Fromseier Mortensen, David Garcia, Millie Holt, Pedro J. Moreno, Pat Rondon, Benyah Shaparenko, and Eugene Weinstein.

References

Alëna Aksënova, Antoine Bruguier, Amanda Ritchart-Scott, and Uri Mendlovic. 2020. Algorithmic exploration of American English dialects. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno. 2015. Bringing contextual information to Google speech recognition. In *Proc. Interspeech 2015*, pages 468–472, Dresden, Germany.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proc. ACM Conference on Fairness, Accountability, and Transparency*, page 249–260.

Ebrahim Ansari et al. 2020. Findings of the IWSLT 2020 evaluation campaign. In *Proc. International Conference on Spoken Language Translation*, pages 1–34.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France.

Mario Barcala, Eva Domínguez, Alba Fernández, Raquel Rivas, M. Paula Santalla, Victoria Vázquez, and Rebeca Villapol. 2018. El corpus ESLORA de español oral: diseño, desarrollo y explotación. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 5(2):217–237.

Fadi Biadsy, Mohammadreza Ghodsi, and Diamantino Caseiro. 2017. Effectively building tera scale maxent language models incorporating non-linguistic signals. In *Proc. Interspeech 2017*, pages 2710–2714, Stockholm, Sweden.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4):243–257.

Jelke Bloem, Martijn Wieling, and John Nerbonne. 2016. *The Future of Dialects: Automatically identifying characteristic features of non-native English accents*, chapter 9. Language Science Press, Berlin, Germany.

Antoine Bruguier, Fuchun Peng, and Françoise Beaufays. 2016. Learning personalized pronunciations for contact name recognition. In *Proc. Interspeech 2016*, pages 3096–3100, San Francisco, CA, USA.

William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajič, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42. Web Download. Philadelphia: Linguistic Data Consortium.

Jean Carletta et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39, Edinburgh, United Kingdom.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops (SPW)*, pages 1–7, San Francisco, CA, USA.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018. Speech recognition for medical conversations. In *Proc. Interspeech 2018*, pages 2972–2976, Hyderabad, India.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proc. International Conference on Language Resources and Evaluation*, pages 69–71, Lisbon, Portugal.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proc. International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain.

Michelle Cohn, Melina Sarian, Kristin Predeck, and Georgia Zellou. 2020. Individual variation in language attitudes toward voice-AI: The role of listeners' autistic-like traits. In *Proc. Interspeech 2020*, pages 1813–1817, Shanghai, China.

Laura van Eerten. 2007. Over het corpus gesproken Nederlands. *Nederlandse Taalkunde*, 12(3):194–215.

Charlie Farrington, Sharese King, and Mary Kohn. 2020. Sources of variation in the speech of African Americans: Perspectives from sociophonetics. *WIREs Cognitive Science*, 12(3):1–17.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. In *Proc. Interspeech 2021 (submitted)*, Brno, Czech Republic.

Eric Fosler-Lussier and Nelson Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2):137–158.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, New Orleans, LA, USA.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520, San Francisco, CA, USA.

Christian Gollan, Maximilian Bisani, Stephan Kanthak, Ralf Schlüter, and Hermann Ney. 2005. Cross domain automatic transcription on the TC-STAR EPPS corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 825–828, Philadelphia, PA, USA.

James Grama, Catherine E. Travis, and Simon Gonzalez. 2019. Initiation, progression, and conditioning of the short-front vowel shift in Australia. In *Proc. International Congress of Phonetic Sciences (ICPhS)*, pages 1769–1773, Melbourne, Australia.

Pavel Grill and Jana Tučková. 2016. Speech databases of typical children and children with SLI. *PLOS ONE*, 11(3):1–21.

Isabelle Guyon, John Makhoul, Richard Schwartz, and Vladimir Vapnik. 1998. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):52–64.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards measuring fairness in AI: the Casual Conversations dataset.

Rattaphon Hokking, Kuntpong Woraratpanya, and Yoshimitsu Kuroki. 2016. Speech recognition of different sampling rates using fractal code descriptor. In *Proc. International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5, Khon Kaen, Thailand.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2020a. What do compressed deep neural networks forget?

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020b. Characterising bias in compressed models. In *Proc. ICML Workshop on Human Interpretability in Machine Learning*.

Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko, and Marcin Witkowski. Structure of pauses in speech in the context of speaker verification and classification of speech type. *Journal on Audio, Speech, and Music Processing*, 2016(18):1–16.

Itorobong A. Inam, Ambrose A. Azeta, and Olawande Daramola. 2017. Comparative analysis and review of interactive voice response systems. In *Proc. Conference on Information Communication Technology and Society (ICTAS)*, pages 1–6, Durban, South Africa.

Paria Jamshid Lou and Mark Johnson. 2020. End-to-end speech recognition and disfluency removal. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2051–2061.

Adam Janin, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2004. ICSI Meeting Speech LDC2004S02. Web Download. Philadelphia: Linguistic Data Consortium.

Naoyuki Kanda, Guoli Ye, Yu Wu, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2021. Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone. In *Proc. Interspeech 2021 (submitted)*, Brno, Czech Republic.

Tyler Kendall and Charlie Farrington. 2020. The corpus of regional African American language. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project.

Phillip Keung, Wei Niu, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Attentional speech recognition models misbehave on out-of-domain utterances.

Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013, Barcelona, Spain.

Birgit Knudsen, Ava Creemers, and Antje S. Meyer. 2020. Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11:1–10.

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. *Proc. of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2:205–254.

William Labov. 1991. The three dialects of English. In *New Ways of Analyzing Sound Change*, pages 1–44.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Bo Li, Shuo-yiin Chang, Tara N. Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu. 2020. Towards fast and accurate streaming end-to-end ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6069–6073, Barcelona, Spain.

Hank Liao, Chris Alberti, Michiel Bacchiani, and Olivier Siohan. 2010. Decision tree state clustering with word and syllable features. In *Proc. Interspeech 2010*, page 2958 – 2961, Makuhari, Chiba, Japan.

Hank Liao, Golan Pundak, Olivier Siohan, Melissa Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N. Sainath, Andrew Senior, Françoise Beaufays, and Michiel Bacchiani. 2015. Large vocabulary automatic speech recognition for children. In *Proc. Interspeech 2015*, pages 1611–1615, Dresden, Germany.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2020. Rethinking evaluation in ASR: Are our models robust enough?

Jane W. S. Liu. 2000. *Real-time systems*. Prentice Hall, Upper Saddle River, NJ.

Anirudh Mani, Shruti Palaskar, and Sandeep Konam. 2020. Towards understanding ASR error correction for medical conversations. In *Proc. ACL 2020 Workshop on Natural Language Processing for Medical Conversations*, pages 7–11.

Peter Mattson et al. 2020. MLPerf training benchmark. In *Proc. Conference on Machine Learning and Systems*, pages 1–14, Austin, TX, USA.

Shannon McCrocklin. 2019. ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation*, 5(1):98–118.

Valentin Mendelev, Tina Raissi, Guglielmo Camporese, and Manuel Giollo. 2021. Improved robustness to disfluencies in RNN-Transducer based speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada.

Adam S. Miner, Albert Haque, Jason Alan Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A. Arnow, W. Stewart Agras, Li Fei-Fei, and Nigam H. Shah. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*, 3(82).

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, page 117–123, New York, NY, USA.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency*, page 220–229, Atlanta, GA, USA.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *Proc. Interspeech 2021 (submitted)*, Brno, Czech Republic.

Mukund Padmanabhan, George Saon, Jing Huang, Brian Kingsbury, and Lidia Mangu. 2002. Automatic speech recognition performance on a voicemail transcription task. *IEEE Transactions on Speech and Audio Processing*, 10(7):433–442.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Australia.

Naveen Parihar and Joseph Picone. 2002. Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02. Technical report, Mississippi State University.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech 2019*, pages 2613–2617, Graz, Austria.

Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an online infrastructure for collecting AI data from people with disabilities. In *Proc. ACM Conference on Fairness, Accountability, and Transparency*, page 52–63.

Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

David Pearce and Hans-Günter Hirsch. 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition. In *Proc. International Conference on Spoken Language Processing*, pages 29–32, Beijing, China.

Zilun Peng, Akshay Budhkar, Ilana Tuil, Jason Levy, Parinaz Sobhani, Raphael Cohen, and Jumana Nassour. 2021. Shrinking Bigfoot: Reducing wav2vec 2.0 footprint. In *Proc. Interspeech 2021 (submitted)*, Brno, Czech Republic.

Cal Peyser, Hao Zhang, Tara N. Sainath, and Zelin Wu. 2019. Improving performance of end-to-end ASR on numeric sequences. In *Proc. Interspeech 2019*, pages 2185–2189, Graz, Austria.

Andrew Rosenberg. 2012. Rethinking the corpus: Moving towards dynamic linguistic resources. In *Proc. Interspeech 2012*, pages 1392–1395, Portland, OR, USA.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2021. The third DIHARD diarization challenge.

Tara N. Sainath, Ruoming Pang, David Rybach, Basi García, and Trevor Strohman. 2020a. Emitting word timings with end-to-end models. In *Proc. Interspeech 2020*, pages 3615–3619, Shanghai, China.

Tara N. Sainath et al. 2020b. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6059–6063, Barcelona, Spain.

Conrad Sanderson and Kuldip K. Paliwal. 1997. Effect of different sampling rates and feature vector sizes on speech recognition performance. In *IEEE Speech and Image Technologies for Computing and Telecommunications*, volume 1, pages 161–164.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English conversational telephone speech recognition by humans and machines. In *Proc. Interspeech 2017*, pages 132–136, Stockholm, Sweden.

Yuan Shangguan, Kate Knister, Yanzhang He, Ian McGraw, and Françoise Beaufays. 2020. Analyzing the quality and stability of a streaming end-to-end on-device speech recognizer. In *Proc. Interspeech 2020*, pages 591–595, Shanghai, China.

Yuan Shangguan, Rohit Prabhavalkar, Hang Su, Jay Mahadeokar, Yangyang Shi, Jiatong Zhou, Chunyang Wu, Duc Le, Ozlem Kalinli, Christian Fuegen, and Michael L. Seltzer. 2021. Dissecting user-perceived latency of on-device E2E speech recognition. In *Proc. Interspeech 2021 (submitted)*, Brno, Czech Republic.

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. 2019. Personalizing ASR for dysarthric and accented speech with limited data. In *Proc. Interspeech 2019*, pages 784–788, Graz, Austria.

Matthew A. Siegler and Richard M. Stern. 1995. On the effects of speech rate in large vocabulary speech recognition systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 612–615, Detroit, MI, USA.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Steintór Steingrímsson, Starkadur Barkarson, and Gunnar Thor Örnólfsson. 2020. IGC-parl: Icelandic corpus of parliamentary proceedings. In *Proc. ParlaCLARIN Workshop*, pages 11–17, Marseille, France.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proc. ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions. In *Proc. Interspeech 2017*, pages 934–938, Stockholm, Sweden.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 577–582, St Thomas, VI, USA.

Steven Weinberger. 2015. Speech accent archive. George Mason University.

Walt Wolfram. 2004. *Handbook of varieties of English: The grammar of urban African American Vernacular English*, pages 111–132. Mouton de Gruyter, Berlin, Germany.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423.

Dong Yu, Jinyu Li, and Li Deng. 2011. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473.

Jiahui Yu, Chung-Cheng Chiu, Bo Li, Shuo-yiin Chang, Tara N. Sainath, Yanzhang (Ryan) He, Arun Narayanan, Wei Han, Anmol Gulati, Yonghui Wu, and Ruoming Pang. 2021. FastEmit: Low-latency streaming ASR with sequence-level emission regularization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada.

Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. In *Proc. NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.