Extended Reality Sign Translating Platform (STP)(version 2)

submitted by :zineb ait el mouaddin

submitted to
jean bosco nsekuye

30. august 2024

1 Introduction

1.1 Context and Motivation

Sign language is an essential mode of communication for deaf and hard-of-hearing individuals. It uses visual gestures, facial expressions, and body movements to convey information and express ideas. Around the world, different deaf communities use various sign languages, each with its own grammar, lexicon, and syntactic structures. These sign languages are as rich and complex as spoken languages and play a crucial role in the social and cultural integration of deaf people.

However, communication between deaf individuals who use sign language and hearing people who do not know this language remains a significant challenge. The language barrier can lead to misunderstandings, lack of access to important information, and social exclusion. To overcome this barrier, it is essential to develop technologies that enable smooth translation between sign language and spoken languages. Such a bidirectional translation system could not only enhance accessibility and inclusivity but also promote better understanding and closer collaboration between deaf and hearing communities.

In the field of research on sign language translation, most efforts have focused on translating sign language to text (Sign Language to Text, SLT). These automatic translation systems attempt to recognize gestures and movements in sign language videos and convert them into text. Although significant progress has been made, these systems still face several limitations.

First, the available sign language datasets are often limited in size and diversity. Most existing datasets contain a relatively small number of signs and phrases, limiting the models' ability to generalize to larger datasets. For example, many datasets focus on specific sign languages and do not include a wide range of contexts or idiomatic expressions. This limitation directly affects the accuracy and robustness of translation models, making it difficult to apply them in varied real-life situations.

Secondly, most current translation systems primarily support a single mode of translation, usually from sign language to text. These systems do not support bidirectional translation, meaning the ability to translate not only from sign language to text but also from speech or text to sign language. This means that communication remains asymmetric: while deaf individuals can benefit from translating their signs into text or speech, hearing people do not have a solution to translate their speech into signs, thus limiting interactivity and reciprocity in communication.

Finally, multilingual translation remains another significant limitation. Most current systems are designed to work with a single target spoken language, typically English. This restricts their usefulness in a global context where deaf people use various sign languages and often need to communicate with speakers of different spoken languages.

This project aims to address these limitations by developing a bidirectional and multilingual sign language translation system using advanced technologies such as Transformers and attention models and leveraging datasets like How2Sign. One of the key advantages of our innovative approach is its ability to operate in real-time, enabling instant and seamless communication between users. Additionally, the system is designed to handle the translation of long speeches, which is essential for applications such as conferences, university lectures, or public speeches, where complex and detailed information must be translated without loss of context or fluency.

Our inclusive translation platform aims to facilitate communication between deaf and hearing individuals, regardless of language barriers, while providing an efficient and scalable solution that meets the diverse needs of users in an increasingly connected and multilingual world.

1.2 Project Objective

The main objective of this project is to develop an advanced translation system capable of converting signs into speech and vice versa while supporting multiple languages. This bidirectional system aims to address the gaps in current sign language translation technologies by providing a comprehensive and inclusive solution that facilitates smooth communication between deaf and hearing individuals, regardless of the spoken or signed language used.

1.2.1 Project Goal

This project aims to design and implement an innovative platform that:

- Translates Sign Language into Speech: By recognizing gestures and movements performed in sign language videos, the system can convert them not only into text but also directly into audible speech. This feature allows for more natural and intuitive communication with hearing people, who can hear the translations in real time.

- Converts Speech into Signs: The system is also capable of translating speech into sign language through text. It does this by recognizing speech, transcribing it into text, and then converting it into a video representation of sign language, thereby facilitating understanding for deaf individuals.

- Supports Multiple Languages: Unlike traditional systems that are generally limited to a single language, our project supports multiple languages, both for sign languages and spoken languages. This means the system can translate signs into English, French, Spanish, and vice versa, making the system adaptable to a global and multilingual context.

- Operates in Real-Time: One of the key advantages of our innovative approach is its ability to operate in real-time, enabling instant and seamless communication between users. This real-time functionality is crucial for effective interaction in dynamic environments where immediate responses are necessary.

- Handles Long Speeches: The system is designed to handle the translation of long speeches, which is essential for applications such as conferences, university lectures, or public speeches, where complex and detailed information must be translated without loss of context or fluency. This capability ensures that even extended dialogues or presentations are accurately and efficiently translated, maintaining the coherence and integrity of the original message.

By incorporating these aspects, the project aims to create a versatile and powerful translation platform that enhances accessibility and inclusivity for deaf and hearing communities worldwide.

2 System Design and Translation Scenarios

In our project, we aim to develop a bidirectional translation system that generally supports two main scenarios: Speech-to-Sign translation and Sign-to-Speech translation. These two processes are designed to maximize accessibility and inclusivity, enabling seamless communication between deaf and hearing users, regardless of the spoken or signed language used.

2.1 Speech-to-Sign Case

The Speech-to-Sign process aims to translate a user's speech into sign language, thereby allowing deaf or hard-of-hearing individuals to understand spoken language visually. Here are the steps and methods used in this case:

- **User Input (Speech):**

- The user speaks in their chosen language (e.g., French, Spanish, German). The system must be able to recognize the spoken language to provide accurate translation.

- **Speech Recognition and Translation to English:**
  - **Method:** Use a speech recognition model (Speech-to-Text, STT) such as Google Speech-to-Text API or Mozilla's DeepSpeech. These models are trained on large speech datasets and can accurately transcribe speech into text in the source language.
  - **Translation to English:** Once the speech is transcribed into text in the source language, the text is then translated into English to match the English annotations of the How2Sign dataset. For this step, we use multilingual automatic translation models such as MarianMT or Google Translate API. These models enable fast and accurate translation between multiple spoken languages and English.

- **Translation of Text to Signs:**
  - **Method:** The English text is then used as input for a Text-to-Sign synthesis model. This model is based on deep learning techniques such as Transformers and is trained on the How2Sign dataset to learn to generate gesture sequences corresponding to the text.
  - **Sign Video Generation:** The Text-to-Sign model produces a video of sign language using either animated 3D avatars or pre-recorded sign videos corresponding to the English words. The user can view this video to understand the translation of the initial speech.

2.2 Sign-to-Speech Case

The Sign-to-Speech process allows for translating gestures into speech, making communication accessible to hearing individuals who do not know sign language. Here are the steps and methods used in this case:

- **Sign Language Video Input:**
  - The user provides a sign language video. This video is then analyzed to detect and extract the gestures and movements of sign language.

- **Translation of Signs to English Text:**
  - **Method:** Use a Sign Language to Text (SLT) translation model. This model is based on Transformer architectures. The SLT model will be trained on the How2Sign dataset, which means it produces translations in English, the language in which the dataset annotations are provided.

- **Translation to Target Language and Speech Synthesis:**
  - **Translation to Target Language:** Once the English text is generated, it is translated into the target language chosen by the user. For this step, we use automatic translation models such as MarianMT or Google Translate API,

which allow for quickly converting English text into multiple target languages (e.g., French, Spanish, German, etc.).

- **Speech Synthesis (Text-to-Speech, TTS):** The translated text is then converted into speech using a text-to-speech synthesis model like Google Text-to-Speech or Amazon Polly. These models use deep neural networks to produce natural and fluid speech synthesis in the target language. The user can then listen to the translation in real time in their chosen language.

3 Methodology

To achieve these objectives, the project will leverage state-of-the-art deep learning techniques for sign language recognition and generation. A critical component of this project is the selection of a suitable dataset that supports both French and English sign languages.

3.1 Dataset Selection Criteria:

The following criteria were considered to select the most appropriate dataset for our project:

- Year: More recent datasets are preferred as they incorporate the latest advancements in technology and methodology.
- Dataset Name: Identifying the dataset allows for further research and understanding of its applications and limitations.
- Country: The origin country provides insight into the cultural and linguistic context of the dataset.
- Number of Classes (CN): A higher number of classes indicates a richer diversity of signs and phrases. Each class represents a unique category of a sign or phrase. In a sign language dataset, a class corresponds to a specific sign or phrase, such as "hello," "thank you," or "how are you?". Multiple instances (videos or images) of different people performing the same sign or phrase are grouped under the same class. The number of classes therefore reflects the variety of distinct signs or phrases covered by the dataset. For example, a dataset with 2000 classes includes 2000 different signs or phrases, providing a comprehensive range for training and testing recognition and generation models.
  *example:
  Class 1: "Hello": Includes multiple videos or images of different individuals signing "hello."
- Number of Subjects (SubN): A diverse set of participants helps create more generalized and unbiased models.
- Number of Samples (SampN):Number of Samples (SampN): This refers to the total number of instances or examples in the dataset. Each sample is an individual occurrence or instance of a sign or phrase, typically represented as a video, image, or other data format. The number of samples indicates the volume of data available for training and testing models. For example, if a dataset contains 31,166 samples,

it means there are 31,166 videos or images of various signs or phrases performed by different individuals. A higher number of samples generally provides a more robust dataset for model training, as it captures more variations in sign execution and context.

- Total Duration: A longer duration of video content offers richer data for model training and validation.
- Language Level (LL): The level of language detail (words, phrases, sentences) must align with project requirements.
- Annotation (A): The types of annotations available (text, hand, face) are crucial for training accurate models.
- Target Language(s): The dataset must support the necessary languages, specifically French and English.

Comparative Table of Datasets for sign recognition (Part 1)

Criterion	Importance for the Project	How2Sign	Boston ASL LVD	DGS Kinect 40	RWTH-PHOENIX-Weather	GSL 20
Year	More recent is preferable	2020	2011	2012	2012	2012
Dataset	Knowing the name for additional research	How2Sign	Boston ASL LVD	DGS Kinect 40	RWTH-PHOENIX-Weather	GSL 20
Country	Cultural and linguistic relevance	USA	USA	Germany	Germany	Greece
Number of Classes (CN)	Diversity of signs and phrases	2000	3300	40	1200	20
Number of Subjects (SubN)	Diversity of participants	10	6	15	9	6
Number of Samples (SampN)	Volume of data for training	31166	9800	3000	45760	840
Total Duration	Richness of data	80 hours	Not specified	Not specified	Not specified	Not specified
Language Level (LL)	Relevance of detail level	Phrase (P)	Word (W)	Word (W)	Sentence (S)	Word (W)
Annotation	Types of annotations available	Face, Hand (F, H)	Hand (H)	–	Face, Hand (F, H)	–
Target Language(s)	Correspondence with the necessary languages (FR/EN)	English (needs to be completed for FR)	English	German	German	Greek
Feasibility for the Project	Evaluation of relevance for the needs of your project	Moderate (Complete for FR)	Moderate (English only)	Low (German)	Low (German)	Low (Greek)

Comparative Table of Datasets for sign recognition(Part 2)

Criterion	**Importance for the Project**	**ASL Fingerspelling A**	**ASL Fingerspelling B**	**LSA16 handshapes**	**PSL Fingerspelling ToF**
Year	More recent is preferable	2011	2011	2016	2015
Dataset	Knowing the name for additional research	ASL Fingerspelling A	ASL Fingerspelling B	LSA16 handshapes	PSL Fingerspelling ToF
Country	Cultural and linguistic relevance	USA	USA	Argentina	Poland
Number of Classes (CN)	Diversity of signs and phrases	24	24	16	16
Number of Subjects (SubN)	Diversity of participants	5	9	10	3
Number of Samples (SampN)	Volume of data for training	131000	–	800	960
Total Duration	Richness of data	Not specified	Not specified	Not specified	Not specified
Language Level (LL)	Relevance of detail level	–	–	–	–
Annotation	Types of annotations available	–	–	–	–
Target Language(s)	Correspondence with the necessary languages (FR/EN)	English	English	Spanish	Polish
Feasibility for the Project	Evaluation of relevance for the needs of your project	Moderate (English only)	Moderate (English only)	Low (Spanish)	Low (Polish)

3.2 Justification for Dataset Selection

Based on the comparative analysis, How2Sign emerges as the most suitable dataset for our project, provided we can supplement it with French annotations. It is the most recent dataset, offering a rich volume of diverse data with detailed annotations for face and hand movements. While it primarily supports English, its comprehensive structure makes it feasible to extend its use to French with additional efforts. Other datasets, while useful,

are limited by their focus on a single language or lack the necessary volume and diversity required for our bilingual sign language recognition and generation system.

3.3 Dataset Categories in How2Sign

The How2Sign dataset comprises several components, including RGB videos, keypoint clips, and English translations. Below is a table summarizing the storage requirements for each part of the dataset along with an explanation of each component and its role:

Dataset Component	**Training Size**	**Validation Size**	**Test Size**	**Total Size**
Green Screen RGB videos (frontal view)	290 GB	16 GB	23 GB	329 GB
Green Screen RGB videos (side view)	290 GB	16 GB	23 GB	329 GB
Green Screen RGB clips (frontal view)	31 GB	1.7 GB	2.2 GB	34.9 GB
Green Screen RGB clips (side view)	22 GB	1.2 GB	1.6 GB	24.8 GB
B-F-H 2D Keypoints clips (frontal view)	21 GB	1.2 GB	1.6 GB	23.8 GB
English Translation (original)	5.6 MB	311 KB	423 KB	0.006334 GB
English Translation (manually re-aligned)	5.6 MB	311 KB	424 KB	0.006335 GB

Table 1: Storage Requirements for How2Sign Dataset Components

- Green Screen RGB videos (frontal view): Full-length videos recorded with a green screen backdrop, capturing the signer from a frontal view. ** Provides a clear and unobstructed view of the sign language being performed, essential for training models to recognize signs accurately.

- Green Screen RGB videos (frontal view): Full-length videos recorded with a green screen backdrop, capturing the signer from a frontal view.
  * Provides a clear and unobstructed view of the sign language being performed, essential for training models to recognize signs accurately.

- Green Screen RGB videos (side view): Full-length videos recorded with a green screen but capturing the signer from a side angle.
  *Offers additional perspective and helps improve the model's understanding of the signs from different angles.

- Green Screen RGB clips (frontal view): Shorter clips extracted from the full-length frontal view videos.
  * Focuses on specific segments of the signing, making it easier to process and analyze individual signs or phrases.

- Green Screen RGB clips (side view): Shorter segments extracted from the side view videos.
  * Provides focused segments from the side view, adding another perspective to aid in sign recognition.

- B-F-H 2D Keypoints clips (frontal view) Clips with 2D keypoint annotations (Body, Face, and Hands) from the frontal view videos.
  * Keypoint data is essential for understanding the precise movements and positions of the signer's body parts, improving the accuracy of the model.

- English Translation (original) Original English translations of the signs and phrases.
  * Helps map the signed content to the corresponding text, aiding in the training of models to understand and translate sign language.

- English Translation (manually re-aligned) English translations that have been manually re-aligned for accuracy.
  * Ensures better synchronization with the signed content, improving the precision of model training.

3.4 Justification for Using All Categories

Using all categories provides a comprehensive understanding of the sign language movements, ensuring that the model captures both detailed hand/facial expressions (frontal view) and spatial movements (side view). The 3D keypoints offer additional spatial information that enhances the model's ability to generalize across different signers and signing styles.

4 System Architecture

In this section, we detail the methodological process followed to develop an integrated bidirectional translation system between sign language and text. This system is designed to work in real-time and be adaptable to environments with varying computational capacities. The methodology includes several key steps: feature extraction, preprocessing and encoding, decoding and post-processing, as well as multimodal translation.

4.1 Feature Extraction

Feature extraction is the first crucial step for translating sign language into text. It aims to capture relevant information from sign language videos necessary for accurate and faithful translation.

4.1.1 Using MediaPipe

Why use MediaPipe:

MediaPipe is used for the rapid extraction of spatial features from sign language videos. It is chosen for its ability to operate efficiently in real-time, even on low-capacity devices. MediaPipe allows for the detection and tracking of hand and joint landmarks on each frame of the video, which is essential for identifying hand and finger movements.

Method:
MediaPipe analyzes each frame of the video to extract the following features:

- **Hand Position:** Coordinates (x, y, z) of the hand and finger joints.
- **Finger Configuration:** Relative positions of the fingers to identify specific gestures.

These features are extracted in real-time, enabling quick and efficient analysis of simple gestures. For each frame, MediaPipe generates a set of feature vectors that represent the position of hand joints.

4.1.2 Using I3D (Inflated 3D ConvNet)

Why use I3D:
I3D (Inflated 3D ConvNet) is used for extracting rich spatiotemporal features from video segments identified as complex by MediaPipe. I3D is particularly effective in capturing both spatial and temporal information in videos, which is crucial for recognizing complex gestures that involve dynamic movements and subtle changes in gesture sequences.

Method:
When a video segment is identified as complex by MediaPipe (e.g., due to rapid movements or difficult-to-interpret finger configurations), I3D is activated to perform a more in-depth analysis:

- **Spatiotemporal Features:** I3D processes complete video segments to capture the dynamics of movements across multiple frames. This includes the speed, trajectory, and continuity of gestures, allowing for the recognition of longer and more complex gesture sequences.
- **Pre-training on Human Actions:** I3D is pre-trained on human action databases (such as Kinetics) to understand a wide range of movements, which enhances the model's ability to interpret varied gestures in sign language.

4.2 Preprocessing and Encoding

After feature extraction, the system must prepare this data for the Transformer model by undergoing preprocessing and encoding steps.

4.2.1 Preprocessing

Tokenization and Lowercasing:
The text resulting from speech recognition (or any input text) is tokenized and converted to lowercase to reduce vocabulary complexity and minimize capitalization variations. Tokenization transforms text into smaller units (tokens), facilitating processing by the Transformer model.

- **Tokenization:** Breaking down text into words or sub-words, thus transforming a sentence into a sequence of tokens.
- **Lowercasing:** Standardizing all characters to lowercase to reduce text variability.

4.2.2 Encoding with Transformers

Why use Transformers:

Transformer models are chosen for their ability to model long-term dependencies in data sequences, an essential feature for understanding syntax and context in sign language translation. Transformers use attention mechanisms to focus on relevant parts of input data, thereby enabling better understanding and translation of gestures into text.

Method:

- **Encoding Extracted Features:** The features extracted by MediaPipe and I3D are passed through the Transformer's encoder, which converts these feature sequences into a vector representation. This representation captures the necessary information about gestures and their context.
- **Multi-Head Attention:** The encoder uses multi-head attention mechanisms to capture different relationships and interactions between gestures in a sequence, allowing it to understand how each gesture contributes to the overall meaning.

4.3 Decoding and Post-processing

After encoding, the vector representations need to be converted into readable text through decoding and post-processing steps.

4.3.1 Decoding with Transformers

Why use the Transformer Decoder:

The Transformer decoder is used to convert encoded vector representations into text sequences. The decoder can generate natural language sentences based on the representations produced by the encoder while considering the overall context provided by attention mechanisms.

Method:

- **Text Generation:** The decoder generates text token by token, using the encoder's outputs and adjusting its predictions based on previously generated tokens.
- **Cross-Head Attention:** The decoder also uses cross-head attention mechanisms to focus on relevant parts of the encoded representations, thereby improving translation accuracy.

4.3.2 Post-processing

Detokenization and Truecasing:

Once the decoder has generated the text sequence, the system performs post-processing steps to improve the readability of the final text.

- **Detokenization:** Transforming tokens into smooth and readable text, eliminating the separations added during tokenization.
- **Truecasing:** Restoring correct text capitalization to reflect grammatical norms and improve readability.

4.4 Sign Language Synthesis (Text-to-Sign)

Why use Sign Language Synthesis:

Sign language synthesis enables the translation of text into sign language videos, which is crucial for making textual content accessible to deaf or hard-of-hearing individuals. This adds a multimodal dimension to the system, thereby increasing its versatility and usefulness.

Method:

- **Animation of Virtual Avatars:** Utilization of 3D avatar models to perform signs based on the input text. This allows for the generation of dynamic and expressive sign language videos.
- **Pre-recorded Sign Videos:** Selection of individual sign video clips from a video library and combining these clips to form complete sentences.

Summary

The methodology described above is designed to maximize the efficiency, accuracy, and flexibility of the translation system for converting sign language to text and text to sign language. Each component is chosen for its ability to effectively process input data and produce accurate outputs, while also being adaptable to different performance and capacity requirements of various devices.