A Voice Tool for Deaf and Hearing Communities

1. **Keerthivasan S**

Pg & Research Department of Computer Science,

Sri Ramakrishna College Of Arts & Science,

skeerthivasan1109@gmail.com

2. **DR. N. Mahendiran,**

Assistant Professor,

Pg & Research Department of Computer Science,

Sri Ramakrishna College of Arts & Science,

mahendiran@srcas.ac.in

Abstract

paper presents WLASL (Word-level American Sign Language) as a dataset for building a voice to sign language translator that translates spoken English into ASL (American Sign Language) through video. The system integrates ASR (Automatic Speech Recognition) to transcribe (convert) real-time speech inputs into text. This transcription process identifies the respective keywords from an input sentence, and these keywords are then mapped to ASL glosses. Sign language videos of the respective ASL signs are streamed to the user through a web-based chatbot interface. The purpose of this digital communication tool is to provide greater accessibility for Deaf and hard of hearing individuals by improving access to various types of voice-to-sign communication systems. In developing the WLASL translator sistema we incorporate scalable architecture, real-time performance, and efficient retrieval of video data without the need for storing output videos on the server. This research informs the development of AI-based assistive technologies that have the potential to connect hearing and non-hearing communities.

Keywords:

Voice-to-Sign Translation, WLASL Dataset, American Sign Language (ASL), Automatic Speech Recognition (ASR), Assistive Technology, Real-Time Video Streaming, Accessibility.

Introduction

The gap between hearing and deaf people continues to be an obstacle that society needs to address. Most deaf people use sign language as their main form of communication, while hearing persons often know little about sign language.

Most current tools provide only text-to-sign translation, with few tools available to allow hearing persons to convert spoken languages into sign language in real-time.

This project will use WLASL to create a real-time voice-to-sign language translator that will enable deaf and hearing individuals to communicate more easily using artificial intelligence/multimedia techniques.

Objective

The goal of the project is to create & implement a real-time voice-to-sign language translation tool that converts spoken English to American Sign Language gestures and to develop AP based speech recognition software that captures spoken input and converts it into text with minimal latency. Also, to process the converted text into usable ASL glosses - converting all speech into transcription with no extra latency. The project will incorporate dynamic video streaming through an easy-to-use web-based interface for users & provide an interactive user experience. A critical component of the project will be to allow for scalability, quick retrieval of video, and minimal performance lag, while still providing accurate output data, all without storing data on the web. The project will also enhance the accessibility of deaf/hard of hearing individuals by improving communication between all communities and encouraging inclusive and diverse participation between deaf/hard of hearing and hearing communities, through the use of AI-based assistive technology.

Existing System

Current systems for communicating in sign language are primarily based on static programmed translations (text-to-speech) and pre-recorded videos of someone signing, and do not support live audio/video integration (same time). Current methods of communication use human sign language interpreters, which can be expensive and not always available. Some text-to-sign applications offer animated sign language from written form, however they still require a user to type in the text they want to sign instead of allowing for capturing live speech. In addition, most current systems utilize a still image sequence or an avatar (moving picture) which may not convey natural expressions and/or be accurate in representation of what is being said.

Many previous studies focused on gesture recognition (sign-to-text) rather than on converting voice signals to sign language, which limits any support for only one-way phone or text-based communication. Other platforms store processed data on central servers, limiting user privacy as well as scalability of the service. As a result of these limitations, current systems experience decreased system efficiencies because of limitations attributed to vocabulary coverage (limited words available) and the time to retrieve the video (slow). Overall, as a result of these limitations, current solutions lack the ability to process spoken language into text-based responses or using video or dynamic streaming of video in real-time (at same time) in a scalable architecture. This indicates that there is a need for an enhanced artifical intelligence-based (AI) speech to sign language conversion system to achieve seamless (non-stop) voice to sign language translation.

Proposed system

- Speech capture module: the user’s voice is collected through a microphone
- Speech recognition module: converts the audio into recognizable text by using automatic speech recognition tools like the Google Speech API and Vosk
- Text processing module: extracts the most relevant keywords from the text that was created in the speech recognition module
- Gloss mapping module: uses mapped identifier values to identify each keyword using WLASL notation
- Video streaming module: will dynamically stream to the user ASL (American Sign Language) videos that correlate with the keywords extracted from their speech through the speech capture and recognition modules
- User interface module: uses a chatbot style web page to allow the user to view the videos they are receiving

The system will have little to no time delay between when the user’s voice is captured and the time that the video is streamed, and will not permanently store any of the video that it streams to the user.

Methodology

structured approach is used to translate voice to sign in the proposed system. A user's speech will be captured by a microphone using an Automatic Speech Recognition (ASR) engine that will convert the audio into written text. The written text will then be pre-processed using natural language techniques (e.g., tokenisation and the extraction of keywords). The extracted keywords will be matched to gloss entries within the WLASL database. Based on the gloss ID identified, relevant ASL video clips are retrieved dynamically and played in sequence through a web interface. The proposed solution provides real-time performance and effective retrieval of video clips to provide seamless ASL communication, providing no storage of output data on the web server.

System Design Overview

The Voice-to-Sign Translator System consists of five components: Voice Input, ASR Module (for Speech Recognition), Text Processing Module, Gloss Mapping Module, and Video streaming Interface. Users speak into a microphone that captures real-time English as voice input. The ASR module processes and recognizes the user’s speech and converts the audio signal into text. These keywords are then processed by the Gloss Mapping module to match each keyword with its corresponding ASL gloss in the WLASL dataset.

Following the matching process, the video streaming component retrieves the appropriate ASL video clip. The overall system will display videos to the user via web-based Chatbot, with instantaneous video streaming to the user’s interaction. The overall system architecture is designed to ensure low latency and scalability while allowing for efficient performance without storing user data, thereby maintaining user privacy and security.

Literature survey

Previous research has explored the use of assistive communication aids utilizing the following methods:

- Animated avatars providing video translation from text to sign language.
- Communication aids based on speech recognition.
- Gesture-based communication aids utilizing deep learning models that support recognition of gestures and signs.

Nonetheless, most past research lacked features such as real-time streaming of their data and used limited size datasets, as well as produced unrealistic video output of American Sign Language (ASL). The WLASL dataset represents a more realistic and extensive alternative to current avatar based translation solutions.

Advantages

- Enhanced Access to Communication - Supports successful interaction between individuals with hearing impairments and people who have difficulty communicating verbally (i.e., Deaf).
- Instantaneous ASL Translations - ASR converts oral English to ASL automatically in near real-time.
- Integrative Chat Bot - Based on a web chat bot design to be as user-friendly as possible.
- Scalable Architecture - Can be scalable in nature such that it is able to accommodate both multiple users and larger volumes of data.
- Instant Retrieval of ASL Videos - Allows for streaming of relevant ASL video dynamically, thereby allowing for video to be streamed over the Internet without using excessive server storage.
- AI Powered Automation - Reduces the amount of human interpretation requirements in simple scenarios.
- Cost Effective Solution - Provides an infrastructure for elevating WLASL (Web-Based American Sign Language) from current provide datasets (WLASL) and open access ASR tools to support the integration of ASL translation into a variety of applications.

Conclusion

Voice-To-Sign Language Translator, which utilizes the WLASL dataset, was developed to assist in reducing the communication gap between the hearing population and the Deaf community. The use of speech recognition, key-word extraction, the development of mapping glosses and the provision of a real-time streaming video capability for the translation system allows for efficient, effective and immediate translation of spoken language into American Sign Language (ASL) signs.

References

1. Schembri T. C. and Gleeson J. S. and Brentari D. (2020), “Dataset for Recognition/Translation of American Sign Language at the Level of the Word”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
2. Graves A., Fernández S., Gomez F., and Schmidhuber J. (2006), “Connectionist Temporal Classification: Using Recurrent Neural Networks to Label Unsegmented Sequence Data”, Proceedings of the International Conference on Machine Learning.
3. Lee D. H. and Park K. H. (2018), “Using the Google Speech API for Real-time Speech Recognition”, International Journal of Computer Applications, vol. 182, no. 23, pp. 15-20.
4. Vosk (2020), “Offline Speech Recognition Toolkit", Alpha Cephei LLC, available online at https://alphacephei.com/vosk/).
5. Goodfellow I., Bengio Y., and Courville A. (2016), Deep Learning, Cambridge, MA: MIT Press.
6. Starner T., and Pentland A. (1995), “Real-time Recognition of American Sign Language from Video using Hidden Markov Models”, Proceedings of the International Symposium on Computer Vision.
7. OpenCV, "Open Source Computer Vision Library", (2023), Available online at: https://opencv.org/.
8. FastAPI, “FastAPI: High-performance web framework for Building APIs in Python”, (2023), Available online at: https://fastapi.tiangolo.com/#/).