Automatic Speech to American Sign Language Video Generation

Team 4

Nikitha Bramadi, Razan Dababo, Priya Khandelwal, Sadakhya Narnur

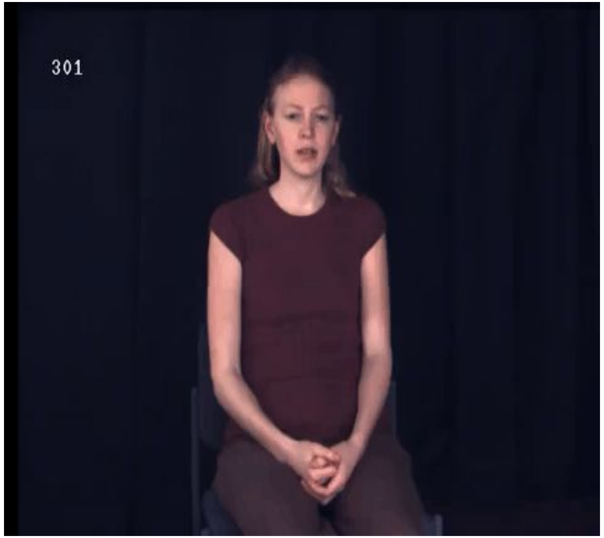
Introduction & Motivation

- Sign language is essential for communicating with the hard-of-hearing or **hearing-impaired**, who make up **5% of the global population**.

- While significant strides have been made in language recognition (SLR), sign language production **(SLP) lags behind**.

- ***Bridge communication gap***: A system that translates spoken language to American sign language.

- *Proposed System*: An approach to SLP using Transformer and GAN models to produce **realistic sign language videos** from speech input.



Source: http://bu.edu/av/asllrp/dai-asllvd.html

System Requirements

Software Requirements:

- Language : Python, Javascript, Html
- Libraries: PyTorch, Tensorflow, DWPose, ngrok
- Framework : Flask
- Google Speech to Text API
- Google Colab
- AWS Sagemaker, Lambda Functions, API Gateway
- Storage tools: AWS S3 Bucket, Google drive

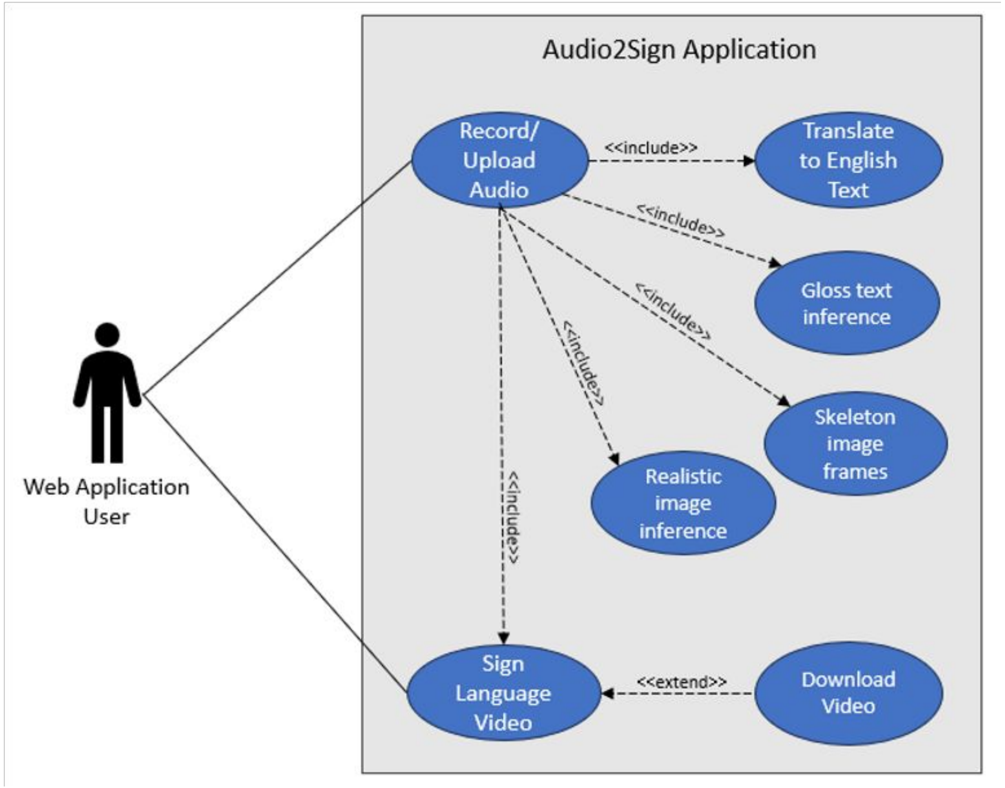
Hardware Requirements :

- Graphics Processing Unit (GPU)
- Memory (RAM)
- Microphone

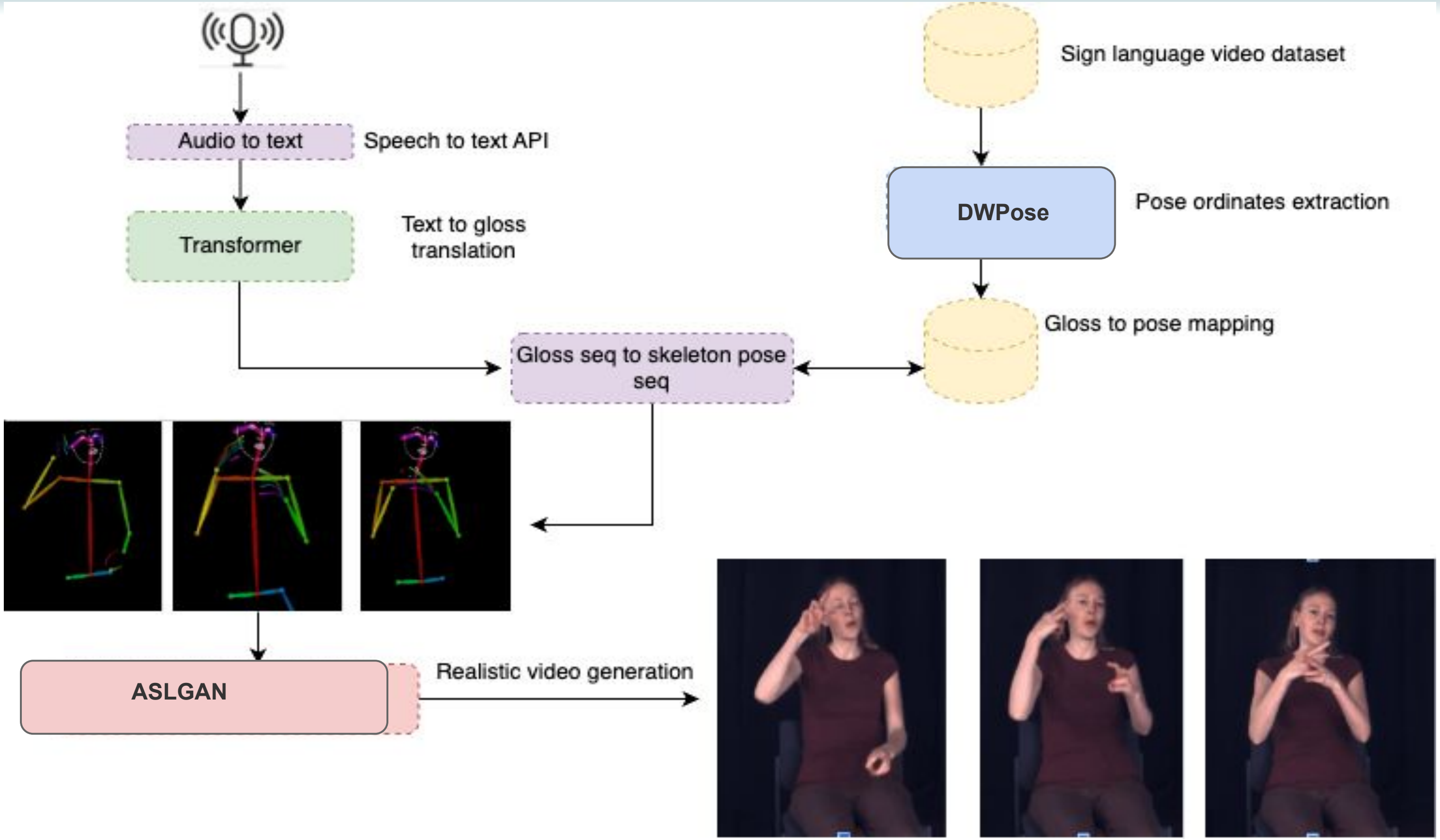
System Requirements

Functional Requirements:

- Microphone Accessibility
- Upload File Accessibility
- Generate recorded Text
- Submit query
- Generate Sign lang Video
- Compatibility with operating systems such as Windows, macOS, or Linux



Pipeline Architecture



Audio to text
Speech to text API
Transformer
Text to gloss translation
Sign language video dataset
DWPose
Pose ordinates extraction
Gloss to pose mapping
Gloss seq to skeleton pose seq
ASLGAN
Realistic video generation

End to End Data Conversion

Spoken Language

The boy likes to read the book that is gifted by his old friend

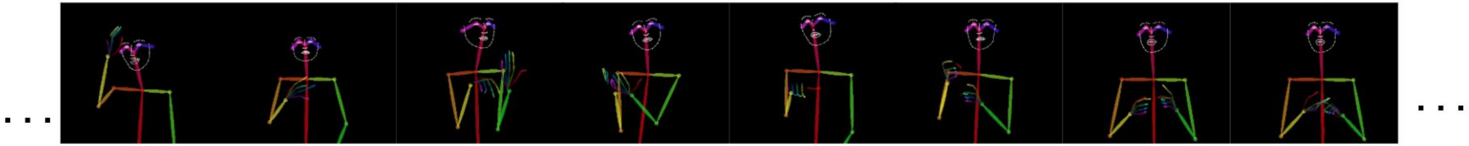
↓ Text to Gloss

Sign Language Glosses

BOY | LIKE | READ | BOOK | THAT | GIFT | OLD | FRIEND

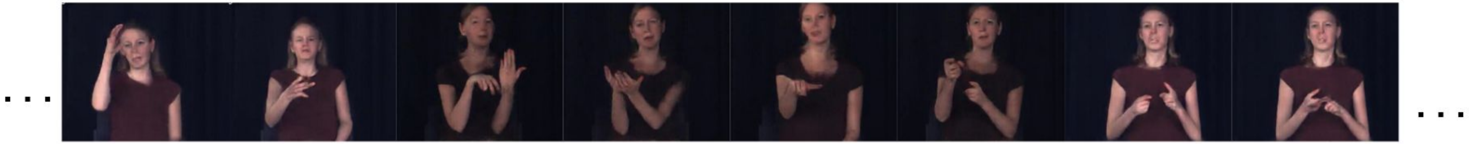
↓ Gloss to Pose

Pose Sequence Frames



↓ Pose to Sign

Sign Language Video



Datasets

ASLG-PC12 - text to gloss dataset

- A large parallel corpus of English written texts and American Sign Language glosses
- More than 80,000 pairs of sentences
- Open access for research
- Latest release in 2012

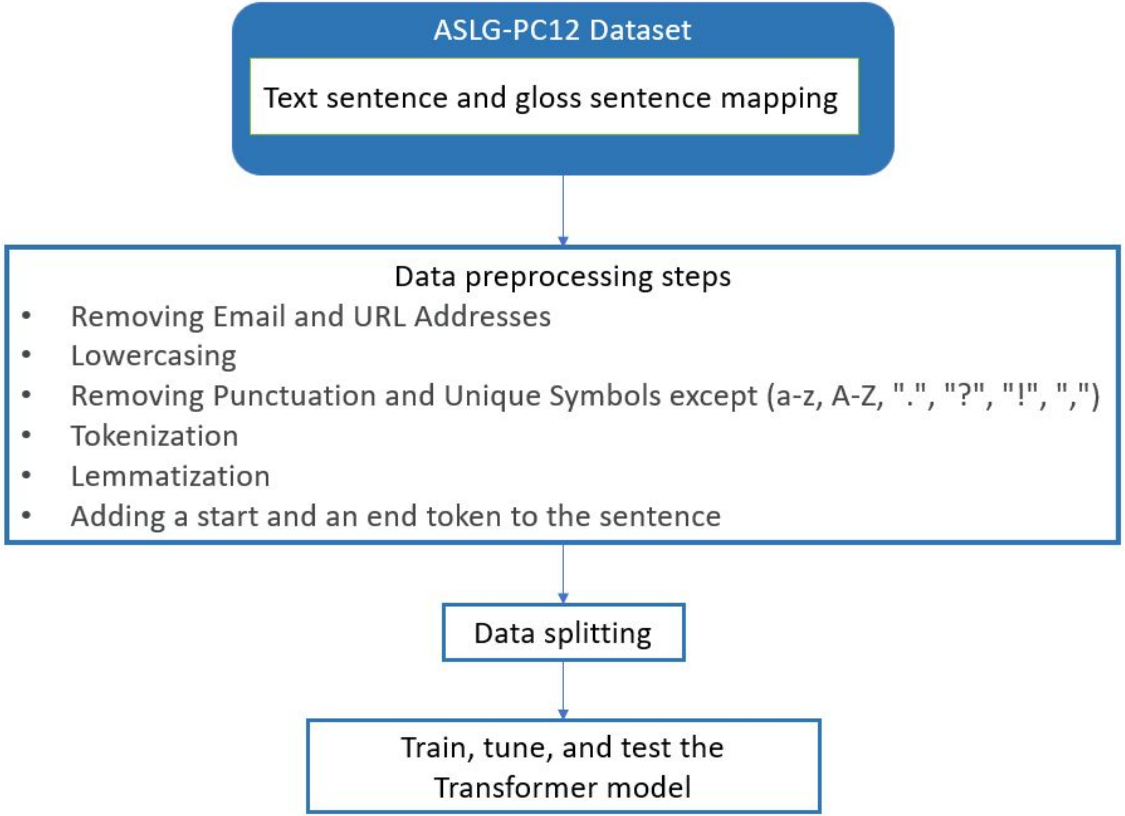
ASLLVD - gloss to sign dataset

- American Sign Language Lexicon Video Dataset
- Created by capturing videos of six signers with four synchronized cameras
- Videos represented for 9800 gloss tokens in more than 3300 video clips
- Video metadata file to map from gloss to video clip using scene id and sessionid along with start and end frame number of the video clip
- Videos accessible in 3 camera angles each of the resolution 640 x 480 and 60 fps
- Open to access for research

Data Preprocessing - Text2Gloss

ASLG-PC12

Text-to-Gloss Dataset



ASLG-PC12 Dataset

Text sentence and gloss sentence mapping

Data preprocessing steps

- Removing Email and URL Addresses
- Lowercasing
- Removing Punctuation and Unique Symbols except (a-z, A-Z, ".", "?", "!", ",")
- Tokenization
- Lemmatization
- Adding a start and an end token to the sentence

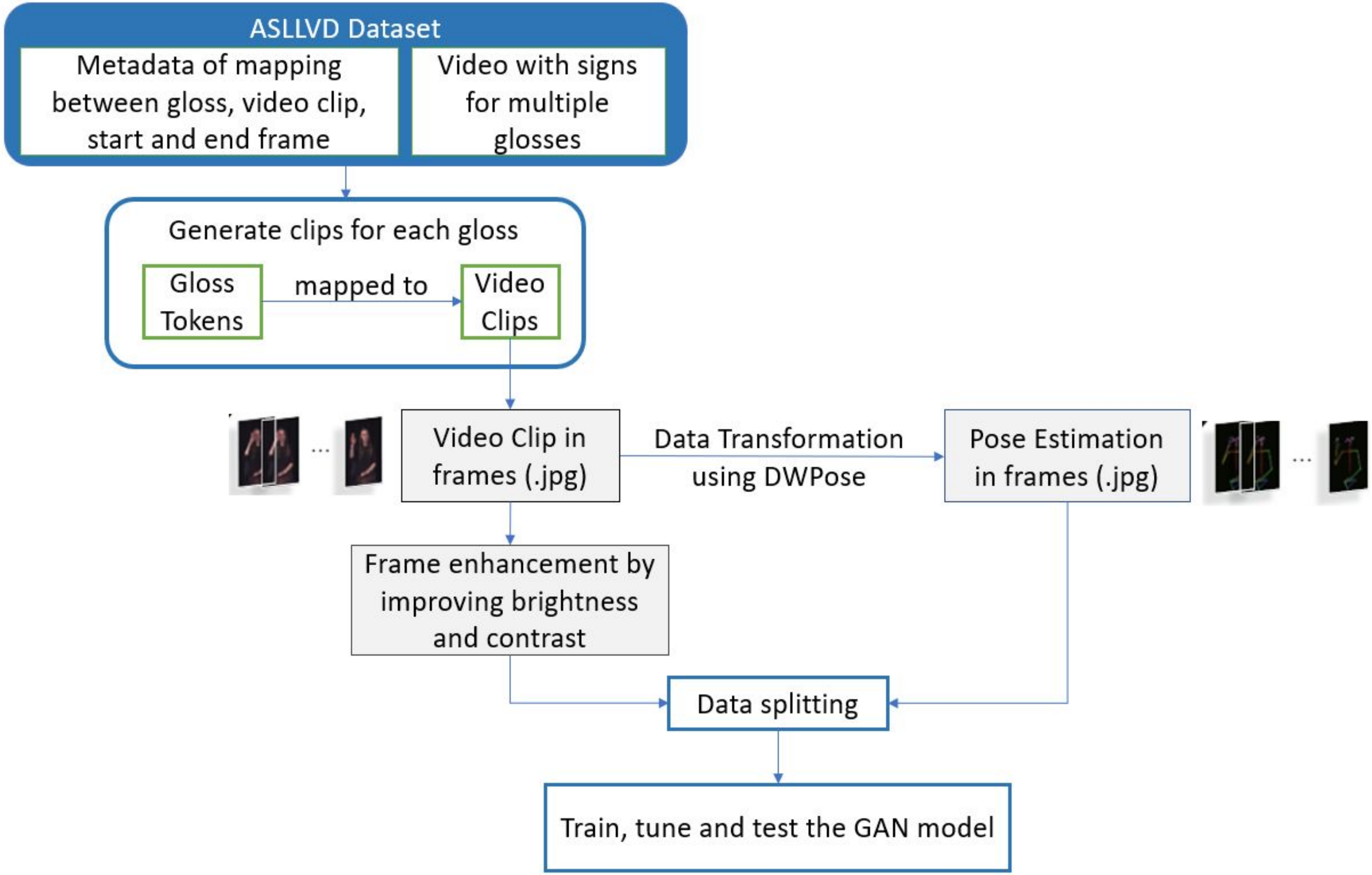
Data splitting

Train, tune, and test the Transformer model

Data Preprocessing - Gloss2Pose

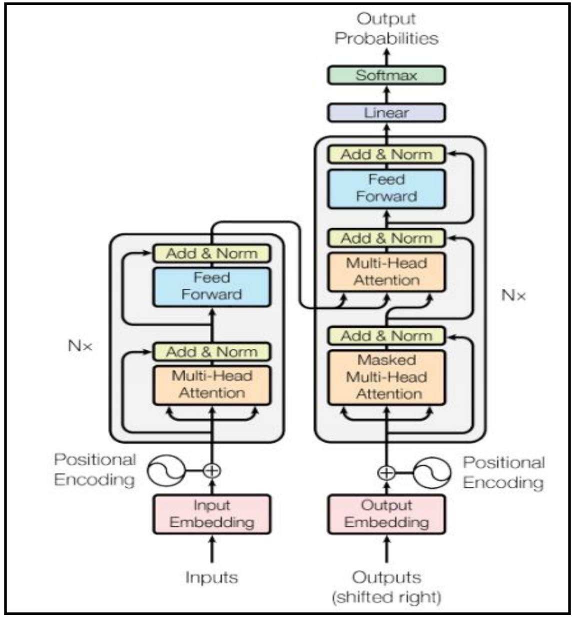
ASLLVD

Gloss-to-Pose Dataset

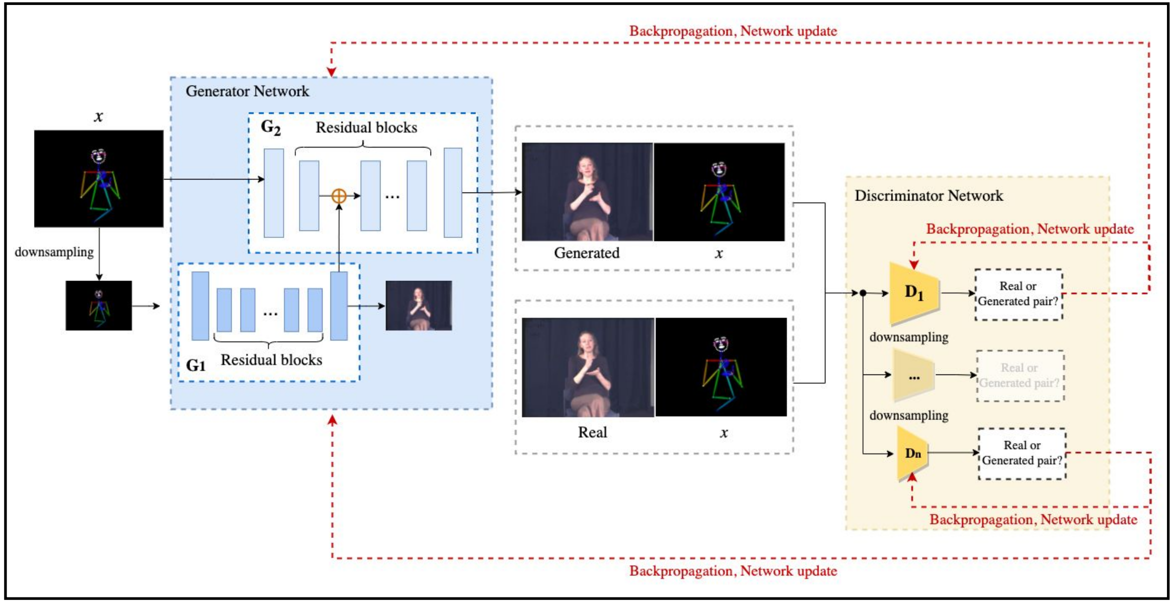

ASLLVD Dataset
Metadata of mapping between gloss, video clip, start and end frame
Video with signs for multiple glosses
Generate clips for each gloss
Gloss Tokens
mapped to
Video Clips
...
Video Clip in frames (.jpg)
Data Transformation using DWPose
Pose Estimation in frames (.jpg)
...
Frame enhancement by improving brightness and contrast
Data splitting
Train, tune and test the GAN model

Models

Transformer



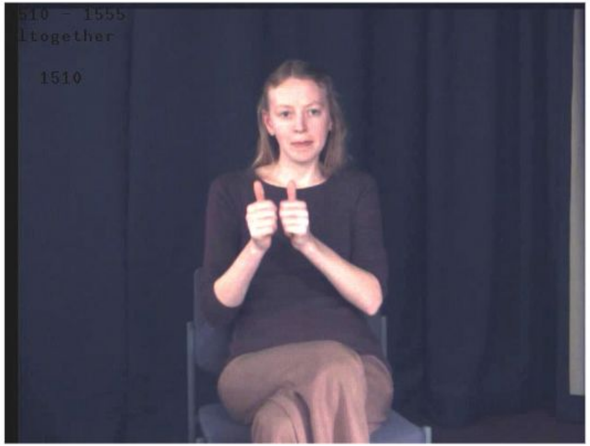
ASLGAN



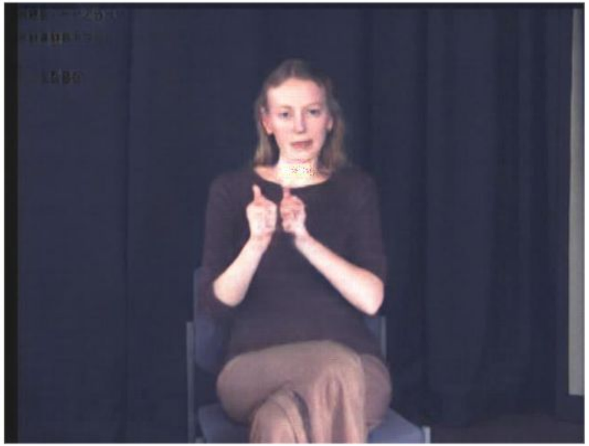
https://www.tensorflw.org/text/tutorials/transforer

Model Evaluation

Model	Dataset	Evaluation
Transformer	ASLG_PC12	**61.23 BLEU**
ASLGAN + 5 Discriminators	ASLLVD	**0.931 SSIM**

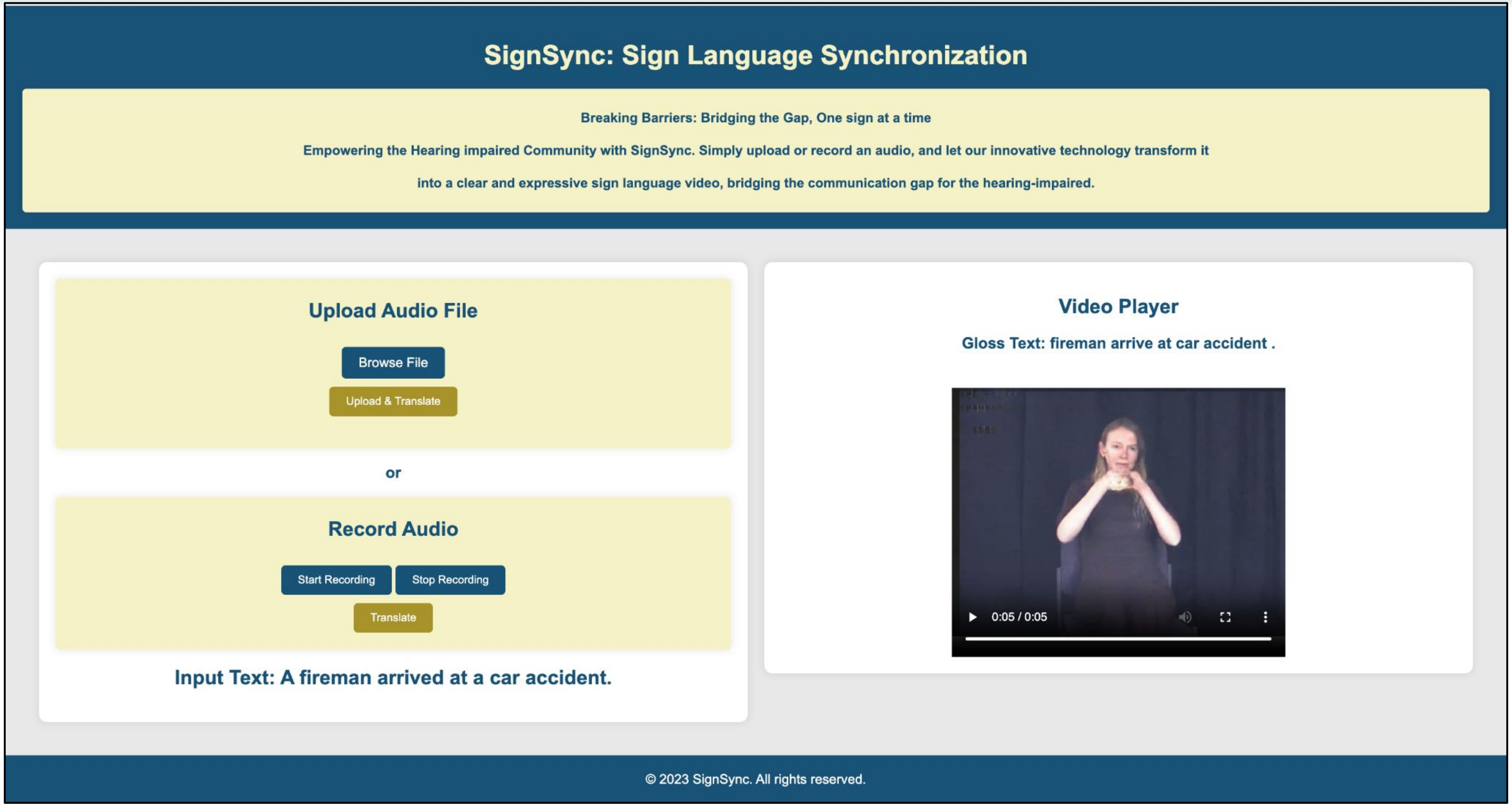


Real Image



Generated Image

System Design - Web UI Design



SignSync: Sign Language Synchronization

Breaking Barriers: Bridging the Gap, One sign at a time

Empowering the Hearing impaired Community with SignSync. Simply upload or record an audio, and let our innovative technology transform it into a clear and expressive sign language video, bridging the communication gap for the hearing-impaired.

Upload Audio File

or

Record Audio

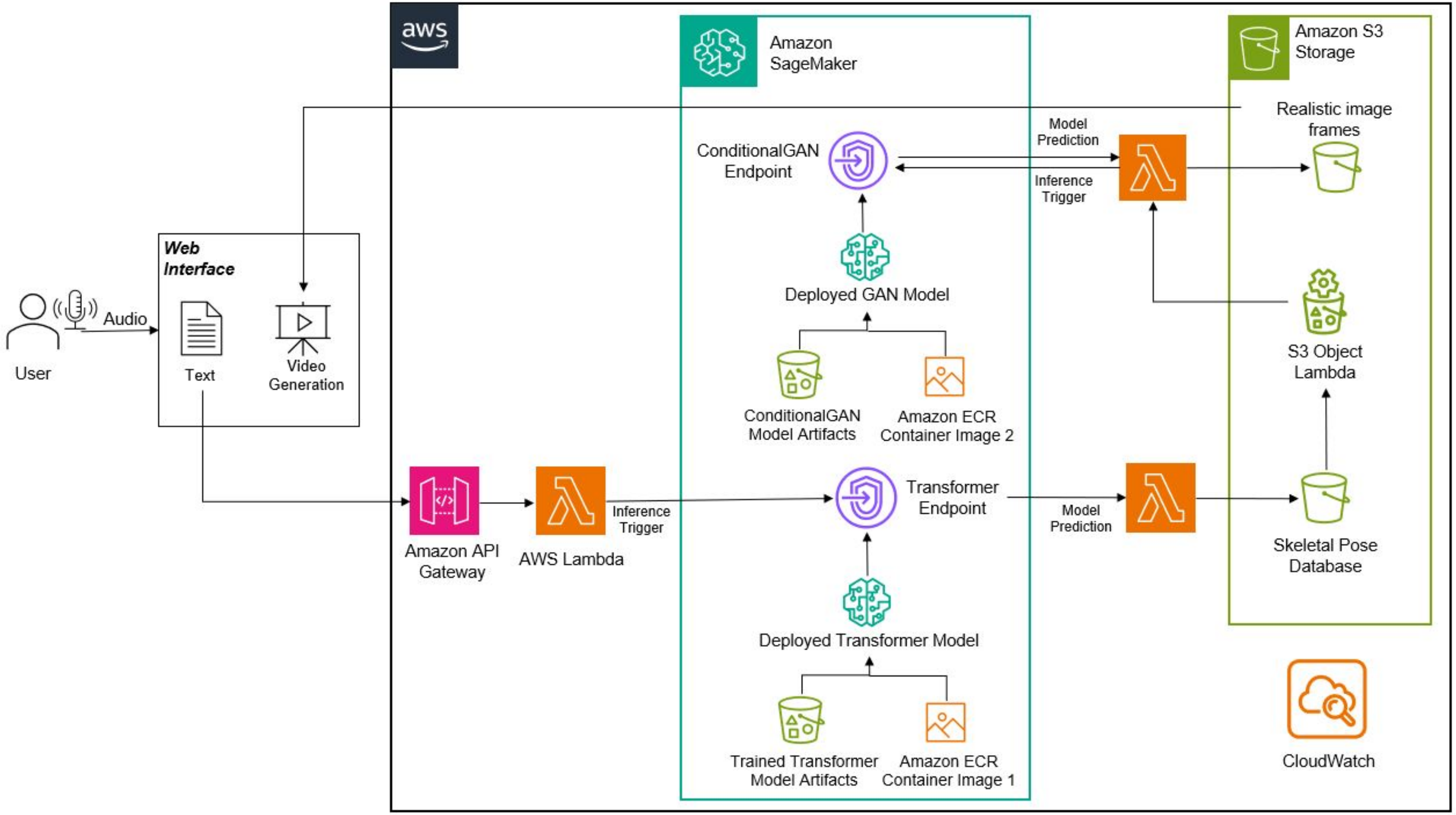
Input Text: A fireman arrived at a car accident.

Video Player

Gloss Text: fireman arrive at car accident .

© 2023 SignSync. All rights reserved.

System Deployment Design



Achievements

Model Achievements:

- Developed a Transformer model from scratch that achieved state-of-the-art results translation results.
- Successfully generated high-quality videos using our proposed system.
- Pioneered the usage of CycleGAN for image generation from poses, reducing the need for extensive paired datasets.
- Integration of DWPose for pose extraction improved skeleton pose quality and inference speed with ASLGAN.

Sign Language Production Achievements:

- Project contributes an easy-to-understand yet foundational approach to sign language video production domain.
- Enhanced the accuracy and efficiency of the deep learning models that drive support to the hard of hearing community.

Constraints

Model Development Constraints

- Model's training was limited to a single signer.
- Hand keypoints need further enhancement.

Dataset Constraints

- Limited vocabulary in available datasets.
- Lack of high-resolution videos.
- Alternative datasets not feasible due to lack of gloss annotations.

System Constraints

- Deep Learning models require high GPU power for training and inference.
- Not possible to deploy SignSync on a single Amazon SageMaker endpoint due to differences in model frameworks (TensorFlow vs PyTorch).

Lessons Learned

Academic Research

- Systematic progress is the only solution to a complex problem.
- Keeping up with a progress made in GenAI (or any fast evolving domain) is essential to publish a promising paper.

Development and Deployment

- Preplanned continuous training and deployment on cloud service.
- Learn to estimate the resource allocation when dealing with high computational deep learning models.

Potential System and Model Applications

On-the-Go Communication

- Practical application for individuals with hearing difficulties in daily life.
- Supports communication needs in various settings, both indoors and outdoors.

Educational Integration

- Seamless integration into learning settings for real-time sign language interpretation.
- Enhances accessibility in online tutorials and courses.

Public Areas and Accessibility

- Integration into public spaces like government buildings and transit hubs.
- Enables real-time sign language interpretation via public address systems.

Contributions and Impacts on Society

- Addresses crucial issues in **communication accessibility** for individuals with hearing impairments.
- **Empowers users** to participate autonomously in various activities.
- Acts as a valuable tool for **instructional purposes**, facilitating easier learning of sign language.
- **Real-time translation** capabilities enhance the immediacy and effectiveness of sign language communication.
- Demonstrates the potential of **responsible AI** in assistive technology development.
- Can potentially curb the **malpractices** with suitable long-term enhancements.

Future Work

- Utilize **Computer Vision** and center the human figure on the GAN generated image and position the human in the center to avoid positional variations among different frames.
- Use of a **Stable Diffusion model** for generating the human-like images which leverages the choice of selecting the signer gender and other specificities as a prompt input.
- Evaluate the possibility of training GAN model on **Multiple Signers**.
- Include a **Transition Mechanism** while combining two different frames to generate a final video to have a smoother transition between poses generated.

Demo

Time to Demonstrate!