Evaluating Cloud Speech-to-Text Services

Luke Connolly
School of Enterprise Computing and Digital Transformation, TU Dublin, Ireland
X00218713@myTUDublin.ie

Introduction

Cloud speech-to-text (STT) services promise accuracy, scalability and low-cost transcriptions, yet vendor accuracy is not widely reported and many available benchmarks use clean, lab-like audio for these accuracy figures. A key question is how these systems perform under real-world conditions, where speech is often degraded by background noise and interference. This thesis presents a comparative evaluation of Amazon Transcribe, Microsoft Azure Speech and Google Cloud Speech-to-Text, focusing on accuracy, latency and cost to provide a trade-off analysis for organisations. Understanding these trade-offs allows organisations seeking to embed STT into production workflows to better understand each vendor's strengths.

Research Question

Which cloud speech-to-text service offers the best trade-off between accuracy, latency and cost when transcribing Irish English speech across varying acoustic conditions?

Methodology

- **Create a reproducible, DevOps-ready evaluation framework:** Implement an automated benchmarking pipeline using vendors' official SDKs, built as modular Ruby scripts and designed for CI/CD integration.
- **Experiment design:** Start with clean Irish English speech audio and apply pink noise to generate controlled acoustic conditions (Clean, 20 dB, 10 dB, 0 dB SNR).
- **Evaluate leading vendors:** AWS Transcribe, Microsoft Azure Speech and Google Cloud Speech-to-Text.
- **Benchmark metrics:**
  - **Accuracy:** Word Error Rate (WER), error-type breakdown (substitutions/deletions/insertions), and misrecognition analysis.
  - **Latency:** Time from job submission to completed transcription.
  - **Cost:** Cost per hour and scaled production workload projections.

Key Findings

Accuracy is consistent across all conditions: **AWS** (1st), **Azure** (2nd), **GCP** (3rd).

Accuracy (Word Error Rate)

- **WER (Clean):** AWS 4.34%, Azure 10.27%, GCP 17.58%.
- **WER (0 dB):** AWS 8.22%, Azure 25.11%, GCP 32.88%.
- **Overall mean WER (all conditions):** AWS 5.71%, Azure 15.13%, GCP 22.49%.

Error fingerprints (What causes the error increases)

- Degradation in quality of transcriptions is **substitution-led** for all vendors.
- **Azure:** Larger **deletion** rise at heavy noise.
- **GCP:** More **insertions** at mid-noise (10 dB), suggesting occasional short-word hallucinations.

Latency

- Total latency: AWS 17.35 s, GCP 41.39 s, Azure 51.24 s → AWS is ∼2.4–3× faster.

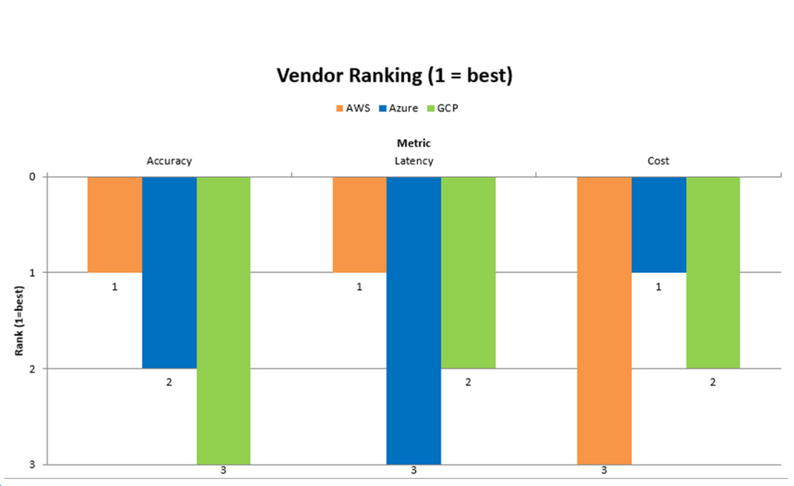
Cost

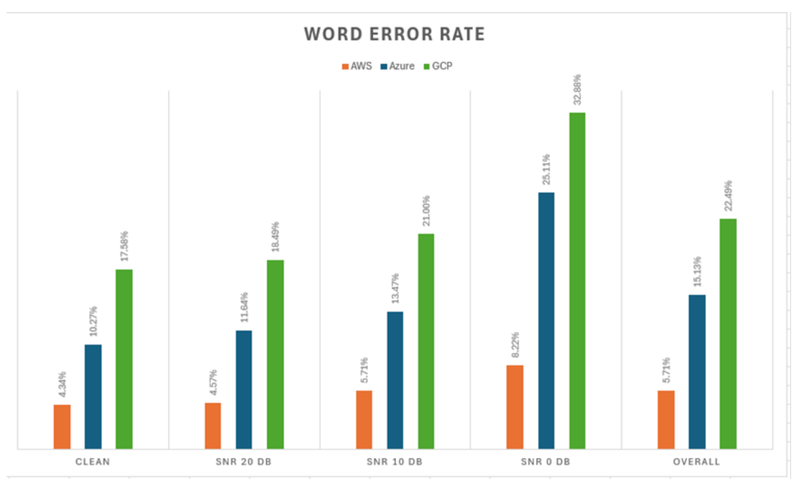
- AWS €1.24/hr, GCP €0.83/hr, Azure €0.16/hr → Azure is the cheapest cost option.

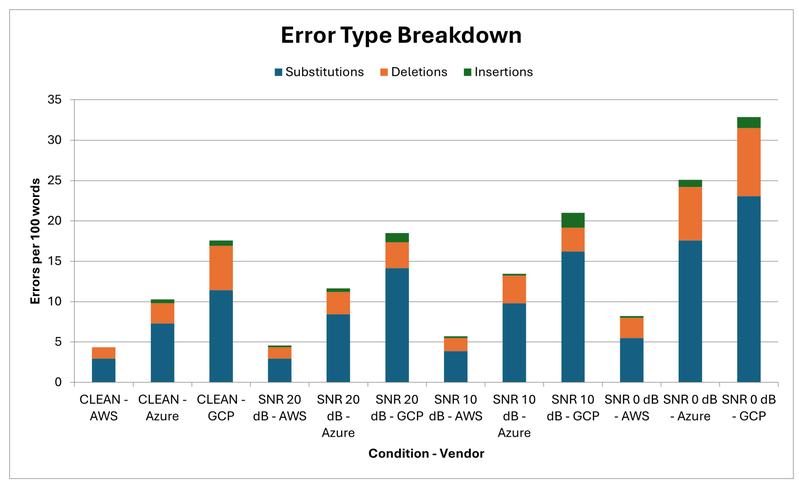
Trade-off Analysis

- Choose **AWS** when **quality** or **turnaround** matters most.
- Choose **Azure** when **budget at scale** dominates and higher WER/slower turnaround is acceptable.
- Choose **GCP** only when accuracy needs are looser or when there is data residency/on-premise constraints.

Topic Overview







Conclusions and Future Work

Results show that **AWS** delivers the strongest accuracy and the fastest turnaround, while **Azure** offers the lowest cost but with higher error rates and longer latency. **Google** trails on all benchmarks. The key takeaway is that organisations should evaluate vendors on their own audio and conditions, then select the service whose accuracy–latency–cost best matches the target use case.

Future Work: Will extend the evaluation to real-time streaming transcription, broader speaker styles to also include multi-speaker overlap and to use the pipeline for continuous regression testing in CI/CD to track performance changes across models, configuration and region updates. Additional experiments will also assess the value of custom vocabularies and review human editing effort required to reach usable transcripts for each vendor.

QR Code for Recording