Speech recognition: Speech-to-Text

Abstract

Speech recognition technology plays a pivotal role in enhancing user experience across various applications, particularly in the context of video chat platforms. The ability to transcribe spoken words into text not only facilitates accessibility for users with hearing impairments but also enables the creation of searchable and indexable context from video conversations. This project seeks to develop robust speech-to-text system tailored for video chats. Due to compactional constraints, our approach revolves around the use of Deep Recurrent Neural Network (RNN) with Connectionist Temporal Classification (CTC). While more advanced models like transformers and conformers have exhibited superior performance, the primary focus here is on achieving reliable results within limited compactional resources. The proposed system holds promise for significantly improving user experience and accessibility in video communication platforms.

Introduction

The growing demand for efficient speech recognition systems is not only driven by the widespread adoption of video communication platforms but is also deep rooted in the changing dynamics of work, education, and collaboration. The prevalence of remote work, virtual meetings, and online collaboration has underscored the crucial role of seamless communication, emphasizing the need for robust speech-to-text solutions.

One significant aspect of these systems is the enhancement of accessibility for users with hearing impairments. The transformation of spoken words into written form empowers individuals facing hearing challenges to fully participate in and comprehend conversations conducted through video chat platforms. This inclusivity not only aligns with societal goals but also contributes to creating equal participation and communication opportunities, fostering a more inclusive digital environment.

Beyond accessibility, the conversion of spoken words into written text has opened new avenues, especially in the realm of education. Students now can access lecture transcriptions, providing them with a valuable resource for review, study and comprehension. This application of speech-to-text technology in the educational domain represents a significant advancement, catering the diverse learning needs of students and promoting more effective knowledge dissemination.

In professional settings, the conversion of spoken words into searchable content has become increasingly vital. This functionality facilitates efficient information retrieval, allowing users to quickly locate specific details, revisit discussions, and extract valuable insights form video conversations. In the context of remote work and virtual collaboration, where effective documentation is crucial, the speech-to-text technology proposed in this project holds promise for streamlining workflows and enhancing overall productivity.

Given the computational constraints in certain environments, the project's emphasis on developing a lightweight speech recognition solution acknowledges the need to balance performance and computational efficiency. While more advanced models like transformers and conformers showcase remarkable accuracy, their resource-intensive nature can be a barrier, particularly in resource-constrained settings.

In essence, the project aligns with the evolving landscape of communication technologies, aiming to democratize the benefits of the speech-to-text technology.by addressing the needs of diverse user base, including individuals with hearing impairments, students, and those with various disabilities, the project contributes to fostering a more inclusive and accessible digital communication environment.

Related Work

The landscape of speech recognition research has evolved significantly, encompassing a spectrum of models and methodologies aimed at transcribing spoken language into written text. This section delves into the diverse body of related works, spanning traditional methods to contemporary approaches, shedding light on the progression of the field and contextualizing the current project within this trajectory.

1. Traditional speech recognition models:

Historically, Hidden Markov Models (HMM's) and Gaussian Mixture Models (GMMs) formed the bedrock of speech recognition systems. These models, though foundational, grappled with capturing the intricate linguistic patterns, resulting in limitations regarding accuracy and adaptability across various environments.

1. Deep Learning approaches.

The advent of deep learning ushered in a new era, prominently featuring Recurrent Neural Networks -RNN. renowned for their capacity to comprehend temporal dependencies in sequential data, RNN's became a linchpin in speech recognition. The integration of Connectionist Temporal Classification -CTC as a a training criterion addressed challenges associated with variable-length sequences, enhancing the model's effectiveness.

1. Transformer architecture.

Vaswani et al. (2017) introduced the transformer architecture, originally designed for natural language processing. Transformers revolutionized speech recognition by capturing long-range dependencies more effectively than RNN's.

However, their computational demands, particularly for self-attention mechanisms, raised concerns about their practicality in resource-constrained environments.

1. Conformer architecture.

Recognizing the computational challenges by transformers, the conformer architecture emerged as a hybrid solution. By combining convolutional neural networks (CNN's) with transformers, models like conformer-Kd achieved improved efficiency without compromising accuracy. This development showcased the potential for optimizing computational resources while maintain high-performance levels in speech recognition tasks.

1. Attention mechanisms.

Attention mechanisms assumed a pivotal role in refining the accuracy of speech recognition systems. Enabling models focus on specific segments of the input sequence, attention mechanisms facilitated the capture of contextual information and nuanced features present in spoken language.

1. Multimodal approaches.

Recent research ha explored the fusion of visual information with audio signals to elevate speech recognition capabilities. Multimodal approaches leverage both auditory and visual cues to enhance robustness and accuracy, particularly in challenging environments or amidst noise.

In this project, a distinctive emphasis is placed on developing a lightweight speech recognition solution. While advanced models like transformers and conformers have demonstrated remarkable accuracy, their computational demands can be prohibitive. Opting for deep RNNs with CTC represents a strategic choice aimed at providing an efficient and practical solution, tailored to the challenges of real-time video communication platforms with limited computational resources. This decision positions the project uniquely within the landscape of speech recognition research, addressing a critical need for accessible and lightweight solutions in contemporary communication technologies.