Video Translation using OCR and Speech Recognition

Atharva Murmure
Department of Computer Science Engineering - AIML
Vishwakarma Institute of Technology
Pune, India

Dhanashri Rajput
Department of Computer Science Engineering - AIML
Vishwakarma Institute of Technology
Pune, India

Aaditya Agarwal
Department of Computer Science Engineering - AIML
Vishwakarma Institute of Technology
Pune, India

Sangeeta Jaybhaye
Department of Computer Science Engineering - AIML
Vishwakarma Institute of Technology
Pune, India

Devashish Kanhere
Department of Computer Science Engineering - AIML
Vishwakarma Institute of Technology
Pune, India

Abstract— Video content is a powerful medium for communication and storytelling, offering a versatile approach to reach diverse audiences effectively. However, language barriers can hinder accessibility and inclusivity, particularly for non-native speakers. To address this challenge, we propose an innovative video translation system that leverages Optical Character Recognition (OCR) and Speech Recognition technologies. This system extracts textual content from video frames and transcribes spoken dialogue, subsequently translating this information using advanced machine translation algorithms. The translated text is overlaid onto the video frames, while the translated audio is synchronized with the original soundtrack, resulting in a fully accessibility of video content, facilitates communication across language barriers, and improves the localization of videos for global distribution. By translating videos without altering the original visuals, the system maintains the integrity of the content while making it comprehensible to a broader audience. This approach not only broadens the reach of video content but also ensures that messages and stories are more inclusive and impactful, catering to a global viewership. The proposed system represents a significant advancement in video translation, promoting inclusivity and enhancing the effectiveness of global communication.

Keywords: Video Translation, Optical Character Recognition (OCR), Speech Recognition, Machine Translation, Accessibility, Global Distribution, Localization, Global Communication, Textual content, Video Frames

INTRODUCTION

The advent of advanced translation technologies has revolutionized how we interact with multimedia content in different languages. Communication is a cornerstone of human existence, facilitating societal progress and cultural exchange. Without effective communication, the global reach of movies and videos would be limited, especially when presented in a foreign language unfamiliar to the audience. Traditional methods of video translation are labor-intensive and time-consuming, often requiring significant human effort to translate a single movie. This work can be streamlined with the development of automated translation programs that leverage Optical Character Recognition (OCR) and Neural Machine Translation (NMT) technologies.

Recent advancements have introduced the task of unsupervised video-to-video translation, which addresses the challenges of maintaining realistic motion and continuity. This is achieved through a spatio-temporal 3D translator that surpasses per-frame methods in generating coherent video sequences. Evaluations on synthetic and realistic datasets demonstrate the superiority of this 3D approach in capturing the complex structure and motion in videos [1]. Moreover, the integration of OCR and translation technologies has led to the development of innovative image translators. These tools detail necessary functionalities and implementation methods to enhance user accessibility to translation, proving their effectiveness through rigorous performance tests [2].

Further enhancing translation capabilities, a hybrid methodology combining NMT with OCR has been proposed. This technique focuses on recognizing text, such as Hindi, and translating it to English and vice versa. By leveraging the strengths of both OCR and NMT, this approach optimizes text detection and translation from images, offering a robust solution for multilingual content translation [3]. By combining these advancements, it becomes feasible to create a seamless video translation system. The process involves dividing the video into frames, extracting text using OCR, translating the text into the preferred language, converting it to audio, and then recombining everything to produce a translated video. This approach significantly reduces the effort and time required for video translation, making it more accessible and efficient.

LITERATURE REVIEW

The growing need for efficient and accurate text extraction and translation from multimedia sources has driven significant advancements in this field. Various approaches have been developed to enhance information storage, accessibility, and understanding from videos and audio files. One system proposes the extraction and conversion of text from educational and news videos into editable text files, thereby enhancing information storage and accessibility for efficient revision and future reference [4]. Another development focuses on a video and audio to text converter aimed at improving documentation processes for software firms, educational institutions, and other organizations. Utilizing Google Speech Recognition for its accuracy and a Python-based interface for ease of use, this system simplifies converting single audio files to text, thus improving access to notes, project details, and presentations [5].

In addressing the challenge of generating natural language descriptions for videos, a transformer network with a deep attention-based encoder and decoder has been proposed. Unlike traditional GRU or LSTM models that rely on the final hidden state of the encoder, this transformer network processes sequences holistically, leveraging attention mechanisms to capture relationships between all elements in the sequence. This method aims to enhance video understanding for applications such as video indexing, retrieval, and sign language translation by providing more accurate and comprehensive descriptions [6].

Further advancements include a Transformer-based approach for Video-to-Text (VTT) tasks aimed at automatically generating descriptions for short audio-visual clips. The proposed method incorporates Fractional Positional Encoding (FPE) to synchronize audio and video features, with experiments on the VATEX dataset showing significant improvements in CIDEr and BLEU-4 scores, achieving state-of-the-art results on the MSR-VTT and MSVD datasets, and an additional 8.6% relative increase in CIDEr score due to FPE [7]. Research on video text recognition has advanced through varied OCR methods tailored for the unique challenges of video frames, including motion, noise, and multi-script content. Uchida [10] emphasizes adaptive segmentation for content-based video retrieval, enhancing indexing and summarization. Hua and Wenyin [11] focus on adaptive binarization to maintain OCR accuracy in dynamic text, essential for applications like live translation. Mirza et al. [12] address multi-script and cursive text with a framework that integrates multiple OCR models, supporting multilingual video indexing. Finally, Elagouni and Garcia [13] employ a recurrent neural network (RNN) for robust recognition under fluctuating video conditions, crucial for real-time applications. Together, these works advance OCR for video, aiding applications from content retrieval to live video analytics.

The reviewed papers collectively highlight the transformative role of advanced technologies like OCR, speech recognition, and deep learning in multimedia processing. Government initiatives, such as India's PIB project, showcase how AI-driven multimedia dissemination can enhance public engagement through multilingual video content, leveraging text summarization, translation, and 2D animation [14]. In regulatory applications, deep learning-based systems efficiently detect and classify promotional content on platforms like TikTok, aiding policy enforcement [15]. Further, multi-modal video segmentation techniques integrate audio transcription and OCR outputs, achieving precise clustering and subtopic extraction [16]. Bilingual text transcription and translation demonstrate how OCR Tesseract combined with deep learning improves accuracy and contextual preservation for multilingual content [17]. For educational and interactive purposes, multilingual temporal answer grounding frameworks employ LLMs, ASR, and OCR to address linguistic gaps and enhance video comprehension, particularly in silent or low-quality audio settings [18]. These advancements underline the potential of integrating machine learning, language processing, and multimedia tools for cross-disciplinary applications spanning public communication, regulatory frameworks, and accessible education.

METHODOLOGY

A. *System Architecture:*

Read File: The process begins by reading an input file, which could be in various formats like PDFs, Word documents, images, or video files.

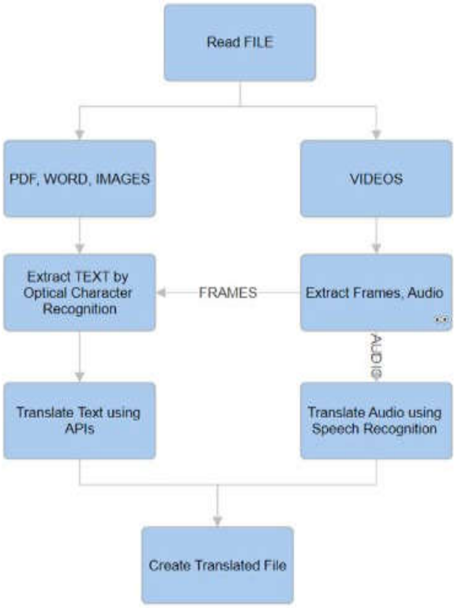
File Type Differentiation:
If the file is a PDF, Word document, or Image, it directly undergoes text extraction using OCR to identify and digitize any text content.
If the file is a Video, it first goes through a step of frame and audio extraction to separate visual frames (which might contain text) and audio components.

Text and Audio Processing:
For visual frames (from images or video), the OCR module extracts text content, such as subtitles or on-screen text.
For audio (from video), speech recognition is applied to convert spoken language into text.

Translation:
Extracted text (whether from OCR or speech recognition) is then translated using APIs, likely involving machine translation services to convert the text into the desired language.

Output: Finally, the translated text is compiled into a translated file, providing a version of the original content in the target language.

Flow chart:


Read FILE
PDF, WORD, IMAGES
VIDEOS
Extract TEXT by Optical Character Recognition
FRAMES
Extract Frames, Audio
AUDIO
Translate Text using APIs
Translate Audio using Speech Recognition
Create Translated File

a. File Input Categorization:

Text Files: This category includes images and Word documents. These files are subjected to Optical Character Recognition (OCR) to extract the textual content. Video Files: This category includes all video formats. The video files are divided into individual frames, and the corresponding audio is extracted from these files.

-Text Extraction and Frame Division: Text Files: For images and Word documents, OCR technology is applied to each file to convert the visual text into editable text. This process involves analysing the text and segmenting it into frames, which are smaller, manageable units of text.

b. Preprocessing:

The preprocessing pipeline involves handling PDFs, PPTs, and audio files to extract and process relevant content. For PDFs, each page is converted into an image using PyMuPDF, which is then processed and compiled back into a PDF after applying necessary transformations. PPT files are processed using a library like Spire.Presentation or python-pptx, converting each slide into an image while ensuring font compatibility with a fallback mechanism to handle missing fonts. For audio files, FFmpeg is used to convert various formats to WAV, which are then split into 10-second chunks. Each chunk undergoes transcription using Google Speech Recognition, followed by translation via Google Translate, and conversion back to audio using gTTS. Finally, all translated audio chunks are concatenated into a single audio file. This unified process ensures efficient handling and processing of diverse input types, maintaining a seamless workflow.

c. Video Files:

The video files are processed to separate the visual and audio components. The video frames are isolated, and the audio track is extracted for further processing.

d. Text Translation:

The extracted text from both text files and video frames is then translated into the desired language using advanced translation APIs. These APIs ensure accurate and contextually appropriate translations, leveraging large language models and neural machine translation techniques. - Audio Recognition and Translation: The extracted audio from video files undergoes speech recognition to convert spoken words into text. This text is then translated into the target language using the same translation APIs employed for text files. -The translated text is converted back into speech using text-to-speech (TTS) technologies, ensuring that the translated audio retains the original intonation and emotional context of the speaker.

e. Integration of Translated Components:

The translated text frames and audio files are systematically combined to form a coherent and integrated output. For videos, this involves synchronizing the translated audio with the video frames to maintain the original timing and sequence. For text files, the translated text is reformatted to match the layout and structure of the original document, ensuring readability and coherence.
-Creation of Final Translated Output:
The final step involves compiling all the translated components into a single, comprehensive file. This file is designed to preserve the original format and content structure while providing the translated text and audio.
For video files, the translated audio is merged with the video frames to produce a fully translated video file. For text files, the translated text is compiled into documents that mirror the original layout and design.

B. Components:

-Input Handling:
File Input Module: This module allows users to upload various types of files, including images, Word documents, and video files. It supports multiple file formats and provides a user-friendly interface for easy file selection and upload.

-Text Extraction:
OCR Module: Utilizes Optical Character Recognition (OCR) technology to extract text from images and Word documents. The module processes each file, identifies text regions, and converts the visual text into editable and searchable text.
Frame Division Module: Processes video files by dividing them into individual frames. This allows for detailed analysis and manipulation of each frame separately, facilitating better text and image handling.

-Audio Extraction:
Audio Extraction Module: Extracts the audio track from video files. This module ensures that the audio is isolated and prepared for subsequent processing, such as transcription and translation.

Translation:

-Text Translation API: Connects to advanced translation APIs to convert the extracted text into the desired language. The API ensures high accuracy and contextual relevance in the translations.

-Speech Recognition API: Converts the extracted audio into text using speech recognition technologies. This module handles various accents and speech patterns to ensure accurate transcription.

-Text-to-Speech (TTS) API: Converts the translated text back into speech. The TTS module uses natural-sounding voices to maintain the original intonation and emotional context of the audio.

Integration:

-Synchronization Module: Aligns the translated audio with the corresponding video frames to ensure proper synchronization. This module ensures that the timing and sequence of the translated audio match the original video.
-Text Formatting Module: Reformat the translated text to match the layout and structure of the original documents. This includes maintaining fonts, headings, and other formatting elements to ensure the translated document is as close to the original as possible.

-Output Generation:

Compilation Module: Compiles all translated components into a single, cohesive output file. For videos, this means merging the translated audio with the video frames. For text documents, this involves compiling the translated text into a formatted document.
-File Export Module: Provides options for users to download the final translated files in various formats. This module ensures that the output is easily accessible and downloadable.

-User Interface:

Input Interface: A user-friendly interface that allows users to upload their files. This interface guides users through the upload process and ensures that files are correctly categorized and processed.
-Progress Tracking Interface: Provides real-time feedback on the status of the file processing. Users can see the progress of each stage, from upload to final output generation.
-Output Interface: Enables users to download the translated files. This interface provides links or buttons to access and download the final output.

-Error Handling and Logging:
Error Detection Module: Monitors the entire process for errors and logs them. This module helps identify issues such as file format incompatibility or processing errors.
Error Resolution Module: Provides solutions or workarounds for common errors. This module offers automatic fixes where possible and gives users guidance on how to resolve issues manually.

-Storage and Management:
Temporary Storage Module: Stores intermediate files during the processing stages. This ensures that all data is securely saved and can be accessed or reprocessed if needed.

Final Output Storage: Saves the final translated outputs for user access. This storage solution ensures that users can retrieve their translated files even after the processing is complete.

-Documentation and Support:
User Guide: Provides detailed instructions on how to use the system, from uploading files to downloading translated outputs. The guide ensures that users can easily understand and utilize the system's features.

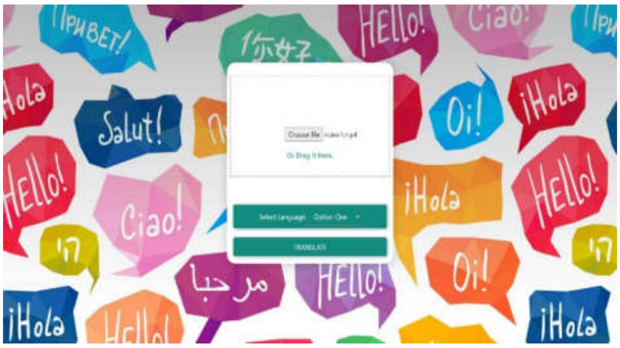


Fig 1. Glimpse of the website

C. Implementation format

a. Development of Environment:

IDEs: Set up Integrated Development Environments (IDEs) such as Visual Studio Code, PyCharm, or Eclipse. These IDEs provide code editing, debugging, and testing tools.
Software Installations: Ensure all necessary software and libraries are installed. This includes Python, pip, OpenCV for video processing, Tesseract for OCR, and FFmpeg for audio extraction and conversion.

b. Text Extraction Module:

OCR Integration: Integrate Tesseract OCR for extracting text from images and documents. Preprocess images to enhance OCR accuracy by adjusting contrast, removing noise, and converting to grayscale.
Document Parsing: Develop parsers for different document formats (e.g., DOCX, PDF) to extract text content. Use libraries like python-docx and PyPDF2.
Text Segmentation: Implement algorithms to segment extracted text into meaningful chunks or sentences for easier translation and synchronization

c. Audio Extraction Module:

FFmpeg Integration: Use FFmpeg to extract audio from video files. Create scripts to automate the extraction process and handle various video formats.
Audio Preprocessing: Preprocess extracted audio to enhance speech recognition accuracy. This includes noise reduction, volume normalization, and format conversion to WAV.

d. Text Translation API:

API Selection: Choose robust translation APIs like Google Translate, Microsoft Translator, or AWS Translate. Evaluate their performance and language support.
API Integration: Implement API calls to translate extracted text. Handle authentication, request formation, and response parsing.

e. Speech Recognition and TTS:

Speech Recognition Integration: Integrate speech recognition APIs such as Google Speech Recognition or IBM Watson to convert audio to text. Preprocess audio files to match the API requirements.
TTS Integration: Use Text-to-Speech (TTS) APIs like Google Text-to-Speech or Amazon Polly to convert translated text back into audio. Ensure the generated audio maintains natural intonation and clarity.

FUTURE SCOPE

The future scope of this project is promising, with numerous enhancements that can significantly improve its utility, performance, and accessibility. Expanding the range of supported languages for translation and text-to-speech (TTS) functionalities, including recognition and translation of different dialects and regional accents, is a key goal. Advanced AI models will be utilized to enhance OCR accuracy, particularly for handwritten text and complex layouts, and to implement more context-aware translations. Real-time processing capabilities will be developed, allowing for the translation of live video feeds during webinars and broadcasts, and instant audio translation during conversations or meetings. User experience will be improved by allowing interactive editing of transcriptions and translations, and offering more customization options such as TTS voice selection, translation quality, and subtitle styling.
The project will integrate with Learning Management Systems (LMS) to automatically translate lecture videos and course materials, and with Content Management Systems (CMS) for seamless translation and transcription of digital content. Accessibility features will include real-time captioning and sign language recognition for the hearing impaired, and improved TTS to support visually impaired users with natural-sounding

voices and better text handling. Finally, scalability and performance will be enhanced by migrating to cloud-based processing to handle larger data volumes and exploring edge computing to reduce latency and improve real-time application performance, especially in remote areas.

CONCLUSIONS

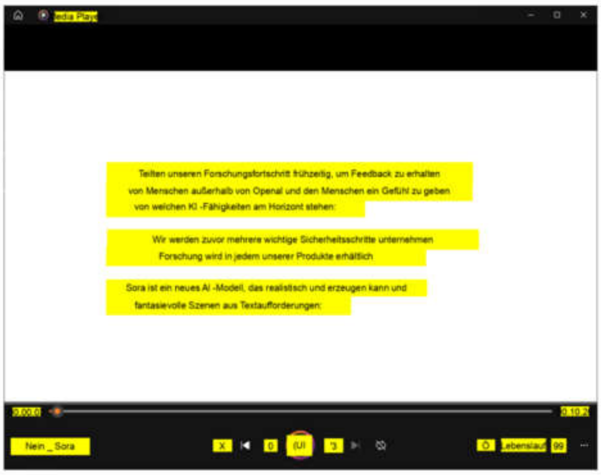
This project introduces a sophisticated system aimed at enhancing the accessibility and usability of multimedia content by integrating advanced OCR, speech recognition, and machine translation technologies. It efficiently converts text from images, documents, and videos into editable and translatable formats, effectively bridging language barriers and democratizing information access.

The system's modular design supports a wide array of file formats and devices, making it an invaluable tool for educational institutions, software firms, and other organizations. Its successful implementation showcases its potential to streamline content management processes, from translating lectures in educational environments to automating transcription in corporate settings, by providing accurate and context-aware translations.

As the system evolves, it holds significant potential to become an essential tool across various fields, fostering inclusivity and understanding in a globalized world. This project not only addresses current demands but also lays a robust foundation for future advancements in multimedia content translation and accessibility.

RESULTS

The results indicate that the transcription accuracy of the mp3 file was much better than that of the video. The overlaid text in the video was accurately identified, and the same goes for the text overlaid on the PDF document.
The results are presented below.



Teilten unseren Forschungsfortschritt frühzeitig, um Feedback zu erhalten von Menschen außerhalb von Openal und den Menschen ein Gefühl zu geben von welchen KI -Fähigkeiten am Horizont stehen:

Wir werden zuvor mehrere wichtige Sicherheitsschritte unternehmen Forschung wird in jedem unserer Produkte erhältlich

Sora ist ein neues AI -Modell, das realistisch und erzeugen kann und fantasievolle Szenen aus Textaufforderungen:

Fig2. This is the video translated file

REFERENCES

[1] D. Bashkirova, B. Usman, and K. Saenko, "Unsupervised Video-to-Video Translation," ArXiv, 2018. [Online]. Available: https://arxiv.org/abs/1806.03698.

[2] S.-M. Hwang and H.-G. Yeom, "An Implementation of a System for Video Translation Using OCR," in Proc. 2021, pp. 4. doi: 10.1007/978-3-030-64773-5_4.

[3] C. Kutur, M. Cross, and V. Vasudevan, "Optical Character Recognition and Neural Machine Translation Using Deep Learning Techniques," in Proc. 2021, pp. 30. doi: 10.1007/978-981-33-4543-0_30.

[4] S. Ramiah, T. Y. Liong, and M. Jayabalan, "Detecting text based image with optical character recognition for English translation and speech using Android," in 2015 IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, Malaysia, 2015, pp. 272-277. doi: 10.1109/SCORED.2015.7449339.

[5] G. S. Hukkeri, R. H. Goudar, P. Janagond, and P. S. Patil, "Machine Learning in OCR Technology: Performance Analysis of Different OCR Methods for Slide-to-Text Conversion in Lecture Videos," Department of CSE, VTU Belagavi, India.

[6] K. Agre, S. Gaonkar, A. Chheda, and M. Patil, "Text Recognition and Extraction from Video," Atharva College of Engineering Mumbai, India.

[7] M. Saraswathi, V. Ronit, and S. S. Pranav, "Implementation of Video and Audio to Text Converter," SCSVMV, Kanchipuram, pp. 1-6.

[8] Uchida, Seiichi. (2014). Text Localization and Recognition in Images and Video. 10.1007/978-0-85729-859-1_28.

[9] Xi, Jie & Hua, Xian-Sheng & Chen, Xiang-Rong & Wenyin, Liu. (2001). A video text detection and recognition system. 2012 IEEE International Conference on Multimedia and Expo. 222. 10.1109/ICME.2001.1237861.

[10] Elagouni, K., Garcia, C., Mamalet, F., Sébillot, P. (2012). Text Recognition in Videos Using a Recurrent Connectionist Approach. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds) Artificial Neural Networks and Machine Learning – ICANN 2012. ICANN 2012. Lecture Notes in Computer Science, vol 7553. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33266-1_22

[11] Detection and recognition of cursive text from video frames Ali Mirza1*, Ossama Zeshan1, Muhammad Atif1 and Imran Siddiqi1 Mirza etal. EURASIPJournalonImageandVideo Processing(2020)2020:34

[12] N. Anitha Devi, M. Mohamed Saleh, K. G. Arvind Srinivas, S. Mathalai Khishan and V. Haresh, "Automated Multilingual Multimedia Dissemination Of Government Press Releases," *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, 2024, pp. 1-7, doi: 10.1109/ICITIIT61487.2024.10580027.

[13] N. V. Raviteja Chappa, C. McCormick, S. R. Gongora, P. D. Dobbs and K. Luu, "Advanced Deep Learning Techniques for Tobacco Usage Assessment in TikTok Videos," *2024 IEEE Green Technologies Conference (GreenTech)*, Springdale, AR, USA, 2024, pp. 162-163, doi: 10.1109/GreenTech58819.2024.10520426.

[14] M. Vasuki, M. Arun Gangadharan, J. T. Daniel, A. Sadashiv, V. Venugopal and S. Vekkot, "Multi-Modal Automatic Video Segmentation with Sentence Transformer Embeddings and KeyBERT-Based Subtopic Extraction," *2024 2nd World Conference on Communication & Computing (WCONF)*, RAIPUR, India, 2024, pp. 1-6, doi: 10.1109/WCONF61366.2024.10692267.

[15] P. D. Cerna, R. J. Cascaro, K. D. E. Laurente, J. V. B. Cabahug, D. W. B. Carino and J. H. Maraguinot, "Transcribing

and Translating Bilingual Text using OCR Tesseract and Deep Learning," *2024 13th International Conference on Educational and Information Technology (ICEIT)*, Chengdu, China, 2024, pp. 30-35, doi: 10.1109/ICEIT61397.2024.10540714.

[16] Zhang, H., Zheng, C., He, Y., Zhao, Y., Lai, Y. (2025). Improving Multilingual Temporal Answering Grounding in Single Video via LLM-Based Translation and OCR Enhancement. In: Wong, D.F., Wei, Z., Yang, M. (eds) Natural Language Processing and Chinese Computing. NLPCC 2024. Lecture Notes in Computer Science(), vol 15363. Springer, Singapore. https://doi.org/10.1007/978-981-97-9443-0_12

[17] M. N and A. James, "Transformer Network for video to text translation," in 2020 International Conference on Power, Instrumentation, Control and Computing (PICC), Thrissur, India, 2020, pp. 1-6. doi: 10.1109/PICC51425.2020.9362374.

[18] P. Harzig, M. Einfalt, and R. Lienhart, "Synchronized Audio-Visual Frames with Fractional Positional Encoding for Transformers in Video-to-Text Translation," in 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 2041-2045. doi: 10.1109/ICIP46576.2022.9897804.