Análise Comparativa de Eficiência Computacional e Word Error Rate (WER) em Ferramentas de Inteligência de Vídeo

Rafael Herculano[1], **Roberta Fagundes**[1]

[1]Universidade de Pernambuco (UPE) – Recife – PE – Brazil

{rhesa@ecomp.poli.br, roberta.fagundes@upe.br}

Resumo. *O crescimento exponencial do YouTube (29 bilhões de vídeos em 2025) exige escalabilidade na análise de mídia digital. Este artigo apresenta uma análise comparativa sistemática de quatro ferramentas: Google Cloud, AWS, TwelveLabs e VOSK. A metodologia avalia o desempenho computacional (throughput, latência, memória) e a qualidade analítica (WER) utilizando um dataset de referência do Kaggle. Os resultados confirmam um trade-off entre eficiência e precisão: a TwelveLabs obteve a melhor performance global (7,15% WER), enquanto o VOSK registrou a menor latência (24,44 s). O estudo fornece diretrizes técnicas para a seleção de ferramentas alinhada aos requisitos de acurácia e disponibilidade de infraestrutura.*

1. Introdução

Em 2025, o Google revelou que cerca de 65 mil vídeos são enviados todos os dias para o YouTube. Isso significa mais de 500 horas enviadas por minuto [Pandeló 2026]. A empresa fechou o ano de 2025 com 29 bilhões de vídeos enviados [Gomez 2025]. Nesse contexto, empresas que realizam pesquisas de mercado a partir de vídeos e comentários enfrentam desafios relacionados à escalabilidade e à eficiência na análise dessas informações.

Diante disso, torna-se fundamental o uso de soluções automatizadas capazes de acelerar o processamento e garantir a qualidade das análises. No entanto, ainda há carência de estudos comparativos que avaliem sistematicamente o desempenho de cada tecnologia. Assim, este trabalho tem como objetivo comparar diferentes artefatos de análise de vídeos no uso prático, considerando critérios como desempenho, eficiência computacional e qualidade dos resultados obtidos.

2. Seleção das ferramentas e critérios

A pesquisa por analisadores de vídeo trouxe ferramentas de empresas de grande renome e ferramentas de código aberto. Por conta das limitações de acesso para testes e limitações de créditos de uso, algumas precisaram ser descartadas. Foram selecionadas 4 opções que representam os mais usados, de grandes empresas, empresas menores, de código fechado e aberto, a fim de contemplar diferentes perfis de soluções disponíveis no mercado.

O Google Cloud Video Intelligence é a primeira ferramenta. Ela tem modelos pré-treinados de machine learning que reconhecem automaticamente os elementos presentes nos vídeos armazenados e em streaming [LLC 2026]. Amazon Rekognition é a segunda das soluções. Além da detecção, o Amazon Rekognition consegue indexar elementos encontrados com timestamp para facilitar uma futura busca por conteúdo

[Services 2026]. A TwelveLabs também foi selecionada para a comparação. Esta trata o vídeo como conteúdo por áudio, visual, texto na tela, contexto temporal simultaneamente para alcançar um entendimento do conteúdo [Daoud 2025]. VOSK é a 4ª e última candidata. Trata-se de uma toolkit de reconhecimento de fala que reconhece diversos idiomas, funciona offline até em dispositivos de baixa performance [Soni 2025].

3. Métrica de Performance de Sistema

Para a realização da comparação, foram selecionadas métricas que avaliam tanto o desempenho computacional quanto a qualidade dos resultados produzidos. Essas métricas subsidiam a tomada de decisão quanto à escolha da plataforma de análise de vídeo mais adequada ao contexto de aplicação. Cada uma das ferramentas transcreveu o que foi falado no vídeo e o desempenho de cada modelo foi medido considerando o processamento dessa transcrição e o quão correto ela está. As métricas desse estudo ajudam na tomada de decisão de qual a melhor solução para o contexto em que o leitor está inserido.

O *throughput* foi adotado como medida de capacidade de processamento, sendo definido como a quantidade de dados processados em um determinado intervalo de tempo. Esse é um fator importante para prever o tempo de espera pelos resultados [Duarte 2019]. A latência *end-to-end* foi utilizada para mensurar o tempo total de processamento, compreendendo todas as etapas do fluxo, da coleta do vídeo à entrega do resultado final, o que evidencia diretamente a velocidade da ferramenta [TOTVS 2023].

Adicionalmente, foi considerado o consumo de memória, que representa a quantidade de memória volátil utilizada durante a execução, impactando diretamente a eficiência e a escalabilidade das soluções. Monitorar essa métrica é essencial para avaliar a viabilidade de execução em diferentes ambientes [Charleaux and Toledo 2025].

Para avaliar a qualidade das transcrições geradas, empregou-se uma abordagem baseada em Processamento de Linguagem Natural (PLN), área da Inteligência Artificial voltada à análise e interpretação da linguagem humana [Software 2023]. A métrica utilizada foi o *Word Error Rate* (WER), amplamente reconhecido como padrão na avaliação de sistemas de reconhecimento automático de fala. O WER quantifica a taxa de erro de uma transcrição ao compará-la com uma referência. Seu cálculo é dado pela razão entre o número total de erros e o número de palavras na transcrição de referência [Ali and Renals 2018]. O cálculo é feito pela soma dos erros dividida pelo número total de palavras presentes na transcrição de referência. Essa métrica é essencial para medir o quão corretas estão as transcrições dos modelos de análise [Park et al. 2008].

O dataset de referência constitui o padrão contra o qual as transcrições geradas pelas ferramentas são avaliadas. Foi necessário um dataset que tivesse a transcrição de diversos vídeos disponíveis no YouTube para facilitar o acesso pelas ferramentas. O dataset escolhido foi o YouTube transcripts kaggle.cvs. Esse dataset tem o título, o autor, a transcrição completa e a playlist onde esse vídeo está inserido. A fim de mensurar a capacidade dos analisadores igualmente várias vezes, foram escolhidos vídeos com duração média de 7 minutos e com média de 2491 tokens no idioma inglês.

4. Execução

A base de dados tem vários vídeos transcritos com diversas durações e sobre diversos assuntos. Foram selecionados vídeos de duração curta ou média devido a limitação de

caracteres das APIs (máximo de 131072 caracteres).

Apesar das limitações impostas pelas ferramentas em suas modalidades de experimentação gratuita, o experimento foi feito com diversos vídeos que estão transcritos no *dataset*. O tempo máximo considerado para execução foi de 5 minutos. Após atingir os 300 segundos de execução, foi adicionado um *timeout* para conduzir o experimento de forma mais fluida. Os experimentos foram conduzidos em um notebook com processador Intel Core i7 de 12ª geração, 16 GB de RAM e conexão de 200 Mbps.

5. Resultados

Os dados experimentais consolidam o perfil de desempenho das ferramentas avaliadas, fundamentando-se nas medianas das métricas de eficiência computacional e qualidade analítica. Conforme ilustrado na Figura 1, os resultados sintetizam o comportamento das soluções em termos de latência *end-to-end*, *throughput*, consumo de memória e *Word Error Rate*(WER).

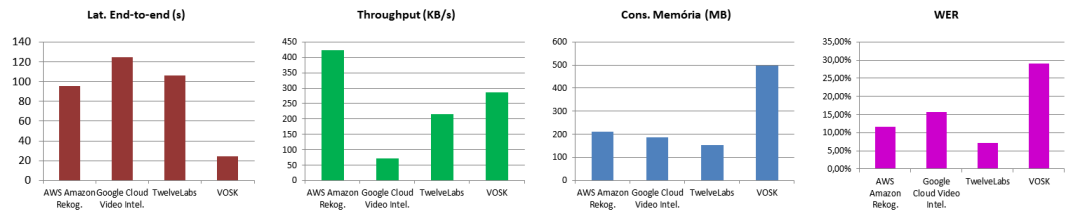

Figura 1. Gráfico da Mediana das execuções por ferramenta

Os resultados experimentais demonstram que as ferramentas TwelveLabs e VOSK apresentaram desempenho superior pelas latências reduzidas e pelos elevados *throughputs*. A agilidade do VOSK — que registrou a menor latência mediana (24,44 s) — é atribuída à sua execução local, que elimina gargalos de comunicação externa e dependência de largura de banda. Contudo, essa rapidez resulta em um alto consumo de memória volátil (497,58 MB) e baixa precisão, evidenciado pelo maior índice de erro do estudo (WER de 29,04%).

Em contraste, a TwelveLabs consolidou-se como a ferramenta de melhor desempenho global, apresentando o menor índice de erro (7,15%) com uma latência competitiva em relação ao VOSK. A superioridade técnica da TwelveLabs sobre o VOSK pode ser explicada pelas disparidades arquiteturais: enquanto o VOSK utiliza modelos compactos baseados na engine Kaldi (uma abordagem mais legada com menor capacidade de generalização), TwelveLabs emprega arquiteturas modernas do tipo Transformer, que integram contextos de áudio, vídeo e texto simultaneamente [Daoud 2025][Soni 2025].

Por outro lado, as soluções em nuvem AWS Rekognition e Google Cloud Video Intelligence exibiram maior latência e menor *throughput*, reflexo do *overhead* inerente ao processamento remoto e às etapas de *upload* e conversão de mídia. Apesar da menor eficiência temporal, essas plataformas garantiram uma qualidade de transcrição significativamente superior ao VOSK, preservando, simultaneamente, os recursos computacionais locais.

Em síntese, os dados evidenciam um *trade-off* crítico entre desempenho e precisão. Ferramentas offline priorizam a velocidade de resposta em detrimento da acurácia, enquanto soluções em nuvem oferecem maior rigor analítico ao custo de um tempo de processamento elevado. A seleção da ferramenta ideal deve, portanto, ser balizada pelos requisitos específicos da aplicação, ponderando a disponibilidade de infraestrutura local frente ao nível de precisão exigido pelo projeto.

6. Conclusão

Com base nos dados experimentais apresentados, conclui-se que a seleção da ferramenta ideal para análise de vídeos do YouTube é determinada por um claro trade-off entre desempenho computacional e precisão analítica. As principais conclusões deste estudo sistemático são:

- **Superioridade da TwelveLabs:** A ferramenta consolidou-se com o melhor desempenho geral, equilibrando uma latência competitiva com o menor índice de erro do estudo (WER de 7,15%). Sua superioridade técnica é atribuída ao uso de arquiteturas modernas do tipo *Transformer*, que permitem um entendimento contextual superior do conteúdo.
- **Agilidade vs. Acurácia no VOSK:** O VOSK destacou-se pela menor latência (24,44 s), favorecido por sua execução offline que elimina gargalos de rede. Contudo, essa rapidez resulta na menor precisão (29,04% de erro) e no maior consumo de memória (497,58 MB), devido ao uso de modelos compactos e da engine Kaldi, que possui menor capacidade de generalização.
- **Estabilidade das Soluções em Nuvem:** AWS Video Rekognition e Google Cloud Video Intelligence apresentaram maior latência e menor throughput em função do processamento remoto e da necessidade de upload de mídia. Entretanto, ambas garantiram níveis de precisão significativamente superiores ao VOSK, com a vantagem adicional de pouparem recursos computacionais locais.

Portanto, este trabalho fornece subsídios técnicos fundamentais para a tomada de decisão estratégica em pesquisas de mercado. A escolha deve ser pautada pelas prioridades do projeto: para aplicações que exigem máxima precisão e contexto, a TwelveLabs é a escolha ideal; para cenários que priorizam a velocidade em tempo real e operam com restrições de conexão, o VOSK é indicado, desde que a baixa acurácia seja tolerável. Para trabalhos futuros, recomenda-se repetir a análise com vídeos de maior duração, tanto nas ferramentas aqui avaliadas quanto em versões futuras destas ou em novos modelos lançados. Os modelos podem ter comportamentos diferentes para textos com mais tokens. Recomenda-se também que os estudos futuros possam avaliar sem a limitação do uso gratuito e com uma variedade maior de idiomas. As limitações do Dataset são inerentes à sua natureza: tem apenas idioma inglês comprometendo a generalização; descarta as informações visuais que seriam importantes para uma possível análise multimodal; está sujeito a erros de ASR, principalmente em terminologias técnicas. Durante a elaboração deste trabalho, os autores utilizaram ferramentas de inteligência artificial generativa como auxílio na revisão linguística e na melhoria da clareza textual. Todo o conteúdo técnico são de responsabilidade exclusiva dos autores, que revisaram e validaram integralmente o texto final.

Referências

Ali, A. and Renals, S. (2018). Word error rate estimation for speech recognition: e-WER. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Charleaux, L. and Toledo, V. (2025). O que é memória ram? veja para que serve, como funciona e quais são os tipos. https://tecnoblog.net/responde/o-que-e-memoria-ram/. Acessado em: 07/04/2026.

Daoud, J. (2025). Twelve labs: Building multimodal video foundation models for better understanding. https://www.youtube.com/watch?v=IMITjAXEqGQ&t=13s. Acessado em: 07/04/2026.

Duarte, A. (2019). Throughput: entenda a importância dessa métrica. https://blog.acelerato.com/artigo/throughput-entenda-o-que-e/. Acessado em: 07/04/2026.

Gomez, V. L. (2025). 20 anos do youtube: quantos vídeos existem na plataforma? descubra esse e outros segredos. https://olhardigital.com.br/2025/02/14/internet-e-redes-sociais/20-anos-do-youtube-quantos-videos-existem-na-plataforma-%descubra-esse-e-outros-segredos/. Acessado em: 07/04/2026.

LLC, G. (2026). Video ai e inteligência. https://cloud.google.com/video-intelligence?hl=pt-BR. Acessado em: 07/04/2026.

Pandeló, N. (2026). Youtube fechou 2025 com 29 bilhões de vídeos; música e shorts são motores do consumo global. https://mundodamusicamm.com.br/youtube-29-bilhoes-videos-2025/#:~:text=O%20YouTube%20se%20aproxima%20dos%2030%20bilh%C3%B5es,pesquisa%20da%20Omdia.%20*%2003/02/2026.%20*%2010:25.. Acessado em: 07/04/2026.

Park, Y., Patwardhan, S., Visweswariah, K., and Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. In *Interspeech 2008*, pages 2070–2073.

Services, A. W. (2026). Trabalhar com operações de análise de vídeo armazenado. https://docs.aws.amazon.com/pt_br/rekognition/latest/dg/video.html. Acessado em: 07/04/2026.

Software, S. (2023). Processamento de linguagem natural: O que é e qual sua importância? https://www.sas.com/pt_br/insights/analytics/processamento-de-linguagem-natural.html. Acessado em: 07/04/2026.

Soni, A. A. (2025). Improving speech recognition accuracy using custom language models with the vosk toolkit. *Cognizant Technology Solutions*.

TOTVS, E. (2023). End to end: o que é, vantagens e como implementar. https://www.totvs.com/blog/gestao-logistica/end-to-end/. Acessado em: 07/04/2026.