Amazon Transcribe - Batch (English-US)

AWS AI Service Cards

AWS AI Service Cards: Amazon Transcribe - Batch (English-US)

Table of Contents

Amazon Transcribe - Batch (English-US)

An AWS AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the responsible design and use of the service. A Service Card will evolve as AWS receives customer feedback, and as the service progresses through its lifecycle. AWS recommends that customers assess the performance of any AI service on their own content for each use case they need to solve. For more information, please see AWS Responsible Use of AI Guide and the references at the end. Please also be sure to review the AWS Responsible AI Policy, AWS Acceptable Use Policy, and AWS Service Terms for the services you plan to use.

This AI Service Card applies to the version of Amazon Transcribe – Batch (English-US) that is current as of November 22, 2023.

Overview

Amazon Transcribe enables AWS customers to add speech-to-text capabilities in their voice-enabled applications. Using Automatic Speech Recognition (ASR) technology, customers can use Amazon Transcribe for a variety of business applications. Features provided by the service include automatic speech recognition, speaker diarization, personally identifiable information (PII) redaction, and language identification; see the Amazon Transcribe Developer Guide for additional detail. This AI Service Card describes one of these features, Transcribe – Batch (English-US), implemented by the Transcribe::StartTranscriptionJob API. This feature performs ASR in the en-US locale at low (8kHz) or high (16kHz) bandwidth. It operates on recorded speech that is available from a static audio file (batch mode). For ASR in near real-time on streaming media, see the Transcribe::StartStreamTranscription API.

We assess the quality of Transcribe Speech by measuring how well the words from an ASR transcript match the words spoken in the speech sample, as transcribed by a human listener. When a speaker says “This system can really recognize speech,” we expect the transcript to contain the spoken words, not “This system can wreck a nice beach.” Three types of errors may appear in a transcription: substitutions (like recognize for wreck), insertions (extra words such as "nice"), and deletions (missing words such as "really"). Correctly-transcribed words are called hits. Quality metrics like precision, recall, F1, and word error rate (WER) depend on the number of hits and errors.

Multiple factors affect the accuracy of any ASR system. The input audio signal consists of the speech itself, modified by a variety of confounding factors. Individual words and utterances differ from speaker to speaker in the frequency with which they are used, in how they are pronounced, and in the ways they are combined with other words. Words that differ in spelling and meaning may not differ in sound. Speakers may overlap or interrupt one another. Recording devices differ in quality, and position relative to the speaker (e.g., farfield vs nearfield). Recording environments differ in the level of background noise, susceptibility to echo, and the presence of other speakers. Transmission lines vary in the level of noise. Transcribe is designed to distinguish between the audio for different words, and ignore the confounding variations.

Intended use cases and limitations

Transcribe – Batch (English-US) is intended for use on audio samples that contain naturally occurring human speech. It is not designed for mechanically or digitally transformed speech, or synthetic speech. It is intended to transcribe US English words; see Supported Languages for additional language locales. Transcribe supports a large general-purpose vocabulary; customers can add custom vocabularies and custom language models for coverage of words and phrases from specialized domains. Transcribe supports speaker partitioning, also known as diarization. Up to 10 unique speakers can be identified by enabling speaker partitioning in the API call.

Transcribe Batch (English-US) has many possible applications, such as contact center analytics (sentiment/categorization/talk speed), transcribing voicemails, meeting captioning, captioning for media content (audio or video), and search/analytics/keyword analysis for media, including cataloging or indexing media archives. These applications vary in their design by 1/ the number of speakers, 2/ the number of speakers per channel (i.e., per recording device such as a laptop or mobile phone), 3/ the style of speech employed by speakers, 4/ recording conditions (such as location and equipment), and other factors. For example, a contact center transcription application might expect two speakers; one speaker per channel; near-field recording (with the speaker's mouth close to the microphone); and high background noise from both the caller's home environment and the contact center operator's work environment. A second example is an application for doing closed-captioning of instructional videos, an entry point to media analytics, indexing, and search. This application would expect multiple speakers; one audio channel shared across all speakers; scripted speech with fewer filler words, pauses, and disfluencies, but more domain-specific jargon; and lower levels of background noise and other audio occlusions.

Design of Transcribe - Batch (English-US)

Machine learning

Transcribe is built using ML and ASR technology. It works as follows: (1) Identify relevant acoustic features of the audio input. (2) Generate a set of candidate word-level strings, based on these features. (3) Apply language modeling to rank the candidates and return the top-ranked transcription. See Amazon Transcribe Developer Guide for details of the API calls.

Performance expectations

Individual and confounding variation will differ between customer applications. This means that performance will also differ between applications, even if they support the same use case. Consider two transcription applications A and B. Application A enables video captioning for a TV talk show, and has multiple voices per recording channel, high-quality boom microphones, and negligible background noise. Application B helps contact centers record customer calls, and has customers speaking close to their mic, one voice per recording channel, and unscripted customer dialog. Because A and B have differing kinds of inputs, they will likely have differing error rates, even assuming that each application is deployed perfectly using Transcribe.

Test-driven methodology

We use multiple datasets to evaluate performance. No single evaluation dataset provides an absolute picture of performance. That’s because evaluation datasets vary based on their demographic makeup (the number and type of defined groups), the amount of confounding variation (quality of content, fit for purpose), the types and quality of labels available, and other factors. We measure Transcribe performance by testing it on evaluation datasets containing audio recordings from a variety of speakers who are representative of the population of end users, where each recording is labeled with ground-truth transcriptions and demographic attributes of the speaker. We represent overall performance on a dataset by several metrics, included word error rate and F1, a percentage that evenly balances the percentage of predicted words that are correct (precision) against the percentage of correct words that are included in the prediction (recall). Groups in a dataset can be defined by demographic attributes (such as gender, age and ancestry), confounding variables (such as recording equipment varieties, the distance of each speaker from recording equipment, post-processing and background noises), or a mix of the two. Different evaluation datasets vary across these and other factors. Because of this, all metrics – both overall and for groups – vary from dataset to dataset. Taking this variation into account, our development process examines Transcribe performance using multiple evaluation datasets, takes steps to increase accuracy for groups on which Transcribe performed least well, works to improve the suite of evaluation datasets, and then iterates.

Fairness and bias

Our goal is for Transcribe – Batch (English-US) to work well for speakers of US English across the variety of pronunciations, intonations, vocabularies, and grammatical features that these speakers may use. We consider speaker communities defined by regions, such as the Midwest or New York City, and communities defined by multiple dimensions of identity, including ancestry, age, and gender. To achieve this, we use the iterative development process described above. As part of this process, we build datasets to capture a diverse range of human speakers under a wide range of confounding factors. We routinely test on datasets for which we have reliable demographic labels. We find that Transcribe performs well across demographic attributes. As an example, on one dataset of natural speech with 65 demographic groups, defined by age, ancestry, gender and regional dialect (such as Female+European, Male+Under 45), we find that F1 word recognition accuracy is 92% or higher for every group of speakers. For transcriptions with speaker partitioning (diarization) enabled, on the same dataset we find that diarization accuracy is 98% or higher for every group of speakers. Because results depend on Transcribe, the customer workflow and the evaluation dataset, we recommend that customers additionally test Transcribe on their own content.

Explainability

When Amazon Transcribe transcribes audio, it creates different versions of the same transcript and assigns a confidence score to each version. If customers enable alternative transcriptions, Amazon Transcribe returns alternative versions of the transcript that have lower confidence levels. Customers can explore alternative transcriptions to gain more insight into candidate words and phrases that were generated for each audio input.

Robustness

We maximize robustness with a number of techniques, including using large training datasets that capture many kinds of variation across many individuals. Ideal audio inputs to Transcribe ASR contain audio with high recording quality, low background noise, and low room reverberation. However, Transcribe is trained to be resilient even when inputs vary from ideal conditions and can perform well in noisy and multi-speaker settings.

Privacy and security

Amazon Transcribe processes only audio input data. Audio inputs are never included in the output returned by the service. Inputs and outputs are never shared between customers.

Customers can opt out of training on customer content via AWS Organizations or other opt out mechanisms we may provide. For more information, see Section 50.3 of the AWS Service Terms and the AWS Data Privacy FAQs. For service-specific privacy and security information, see Amazon Transcribe FAQs and Amazon Transcribe Security.

Transparency

Where appropriate for their use case, customers who incorporate Amazon Transcribe in their workflow are encouraged to disclose their use of ML and ASR technology to end users and other individuals impacted by the application, and give their end users the ability to provide feedback to improve workflows. In their documentation, customers can also reference this AI Service Card.

Governance

We have rigorous methodologies to build our AWS AI services in a responsible way, including a working backwards product development process that incorporates Responsible AI at the design phase, design consultations and implementation assessments by dedicated Responsible AI science and data experts, routine testing, reviews with customers, and best practice development, dissemination, and training.

Deployment and performance optimization best practices

We encourage customers to build and operate their applications responsibly, as described in the AWS Responsible Use of AI Guide. This includes implementing Responsible AI practices to address key dimensions including fairness and bias, robustness, explainability, privacy and security, transparency, and governance.

Workflow design: The performance of any application using Transcribe depends on the design of the customer workflow. Conditions like background noise, recording equipment, and others are discussed in the Intended Use Cases section. Depending on the application, these conditions may be optimized by Transcribe customers, who define the workflow where audio is captured from end users. Transcribe provides features for customers to optimize their recognition performance within the API. These features include recording conditions, sample rates, custom vocabularies, custom language models, and filtering for vocabulary or personally identifying information (PII). Human oversight, workflow consistency and periodic testing for performance drift are also critical considerations that are under the control of customers, and that contribute to accurate, fair outcomes.

1. **Recording conditions:** Workflows should include steps to address variation in recording conditions, such as speaking far from the microphone or in noisy conditions. If variation is high, consider providing help and instructions that are accessible to all end users, and monitor recording quality by periodically and randomly sampling inputs.
2. **Sample rates:** Customers have an optional parameter to specify the sample rate of their input audio, either lower bandwidth (8kHZ) or broadband (16kHz) inputs.
3. **Custom vocabularies:** Transcribe recognizes vocabulary used in a variety of speaker communities (dialect regions, demographic groups). In cases where customers want to provide additional support for words specific to their domain or situation such as brand names or proper nouns and acronyms, customers can deploy custom vocabularies to improve transcription accuracy for such words. For more information, see the documentation for Custom Vocabularies.
4. **Custom language models:** When a customer application must handle domain-specific speech that is more complex than just single words, customers can use custom language models to improve transcription accuracy. For example, when transcribing recordings of climate science talks, it may be possible to increase transcription accuracy by learning the context in which words appear (such as “ice flow” vs “ice floe”). In this case, customers can train a custom language model to recognize specialized terms. For more information, see the documentation for Custom Language Models.
5. **Vocabulary filtering and PII redaction:** These optimizations can improve the security and privacy of the language produced in transcriptions. Vocabulary Filtering enables customers to mask or remove words that are sensitive or unsuitable for their audience from transcription results, based on a customer-defined list. PII Redaction enables customers to generate a transcript where PII has been removed, based on PII types that Transcribe – Batch (English-US) identifies. These include name, address, credit card number, SSN, and others. For more information, including a complete list of PII types and considerations on using PII redaction for regulated workloads, see the documentation for Vocabulary Filtering and for PII Redaction.
6. **Human oversight**: If a customer's application workflow involves a high risk or sensitive use case, such as a decision that impacts an individual's rights or access to essential services, human review should be incorporated into the application workflow where appropriate. ASR systems can serve as tools to reduce the effort incurred by fully manual solutions, and to allow humans to expeditiously review and assess audio content.
7. **Consistency:** Customers should set and enforce policies for the kinds of workflow customization and audio inputs permitted, and for how humans use their own judgment to assess Transcribe outputs. These policies should be consistent across demographic groups. Inconsistently modifying audio inputs could result in unfair outcomes for different demographic groups.

8. **Performance drift:** A change in the kinds of audio that a customer submits to Transcribe, or a change to the service, may lead to different outputs. To address these changes, customers should consider periodically retesting the performance of Transcribe, and adjusting their workflow if necessary.

Further information

- For service documentation, see Amazon Transcribe.
- For a list of supported languages, see Supported Languages.
- For an example of a Contact Center Analytics workflow design, see Amazon Transcribe Call Analytics.
- For details on privacy and other legal considerations, see the following AWS policies: Acceptable Use, Responsible AI, Legal, Compliance, and Privacy.
- For help optimizing workflows, see Generative AI Innovation Center, AWS Customer Support, AWS Professional Services, Ground Truth Plus, and Amazon Augmented AI.
- If you have any questions or feedback about AWS AI service cards, please complete this form.

Glossary

Controllability: Steering and monitoring AI system behavior.

Privacy & Security: Appropriately obtaining, using and protecting data and models.

Safety: Preventing harmful system output and misuse.

Fairness: Considering impacts on different groups of stakeholders.

Explainability: Understanding and evaluating system outputs.

Veracity & Robustness: Achieving correct system outputs, even with unexpected or adversarial inputs.

Transparency: Enabling stakeholders to make informed choices about their engagement with an AI system.

Governance: Incorporating best practices into the AI supply chain, including providers and deployers.