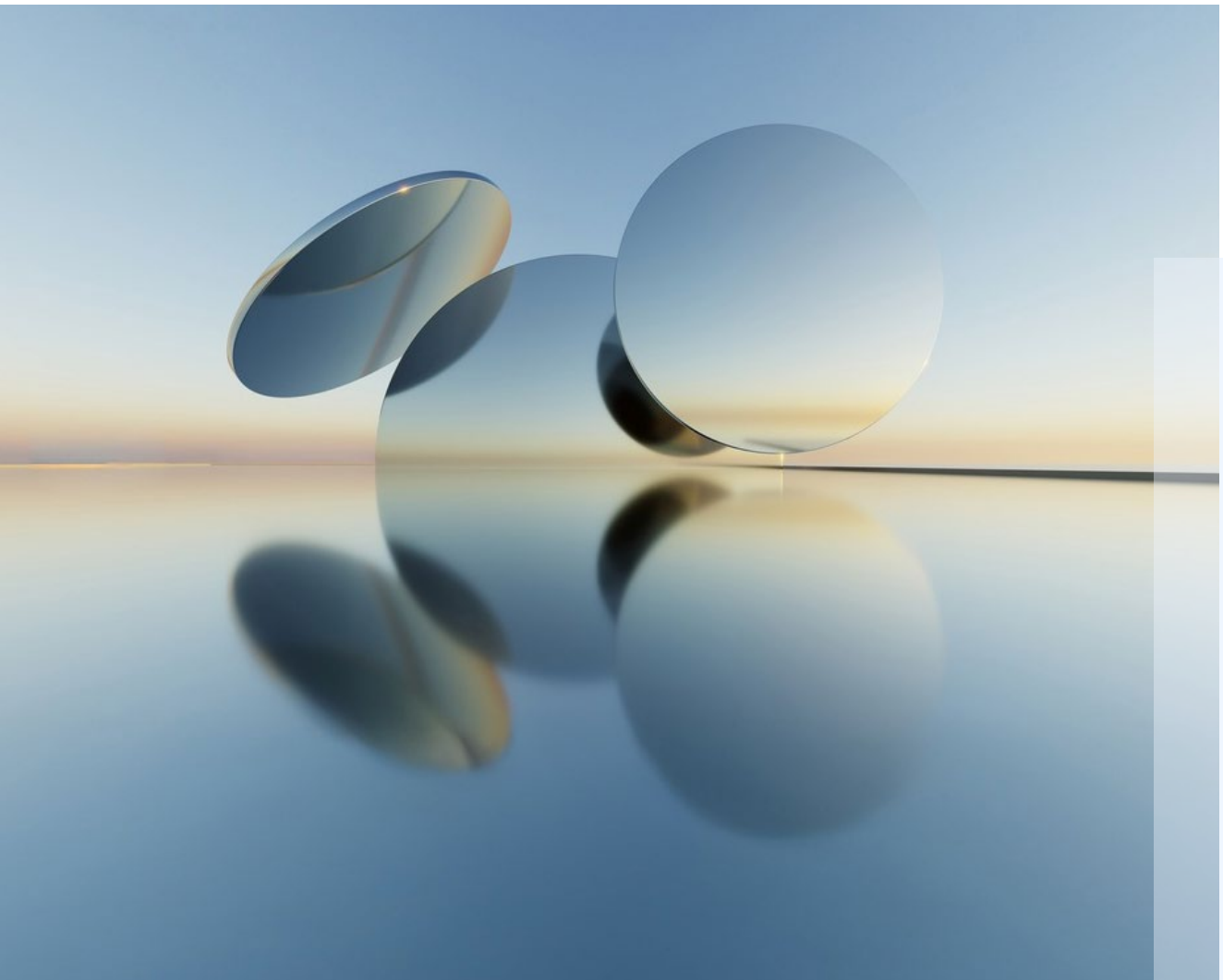
Appen

BUILDING PRODUCTION-REPRESENTATIVE SPEECH BENCHMARKS TO IMPROVE SPEECH MODEL PERFORMANCE



Executive Summary

Automatic Speech Recognition (ASR) has become foundational infrastructure for a generation of AI products such as voice assistants, meeting transcription, and voice agents. However, the benchmarks used to evaluate these systems no longer reflect the conditions under which they are deployed. Widely cited benchmarks like LibriSpeech are predominantly built on clean, scripted, accent-narrow speech that bears little resemblance to production conditions, where speech is spontaneous, accents are diverse, and environments are noisy. This is compounded by benchmaxxing where ASR models are tuned to climb public ASR leaderboards without delivering corresponding gains in real-world performance. The gap between reported benchmark performance and actual user experience is large, empirically documented, and systematically under-measured by the benchmarking datasets the field currently anchors on.

This whitepaper presents Appen's methodology for constructing high-fidelity, production-representative speech benchmarking datasets designed to close this gap. The methodology is structured around a five-stage workflow covering benchmark scoping, contributor sourcing and qualification, speech design, speech recording, and speech transcription. We further outline our end-to-end approach to operationalising these benchmarks once they are built which encompasses normalisation of model inference outputs, computation of error metrics, and dimension-level decomposition.

We highlight our ongoing partnership with Hugging Face, where Appen has developed high-quality scripted and conversational English speech datasets spanning multiple accents to support the expansion of the Hugging Face Open ASR Leaderboard. These speech datasets are benchmaxxing repellent because Hugging Face deliberately keeps them private rather than publishing them openly. Since ASR model builders cannot see or train on benchmarking data they don't have access to, these datasets are far less vulnerable to benchmark-specific optimization which makes the leaderboard a more trustworthy measure of real-world ASR performance.

Appen combines over two decades of speech-data expertise with the operational infrastructure to deliver realistic speech benchmarks. With a pre-vetted global workforce covering 500+ languages and 100+ markets, a rigorous quality-assurance framework, and multilingual expertise built over a 30+ year history, Appen delivers the benchmarking infrastructure that ASR and speech foundation-model teams need to drive measurable performance improvements in their speech models.

Current State of Speech Benchmarks

Current speech benchmarks face two distinct but compounding problems that together undermine their value as signals of real-world ASR performance. The first is benchmaxxing where models are tuned to climb public leaderboards without delivering corresponding gains in real-world performance. The second is insufficient representation of real-world production conditions: widely cited benchmarks underrepresent the speaking styles, accent distributions, and acoustic environments that define actual deployment. The result of these two issues is a large and empirically documented gap between reported benchmark performance and actual user experience. SOTA ASR models report near-human accuracy on public test sets, yet their performance degrades significantly under real-world production conditions. Closing this gap requires both benchmaxxing-resistant evaluation practices and benchmark data that genuinely reflects the conditions models will encounter in production.

Benchmaxxing

Benchmaxxing refers to benchmark-specific optimization where models are tuned to improve their leaderboard performance without delivering corresponding gains in real-world performance. It's essentially Goodhart's Law in action where a measure ceases to be a good measure once it becomes a target. The very qualities that make a benchmark valuable such as standardization and openness, also make it more susceptible to this kind of gaming. On a public benchmark like the Hugging Face Open ASR Leaderboard where test sets, evaluation scripts, and UI code are all openly available, ASR model builders

can deliberately or inadvertently optimize against those public test sets, or seek out training data that closely resembles particularly difficult public datasets in order to boost their ranking. The result is an ASR model that climbs the leaderboard but doesn't generalize to real-world tasks.

To counteract benchmaxxing, the Hugging Face Open ASR Leaderboard now benchmarks ASR models against private datasets provided by data providers such as Appen that are not published openly. Because model developers cannot see, train on, or directly target data they don't have access to, these private sets are far less vulnerable to benchmark-specific optimization or test-set contamination, which in turn increases the trustworthiness of the leaderboard. Several additional design choices reinforce this "benchmaxxing repellent" property: per-split scores are intentionally withheld so developers can't tune to a specific data provider or accent, the data providers have agreed not to supply this exact data to their clients, and using multiple data providers balances out any residual advantage a model might gain from training on a similar distribution.

Insufficient representation of real-world production conditions

A) Limited coverage of spontaneous conversations

Widely cited benchmarks like LibriSpeech are predominantly built on scripted speech leading to modern ASR systems achieving near-zero WERs on LibriSpeech while struggling to generalize to real-world scenarios (Parcollet et al., 2025). Recent empirical work bears this out: one study reported that Whisper's WER climbed to roughly 31% when evaluated on a dataset of spontaneous conversations between pairs of participants (Xiao et al., 2025). Another observed a similar pattern reporting that a zero-shot Whisper model performed acceptably on Austrian German read speech at 11.8% WER but degraded sharply to 41.8% WER on casual conversations between closely-related speakers (Linke et al., 2025). These empirical findings underscore the need for more realistic, conversational ASR benchmarks that provide an accurate assessment of performance in a production environment where speech contains informal phrasing, disfluencies such as pauses and interruptions, and is unstructure.

B) Limited coverage of multi-speaker and overlapping speechs

Most public benchmarks evaluate single-speaker, pre-segmented utterances, yet a large share of production audio such as meetings, customer-service calls, and medical consultations involve two or more speakers with naturally occurring overlap. The MISP 2025 Challenge demonstrated that leading systems struggle on multi-party meetings, with the winning entry posting a concatenated minimum-permutation character error rate of ~11% on its multi-speaker recognition track (Huang et al., 2025). This echoes another study where a SOTA LLM-based multi-talker ASR system achieved 7.3% WER on the LibriMix two-speaker overlap test set and 32.9% WER on the AMI single-distant-microphone meeting test set, an order of magnitude higher than the sub-2% WERs the same class of SOTA models report on LibriSpeech test-clean (He et al., 2025). Without benchmarks that include realistic two-speaker dialogue and three-plus-party scenarios with overlap, aggregate WER on standard test sets systematically overstates how well models will perform on multi-speaker production audio.

C) Lack of accent and demographic diversity

Commonly cited English benchmarks overwhelmingly reflect Native American and British English accents. Leading ASR models that achieve WERs below 10% on LibriSpeech and GigaSpeech see their WERs balloon to 20–40%+ when tested on African-accented English (Ashungafac et al., 2025). This represents a 2 - 4x performance degradation that standard benchmarks fail to surface. Additionally, the ASR-FAIRBENCH evaluation framework, applied to leading ASR systems on the Fair-Speech corpus of 26,500 utterances from 593 demographically self-identified U.S. speakers, found systematic disparities across age, gender, and ethnicity that standard accuracy-only leaderboards do not surface (Rai et al., 2025). ASR systems deployed globally need benchmarks that reflect the full accent distribution of their user base which today's public benchmarks do not provide.

Methodology for Developing Speech Benchmarking Datasetse

Appen combines over two decades of speech-data expertise with a rigorous methodology for constructing high-fidelity, production-representative speech benchmarking datasets. Our methodology is built around the 5-step workflow outlined below, which is designed to deliver granular, diagnostically meaningful signals that teams need to improve real-world speech model performance. We have deep expertise applying this approach across both high and low resource languages, by leveraging our pre-vetted global workforce that covers 500+ languages and 100+ markets globally.

5-step workflow for developing high quality speech benchmarking datasets

01 **Benchmark scoping**

The first and most important stage is scoping. A benchmark is only as useful as its alignment to the production environment it is meant to test, and scoping is where that alignment is defined. Appen works with customers to produce a detailed benchmark specification that captures the target distribution across multiple dimensions for each language in scope.

Dimension	Description
Speaking style	Scripted or conversational speech
Speaker demographics	Demographic attributes reflecting the target user population such as location, age, gender, and race
Accent & dialect	Multiple accents per language (e.g., English-US, English-India, English-Australia, English-Canada) to reflect the user base in production
Environment	Professional recording studio or realistic ambient conditions such as home, office, cafe, street, and vehicle
Device	Covers devices such as professional recording devices, headsets, built-in laptop mics, phone mics, and conference-room devices
Speaker configuration	Single speaker, two-speaker dialogue, multi-party (3+) with naturally occurring overlap
Utterance length	Short commands (<5s), sentences (5–30s), and long-form recordings
Domain	Covers domain specific scenarios and terminology such as technical jargon, product names, and abbreviations

02 Contributor sourcing, recruitment and qualification

Once the benchmark scope is defined, contributors are sourced according to the specification. The recruitment pipeline typically includes demographic screening, a spoken-language assessment that validates proficiency in the target language and accent, and a recording-environment check that confirms the contributor has access to the device and environment types required by the benchmark scope (e.g., a contributor recruited for "English-India, cafe environment, Bluetooth headset" is specifically validated against all three attributes, not just the first).

Appen's robust approach to contributor recruitment is designed to directly address the data quality failures in current crowdsourced speech corpora. Independent audits of widely used crowdsourced speech benchmarks have identified systematic quality issues such as lack of speaker diversity and variable recording quality (Lau et al., 2025).

03 Speech design

Appen designs separate protocols for scripted and spontaneous speech because the two speaking styles require fundamentally different approaches to produce useful benchmarking data.

I) Scripted speech

Scripted speech protocols are used to build benchmarks where specific linguistic phenomena need to be measured with precision. Scripts are drafted to include targeted distributions of phonemes, named entities, numbers, and domain vocabulary.

II) Conversational speech

Appen designs elicitation prompts and conversation topics that reliably produce natural conversational speech – including the disfluencies, turn-taking, overlap, and informal register that characterise realistic production scenarios. Conversations could cover numerous scenarios and configurations such as two-speaker dialogue tasks, multi-party problem-solving scenarios, and domain-specific scenarios such as simulated customer-service interactions or medical-consultation role plays.

04 Speech recording

Audio is recorded in strict accordance with the requirements defined in the benchmark specification. Each recording is accompanied by structured metadata which details dimensions such as speaking style, speaker demographics, environment setting and device type. Without the metadata, a benchmark can report aggregate WER but is unable to provide deeper insight on which conditions are driving failures.

Once recorded, audio recordings undergo thorough quality assurance to verify that it meets the benchmark specification before it enters the downstream transcription pipeline. Automated validation scripts check each submitted file against technical acceptance criteria including sample rate, codec format, and signal-to-noise ratio. In parallel, an experienced human QA team verifies that the audio recordings are aligned with the benchmark specifications. Files that fail any criterion are flagged for re-recording, with only recordings that pass both automated and manual QA gates proceeding to transcription.

05 **Speech transcription**

The speech transcription stage combines automated quality estimation with human review to maximise efficiency and accuracy. Once a first-pass transcript is produced, each segment is scored by an automated quality estimation system on a 0–100% scale. Segments that meet or exceed the quality threshold are classified as locked and proceed directly to QA. Segments that fall below the threshold are classified as unlocked and routed to qualified human post-editors for correction before advancing to QA.

All the transcripts used for the benchmark are audited by senior auditors that review transcripts against the benchmark style guide and validate speaker attribution and turn boundaries for multi-speaker recordings. This ensures that the quality of the delivered ground truth transcription is uniformly high across all segments, regardless of the path each segment took through the workflow.

Our multilingual benchmarking expertise

Appen has a proven track record of building multilingual benchmarks at scale. In a recent project with a frontier research lab, we delivered a benchmarking dataset that spanned 31 diverse languages, including low-resource languages like Czech, Bengali, and Farsi. The benchmark required locally relevant, culturally tailored questions for each language, demanding native-speaker expertise. Appen sourced contributors from its pre-vetted global crowd, applying a multi-step qualification pipeline that included KYC checks, language comprehension tests, and project-specific calibration tasks to ensure every contributor met the linguistic and domain standards required by the customer.

This approach is the same methodology Appen applies to speech benchmarking programmes, where we leverage our multilingual infrastructure and QA rigour to produce production-representative benchmarking data across the full range of languages and accents required by customers.

Approach to Operationalising Speech Benchmarks

Once a benchmarking dataset has been delivered, Appen operationalises it through a standardised, end-to-end evaluation workflow that turns the dataset into a reliable, reproducible measurement system that drives performance improvements for speech models.

01

Model inference

Each candidate ASR model is run against the benchmark under controlled inference conditions. Inference configuration is documented comprehensively given modern ASR models exhibit different behaviour under different inference configurations, and uncontrolled inference variation is a common source of contradictory benchmark results.

02

Text normalisation

Text normalisation is a prerequisite for interpretable WER given raw WER on un-normalised text substantially overstates error because it penalises semantically equivalent formatting differences such as punctuation, casing, abbreviations, and digit vs. word form. Academic research has indicated that text normalization yields substantial reductions in WER, with a paper from Microsoft showing that application of an automated WER normalization systems on 35K utterances across four languages yielded an average WER reduction of ~13 % (Guha et al., 2023).

Appen's normalisation pipeline is designed per-language by in-region linguists, and normalised and un-normalised WERs can both be reported so that customers can audit the effect of text normalisation on their specific use case.

03

WER calculation

Appen calculates Word Error Rate (WER) using the standard formulation derived from the Levenshtein edit distance between the reference transcript and the model hypothesis. For each utterance, the algorithm computes the minimum number of substitutions, insertions, and deletions required to transform the hypothesis into the reference, and WER is expressed as the sum of these three error types divided by the total number of words in the reference.

Appen reports these three error components separately alongside the aggregate WER, because each has distinct diagnostic value: a high substitution rate typically indicates acoustic-model confusion between phonetically similar words, a high deletion rate suggests the model is dropping speech segments (common in noisy or overlapping audio), and a high insertion rate may signal hallucination or erroneous repetition.

04

Dimension level diagnostics and error analysis

The most important output of the evaluation workflow is not the aggregate WER but the granular dimension-level diagnostics. For every benchmarking dataset, Appen can decompose WER along every dimension captured in scoping: by accent, by speaking style, by acoustic environment, by device, by speaker configuration, by utterance length, and by domain category. Dimension level reporting provides the granular insight that drives performance improvements in speech models (e.g., "is my model failing on this specific accent?").

Appen's Partnership with Hugging Face

What Hugging Face is building

The Hugging Face Open ASR Leaderboard is the industry standard for transparent, reproducible evaluation of speech recognition models. As the leaderboard matures, Hugging Face is actively expanding evaluations across additional accents, languages, and domains to better reflect real-world deployment conditions. Hugging Face is also pursuing the use of private, non-public evaluation speech sets to minimise the risk of test-set contamination and benchmark optimization (Srivastav et al., 2025).

How Appen supported

As part of the partnership, Appen developed and provided Hugging Face with high-quality scripted and conversational speech datasets covering multiple accents of English including English-US, English-Australia, English-Canada, and English-India. These datasets were built using the same methodology described in this paper: targeted contributor sourcing against specific demographic and accent requirements, scripted and conversation speech protocols, and rigorous multi-stage quality assurance conducted by experienced linguists. Additionally, these speech datasets are benchmaxxing repellent because Hugging Face deliberately keeps them private rather than publishing them openly.

Results

With the inclusion of Appen's private datasets, the Hugging Face Open ASR Leaderboard has become more trustworthy as ASR model builders are unable to use public test sets or find training data that closely resembles a difficult dataset to boost their leaderboard ranking.

Conclusion

Current speech benchmarks are insufficient because they are predominantly built on clean, scripted, accent-narrow speech that no longer reflects the conditions under which ASR systems are deployed. Until benchmarks are designed around production-representative conditions, the gap between reported performance and actual user experience will remain invisible to the teams responsible for closing it.

Equally important is the rigour with which benchmarks are constructed and operationalised. A benchmark is only as useful as the methodology behind it. The five-stage workflow outlined in this paper, covering benchmark scoping, contributor sourcing and qualification, speech design, speech recording, and speech transcription, is designed to produce benchmarking datasets that reflect the full complexity of production speech. Once built, the end-to-end evaluation workflow encompassing model inference, text normalisation, WER calculation, and dimension-level diagnostics turns these datasets into reliable measurement systems that surface the specific conditions driving model failures and track improvement over time.

We're proud to have partnered with Hugging Face to expand their Hugging Face Open ASR Leaderboard by providing benchmaxxing repellent speech datasets that are kept private rather than publishing openly. The organisations that build reliable voice AI products are those that invest in benchmarking infrastructure with the same rigour they apply to model development. Appen is uniquely positioned to deliver this infrastructure at scale. By combining over two decades of speech-data expertise, a pre-vetted global workforce covering 500+ languages and 100+ markets, and multilingual capabilities, Appen delivers production-representative speech benchmarks that ASR and speech foundation-model teams need to drive systematic, measurable performance improvements.

Sources

Ashungafac, G. Z., Sanni, M., Awobade, B., Gichamba, A., & Olatunji, T. (2025). AfriSpeech-MultiBench: A Verticalized Multidomain Multicountry Benchmark Suite for African Accented English ASR. https://arxiv.org/abs/2511.14255

Guha, S., Ambavat, R., Gupta, A., Gupta, M., & Mehta, R. (2023). Unsupervised Language Agnostic WER Standardization. https://arxiv.org/abs/2303.05046

He, J., Sawada, N., Miyazaki, K., & Toda, T. (2025). CMT-LLM: Contextual Multi-Talker ASR Utilizing Large Language Models. https://arxiv.org/abs/2506.12059

Huang, S., Du, Y., Yang, J., Zhang, D., Jia, X., Deng, J., Kang, J., & Zheng, R. (2025). Overlap-Adaptive Hybrid Speaker Diarization and ASR-Aware Observation Addition for MISP 2025 Challenge. https://arxiv.org/abs/2505.22013

Lau, M., Chen, Q., Fang, Y., Xu, T., Chen, T., & Golik, P. (2025). Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning. https://arxiv.org/abs/2506.17525

Linke, J., Winkler, J., & Schuppler, B. (2025). Context is all you need? Low-resource conversational ASR profits from context, coming from the same or from the other speaker. https://www.isca-archive.org/interspeech_2025/linke25_interspeech.pdf

Parcollet, T., Tseng, Y., Zhang, S., & van Dalen, R. (2025). The Loquacious Set: 25,000 Hours of Transcribed and Diverse English Speech Recognition Data for Research and Commercial Use. https://arxiv.org/abs/2505.21578

Rai, A., Rahangdale, S., Anand, U., & Mukherjee, A. (2025). ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems. https://arxiv.org/abs/2505.11572

Srivastav, V., Zheng, S., Bezzam, E., Le Bihan, E., Koluguri, N., Żelasko, P., Majumdar, S., Moumen, A., & Gandhi, S. (2025). Open ASR Leaderboard: Towards Reproducible and Transparent Multilingual and Long-Form Speech Recognition Evaluation. https://arxiv.org/abs/2510.06961

Xiao, C., Liang, R., Zhang, X., Tiryaki, M. E., Bae, V., Shankar, L., Yang, R., Poon, E., Dupoux, E., Khudanpur, S., & Garcia Perera, L. P. (2025). CASPER: A Large Scale Spontaneous Speech Dataset. https://arxiv.org/abs/2506.00267