PROPOSTA DE UM MÉTODO DE SUMARIZAÇÃO DE VÍDEO DIGITAL EM TEXTO

BRUNO MASCIOLI DE SOUZA[1], TIAGO CATOIA DE SOUZA[2], TIAGO HENRIQUE TROJAHN[3]

[1] Graduando em Bacharelado em Engenharia de Software, Bolsista PIBITI, IFSP, Campus São Carlos, bruno.mascioli@aluno.ifsp.edu.br.
[2] Graduando em Bacharelado em Engenharia de Software, Voluntário PIVICT, IFSP, Campus São Carlos, catoia.t@aluno.ifsp.edu.br.
[3] Doutor em Ciências da Computação e Matemática Computacional, docente, IFSP, Campus São Carlos, tiagotrojahn@ifsp.edu.br"
Área de conhecimento (Tabela CNPq): 1.03.03.00-6 Metodologia e Técnicas da Computação

RESUMO: A produção massiva de vídeos na internet cria um desafio crescente: como encontrar conteúdos relevantes em meio ao excesso de informações. Esse aumento no volume de vídeos agrava o problema da sobrecarga informacional, dificultando a localização de conteúdos de interesse. Este estudo propõe uma técnica de sumarização de vídeos, que extrai os principais tópicos de um vídeo em formato textual. A técnica utiliza um modelo de Reconhecimento Automático de Fala (ASR) para transcrever o conteúdo e, em seguida, aplica modelos baseados na arquitetura Transformer para refinar e resumir as informações, facilitando o acesso a conteúdos relevantes.

PALAVRAS-CHAVE: sumarização de vídeo; segmentação de vídeo; machine learning; aprendizagem profunda;

PROPOSAL OF A METHOD FOR DIGITAL VIDEO SUMMARIZATION INTO TEXT

ABSTRACT: The massive production of videos on the internet creates a growing challenge: how to find relevant content amidst the information overload. This increase in video volume exacerbates the problem of information overload, making it difficult to locate content of interest. This study proposes a video summarization technique that extracts the main topics of a video in textual format. The technique uses an Automatic Speech Recognition (ASR) model to transcribe the content and then applies Transformer-based models to refine and summarize the information, making it easier to access relevant content.

KEYWORDS: video summarization, video segmentation, machine learning, deep learning

INTRODUÇÃO

A sobrecarga de informação, popularizada por Toffler em Future Shock (1970), destaca o impacto negativo do excesso de dados disponíveis, gerando dificuldades em encontrar conteúdo relevante e causando desconforto mental. Com a popularização de dispositivos e serviços de gravação e distribuição de vídeo digital online, os usuários enfrentam uma crescente dificuldade em localizar vídeos de interesse devido ao grande volume disponível. Entre as soluções para amenizar esse problema, como a indexação e recomendação de conteúdo (Manzato, 2011), destaca-se a sumarização de vídeos. Essa técnica visa gerar um resumo condensado do conteúdo, facilitando a identificação de vídeos relevantes para o usuário e auxiliando na filtragem automática de vídeos indesejados.

A sumarização automática de vídeos possui aplicações relevantes, como no ambiente educacional, onde facilita a revisão de conteúdos essenciais de aulas online, e no contexto corporativo, onde pode sintetizar reuniões extensas, destacando os principais temas discutidos.

Recentemente, avanços em aprendizagem profunda, especialmente com a arquitetura Transformer (Vaswani et al., 2017), têm proporcionado melhorias significativas na modelagem de dependências complexas entre entradas e saídas, demonstrando eficácia em tarefas de reconhecimento automático de fala (ASR) e geração de texto. O trabalho de (Radford et al., 2023) exemplifica esse progresso com a criação de modelos altamente eficazes, como o Whisper, que consegue transcrever áudio com alta precisão em diversas línguas.

Assim, este projeto propõe o desenvolvimento de uma técnica de sumarização de vídeos digitais, utilizando e adaptando técnicas de extração de informações multimídia, como a conversão de fala em texto. A geração do resumo será realizada por meio de redes neurais generativas de última geração, garantindo uma síntese precisa e eficiente do conteúdo original.

MATERIAL E MÉTODOS

Inicialmente, foi realizado um levantamento bibliográfico de estudos e pesquisas realizadas no âmbito da sumarização de vídeos, com o objetivo de introduzir, compreender e explorar técnicas já presentes na literatura. Com base nesse embasamento teórico, este projeto propõe uma técnica de sumarização de vídeos automática, baseada em transcrição. O método proposto consiste em um *pipeline* de duas etapas, conforme ilustrado na Figura 1. Na primeira etapa, o áudio do vídeo é transcrito utilizando um modelo de reconhecimento automático de fala (ASR), convertendo o discurso em texto. A saída gerada é então processada por uma rede neural generativa de transformadores, que realiza a sumarização do conteúdo conforme critérios e parâmetros pré-definidos.

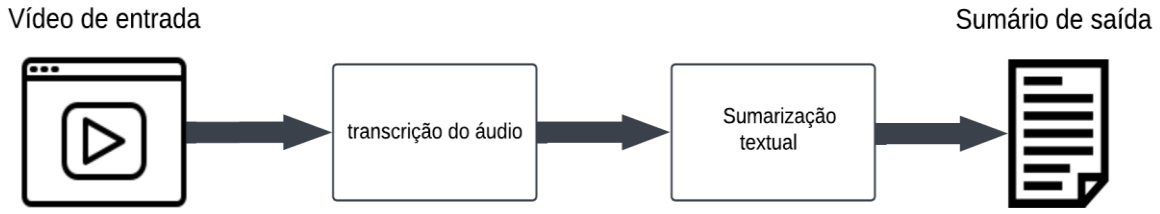


FIGURA 1. Ilustração do pipeline da técnica proposta

Para implementar essa técnica, foi conduzida uma pesquisa sobre métodos de extração de texto a partir de vídeos, com o objetivo de identificar as ferramentas mais eficazes e adequadas para o projeto. Após a avaliação das opções disponíveis, o modelo Whisper AI (Radford et al., 2023) foi selecionado. Esse modelo de código aberto foi treinado em 680.000 horas de dados supervisionados multilíngues e multitarefa coletados da web, demonstrando alta eficácia para a transcrição de áudio em diversas línguas. O Whisper é composto por cinco modelos de tamanhos principais: *tiny, base, small, medium e large*, que variam em número de camadas, comprimento, cabeças e parâmetros. Cada versão disponível foi treinada de forma independente com suas próprias características e capacidade de generalização.

Com a transcrição gerada a partir do Whisper, foi possível alimentar modelos generativos, como o GPT-3.5 e o Gemini-1.5-Pro para que uma síntese dos principais tópicos do texto fosse gerada. A escolha desses modelos se deve à sua alta eficiência e ao fato de serem opções robustas disponíveis para uso gratuito. No entanto, essas opções possuem limitações, como a necessidade de uma chave de acesso para utilizar a aplicação e restrições no número de utilizações na versão gratuita.

Foram definidos parâmetros como temperatura, top_p, top_k e max_tokens para garantir que os resumos gerados sejam informativos, diretos e livres de desvios do tema principal. A temperatura foi configurada para um valor baixo de 0,3, o que ajuda a manter a geração do texto mais focada e determinística, reduzindo a aleatoriedade e evitando informações irrelevantes. O parâmetro top_k foi ajustado para 50, limitando as opções de palavras para as 50 mais prováveis e assegurando que o modelo mantenha a concisão e relevância do resumo. O top_p foi definido em 0,9 para permitir uma escolha flexível dentro das palavras de alta probabilidade, equilibrando diversidade com precisão. Já o max_tokens foi definido para garantir que os resumos sejam suficientemente longos para cobrir os

principais pontos sem se tornarem excessivamente extensos.

Além disso, foi definido um *prompt* estruturado para guiar o modelo na extração dos pontos mais críticos dos vídeos transcritos. Nele, foram incluídas instruções para que o modelo seja capaz de identificar e sumarizar os principais tópicos de cada seção do vídeo de entrada, assegurando que a síntese final represente fielmente o conteúdo essencial sem perda de informações importantes.

A partir do método proposto, uma aplicação desktop foi desenvolvida, com o objetivo de facilitar a utilização e a personalização do *pipeline* proposto. A aplicação foi desenvolvida em dois repositórios separados, um para o *backend* e outro para o *frontend*. Ambos os repositórios são de código aberto e estão disponíveis para consulta e contribuição. A escolha por dividir o projeto em dois repositórios distintos reflete a necessidade de modularidade e organização, facilitando a manutenção e atualização de cada componente individualmente.

O *backend* da aplicação foi desenvolvido inteiramente em Python, e o serviço foi disponibilizado através de uma interface de programação de aplicações (Application Programming Interface - API) construída com o framework FastAPI. O código foi elaborado seguindo a arquitetura de Orientação a Objetos, tornando mais fácil a manutenção e a inserção de novos modelos a partir da criação de novas classes, que poderão ser adicionadas em trabalhos ou contribuições futuras.

Para o *frontend*, foi desenvolvida uma interface ilustrada na Figura 2, utilizando tecnologias como Electron, React, JavaScript, HTML e CSS. A interface permite diversas opções de configuração, como a entrada do vídeo, que pode ser feita por meio de uma URL (Uniform Resource Locator) de um vídeo disponível no YouTube ou pelo upload de um arquivo de vídeo local. Os usuários podem escolher o tamanho do modelo Whisper a ser utilizado para a transcrição. Além disso, a interface oferece a escolha entre diferentes modelos de linguagem natural (LLM - Large Language Models). para a sumarização ou classificação do vídeo, que incluem o GPT-3.5 Turbo e o Gemini-1.5-Pro.

Cada tarefa conta com um *prompt* específico, permitindo que a aplicação seja configurada para gerar resumos ou realizar classificações conforme a necessidade do usuário. O processo de classificação de vídeo, embora integrado nesta aplicação, foi elaborado em um projeto paralelo focado na criação de um sistema de categorização de vídeos. Essa funcionalidade adicional permite a classificação do vídeo em uma lista de categorias como, por exemplo Notícia, Esporte, Entretenimento, Educação, Tecnologia, entre outras. Com base na transcrição do vídeo o modelo generativo identifica os principais aspectos do texto, facilitando a categorização e organização dos vídeos de acordo com critérios pré-definidos, o que enriquece a aplicação e proporciona uma maneira eficiente de localizar e analisar conteúdos relevantes para os usuários.

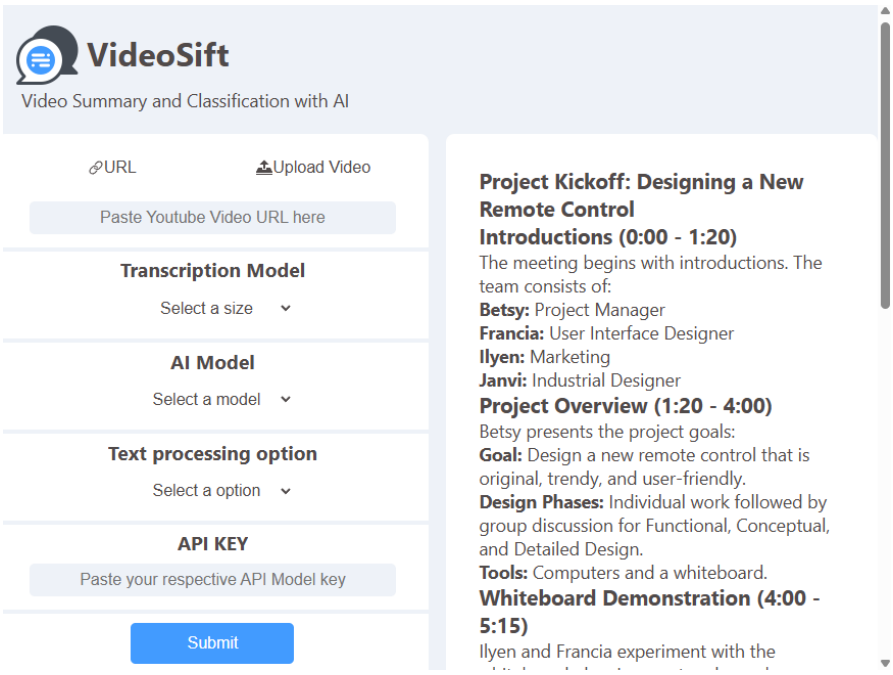


FIGURA 2. Interface da aplicação *Desktop*

RESULTADOS E DISCUSSÃO

Os resultados obtidos demonstram a eficácia da técnica proposta, com a transcrição de áudio utilizando o modelo Whisper. A transcrição resultante é, em geral, fiel ao áudio original, especialmente em vídeos com boa qualidade sonora e linguagem clara. Embora o modelo suporte diversas línguas, incluindo português e inglês, o *prompt* (Figura 3) foi definido para que o processo de sumarização ocorra na mesma língua em que o vídeo foi transcrito, mantendo assim a coerência linguística com o material original. A técnica foi projetada para assegurar que o resumo capture os principais tópicos de cada seção do vídeo, mantendo a clareza e a concisão necessárias. Ela orienta a criação de resumos que sejam fáceis de ler e compreender, com uma estrutura que inclui títulos e subtítulos para organizar a informação. A utilização de linguagem simples, sem jargões técnicos, e a manutenção do idioma original do texto garantem que o conteúdo essencial seja representado de forma fiel e acessível.

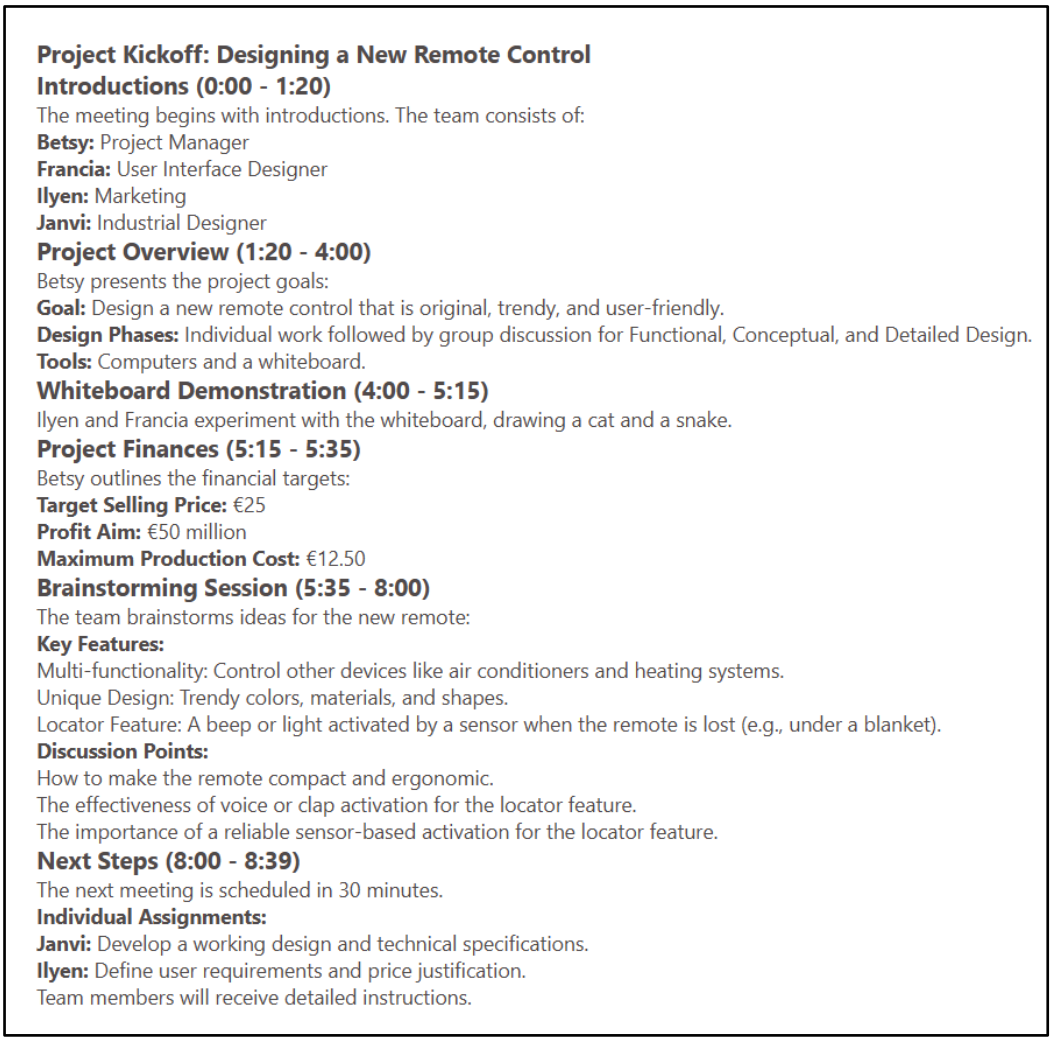
```
Summarize the following text, making it easy to read and comprehend.
The summary should be concise, clear, and capture the main points of the text.
Add clear headings and subheadings to guide the reader through each section.
Avoid using complex sentence structures or technical jargon.
Ensure that the summary is written in the same language as the original text.
Please begin by editing the following text:
```

FIGURA 3. *Prompt* criado para a tarefa de sumarização

A aplicação realiza a diarização dos áudios dos vídeos processados usando o toolkit de código aberto pyannote.audio (Bredin et al., 2020). O modelo pyannote/speaker-diarization-3.1 divide o áudio em segmentos por locutor, permitindo identificar quem está falando em cada parte do vídeo. Isso é útil para análises detalhadas em cenários com vários participantes, como reuniões ou entrevistas.

O processo de diarização inclui a aplicação de um modelo de detecção de atividade vocal (VAD) para remover partes silenciosas e identificar regiões com fala. O áudio é então convertido em *embeddings* de fala, que são agrupados em clusters para distinguir os locutores.

Para demonstrar a eficácia, foi usada uma amostra do AMI Meeting Corpus (Carletta et al., 2006). A aplicação gerou uma sumarização detalhada com a identificação dos locutores, evidenciando a capacidade do sistema em cenários com múltiplos falantes, como ilustrado na Figura 4.



Project Kickoff: Designing a New Remote Control

Introductions (0:00 - 1:20)

The meeting begins with introductions. The team consists of:

Betsy: Project Manager

Francia: User Interface Designer

Ilyen: Marketing

Janvi: Industrial Designer

Project Overview (1:20 - 4:00)

Betsy presents the project goals:

Goal: Design a new remote control that is original, trendy, and user-friendly.

Design Phases: Individual work followed by group discussion for Functional, Conceptual, and Detailed Design.

Tools: Computers and a whiteboard.

Whiteboard Demonstration (4:00 - 5:15)

Ilyen and Francia experiment with the whiteboard, drawing a cat and a snake.

Project Finances (5:15 - 5:35)

Betsy outlines the financial targets:

Target Selling Price: €25

Profit Aim: €50 million

Maximum Production Cost: €12.50

Brainstorming Session (5:35 - 8:00)

The team brainstorms ideas for the new remote:

Key Features:

Multi-functionality: Control other devices like air conditioners and heating systems.

Unique Design: Trendy colors, materials, and shapes.

Locator Feature: A beep or light activated by a sensor when the remote is lost (e.g., under a blanket).

Discussion Points:

How to make the remote compact and ergonomic.

The effectiveness of voice or clap activation for the locator feature.

The importance of a reliable sensor-based activation for the locator feature.

Next Steps (8:00 - 8:39)

The next meeting is scheduled in 30 minutes.

Individual Assignments:

Janvi: Develop a working design and technical specifications.

Ilyen: Define user requirements and price justification.

Team members will receive detailed instructions.

FIGURA 4. Resultados da sumarização após as etapas de transcrição e diarização

CONCLUSÕES

Este trabalho apresentou o desenvolvimento de uma técnica para a sumarização automática de vídeos digitais, baseada na integração de transcrição de fala e modelos generativos de última geração. Os resultados alcançados demonstraram a capacidade do algoritmo na geração de resumos textuais concisos e informativos a partir de vídeos de entrada. A transcrição realizada pelo modelo Whisper, combinada com a diarização fornecida pelo pyannote.audio, permitiu uma separação dos diferentes participantes em um vídeo.

Além disso, o método mostrou-se eficiente na criação de sumários segmentados por seções, representando de forma clara o conteúdo abordado em diferentes *timestamps* do vídeo. No entanto, foram observadas algumas limitações, como a lentidão no processamento durante as etapas de diarização e transcrição, que pode ocorrer devido à complexidade computacional envolvida e à qualidade dos dados de entrada.

Apesar dessas limitações, a contribuição deste trabalho é significativa, pois oferece uma solução robusta para o problema da sobrecarga de informações em vídeos digitais. A aplicação *desktop* tem potencial de aplicação no processamento de reuniões, vídeos multilíngues, podcasts, entre outros, onde a capacidade de obter um resumo claro e preciso é crucial.

CONTRIBUIÇÕES DOS AUTORES

Bruno Mascioli de Souza e Tiago Catoia de Souza contribuiram com a pesquisa, desenvolvimento e implementação do software, bem como com a redação do trabalho. Tiago Henrique Trojahn foi responsável pela orientação e supervisão.

Todos os autores participaram da revisão e confecção do trabalho e deram sua aprovação à versão submetida.

AGRADECIMENTOS

Gostaria de expressar meu sincero agradecimento ao CNPq pelo apoio financeiro concedido através da bolsa PIBITI, que foi essencial para o desenvolvimento deste projeto. Agradeço também ao professor orientador do projeto, cujo apoio foi fundamental para a realização deste trabalho, e ao colaborador técnico, cuja dedicação e contribuição na implementação do software foram essenciais para o sucesso do projeto.

REFERÊNCIAS

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7124-7128). IEEE.

Carletta, J. et al. (2006). The AMI Meeting Corpus: A Pre-announcement. In: Renals, S., Bengio, S. (eds) Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science, vol 3869. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11677482_3

Manzato, M. G. Uma arquitetura de personalização de conteúdo baseada em anotações do usuário. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2011. Disponível em: https://dx.doi.org/10.11606/T.55.2011.tde-11042011-160836

TOFFLER, Alvin. Future Shock. Nova York: Random House, 1970.

RADFORD, Alec et al. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, 2023. p. 28492-28518.

SINEK, Simon. How great leaders inspire action. YouTube, 2010. Disponível em: https://www.youtube.com/watch?v=qp0HIF3SfI4. Acesso em: 09 set. 2024.

VASWANI, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems.