Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777443

REAL-TIME VOICE TRANSLATION IN VIDEOS

RajBharath[a]*, V. Mahalakshmi[b], S.R.Rajeshwari[b], S. Hemamalini[b]

[a] Associate Professor, Department of Computer Science Engineering, Manakula Vinayagar Institute of Technology,Puducherry; [b] UG Scholar, Manakula Vinayagar Institute of Technology, Puducherry.

Abstract—Real-time Voice Translation is an AI-powered technology that can translate speech from one language to another. Voice translation consists of three processes namely speech recognition, machine translation and speech synthesis. Speech recognition is a machine's ability to recognize words spoken aloud and convert them into readable text. Machine translation converts the text of one language to another of the user's choice. Speech synthesis acts as a text-to-speech translator, generating an automated replication of human speech. Voice translation application works by integrating all these three processes and gives the best output to the user. The project aims to develop a real-time language translation system using the Python programming language. First, the audio is extracted from the video using Python's Moviepy library. The speech recognition module helps convert source audio to text and uses the Hidden Markov Model algorithm. The source text is converted to text in the target language using the GoogleTrans API. Google Text to Speech (GTTS) helps with speech synthesis. The generated target audio is then merged with the original video to get the final output. Real-time language translation is a promising technology that can break down language barriers and improve communication between people from different language backgrounds.

Index Terms—.text-to-speech translator, Google Text to Speech (GTTS), GoogleTrans API, Python's Moviepy library.

I. INTRODUCTION

An increasing number of live events are nowadays being live streamed on video channels and social networks. These events are transmitted in real-time to a large audience, on all types of devices and anywhere in the world. Captioning and live translation are seen as essential in order to ensure that these events reach a growing international audience. Real-time translation is the technology that can help you translate one language to another instantly. A real-time voice translator that can translate voice input and give translated voice output generated from it. Voice Translation, also referred to as speech-to-speech translation, is software that can translate human speech from one language into another. Once thought to be the work of science fiction (such as the Babel Fish from the Hitchhiker's Guide to the Galaxy or the Universal Translator from the Star Trek universe), modern-day Voice Translation technology relies on AI, neural processing, natural language understanding, speech recognition, and text-to-speech conversion. As of 2021, several companies offer voice translation technology through various formats, most commonly with apps, hand-held devices, or in-ear or over-the-ear headphones. Our project aims at converting a video in the English language to the destination language selected by the user

An increasing number of live events are nowadays being live streamed on video channels and social networks. These events are transmitted in real-time to a large audience, on all types of devices and anywhere in the world Captioning and live translation is seen as essential in order to ensure that these events reach a growing international audience Real-time translation is the technology that can help you translate one language to another instantly A real-time voice translator that can translate voice input and give translated voice output generated from it. may provide users with immediate feedback on their lip motions, which they can utilize to gradually enhance their speech and communication abilities.

Speech Translation has always been about giving source text/audio input and waiting for the system to give translated output in the desired form. In this paper, we present the Acoustic Dialect Decoder (ADD) – a voice-to-voice ear-piece translation device. We introduce and survey the recent advances made in the field of Speech Engineering, to employ in ADD, particularly focusing on the three major processing steps of Recognition, Translation and Synthesis. We tackle the problem of machine understanding of natural language by designing a recognition unit for source audio to text, a translation unit for source language text to target language text, and a synthesis unit for target language text to target language speech.

Artificial Intelligence:

Artificial intelligence allows machines to model, and even improve upon, the capabilities of the human mind o Artificial speech translation is a rapidly emerging artificial intelligence (AI) technology o Initially created to aid communication among people who speak different languages, this speech to-speech translation technology (S2ST) has found its way into several domains

As Artificial Intelligence is used in real-time voice translation, the projected turnaround time for a translation is currently less than 2 to 5 seconds. Due to the fact that many translation technologies rely on cloud-based data, there is a gap between voice and translation. When network speeds increase, this will probably rise. On October 21, 2020, Alibaba presented the first e-commerce live stream with real-time translation in 214 2 languages. Facebook employs artificial intelligence on social media platforms like Messenger and Instagram, resulting in more than 6 billion translations daily. Google Translate leads the internet translation market. Google Translate is a multilingual neural machine translation tool that enables users to translate text, documents, and web pages between languages. By the end of 2021, it will serve 109 languages at varying levels. As of April 2016, there were more than 500 million users who were translating more than 100 billion words daily. Google has produced items with real-time language translation capabilities, such as Bluetooth earbuds and the Pixel Buds. Real-time translation of a discussion between two speakers of different languages is possible with this feature. Additionally, Google just announced that this real-time translation will be accessible to everyone on all Google Assistant-enabled smartphones and headphones.The S2ST model, which can translate speech between two languages directly without the necessity of numerous intermediary subsystems, is the second iteration of Translatotron that Google AI has produced. In 2019, the Google AI team unveiled Translatotron, a voice-to-speech translation model that it claimed was the first end-to-end framework to directly translate speech from one language into speech in another.

Natural Language Processing:

The field of computer science known as "natural language processing" (NLP) is more particularly the field of "artificial intelligence" (AI) that is concerned with providing computers with the capacity to comprehend written and spoken words in a manner similar to that of humans. It is extremely challenging to create software that reliably ascertains the intended meaning of text or voice data since human language is rife with ambiguity. Homonyms, homophones, sarcasm, idioms, metaphors, exceptions to the rules of grammar and usage, and changes in sentence structure are just a few examples of the irregularities in human language that take humans years to learn but that programmer must teach natural language-driven applications to recognise and understand accurately from the beginning if those applications are to be useful. In order to help the computer understand the text and speech data it is absorbing, several NLP activities deconstruct human text and voice data. These are only a few of these jobs: 3 The process of accurately translating voice data into text is known as speech recognition, commonly referred to as speech-to-text. Any programme that responds to voice commands or questions must use speech recognition. The way individuals speak—quickly, slurring words together, with varied emphasis and intonation, in various dialects, and frequently using improper grammar—makes speech recognition particularly difficult. The act of identifying a word's part of speech based on its use and context is known as part of speech tagging, also known as grammatical tagging. In the sentences "I can create a paper plane" and "What make of car do you own?" the word "make" is classified as a verb and a noun, respectively. Word sense disambiguation is the act of choosing a word's meaning from among its possible meanings using semantic analysis to discover which word makes the most sense in the context at hand. Word sense disambiguation, for instance, clarifies the difference between the meanings of the verbs "make" and "make the grade" (achieve) and "make a bet" (place). Words or phrases are recognised as useful entities using named entity recognition, or NEM. NEM identifies "Kentucky" as a place or "Fred" as the name of a guy. The task of determining whether and when two words refer to the same item is known as co-reference resolution. The most typical example is figuring out who or what a certain pronoun refers to (e.g., "she" = "Mary"), but it can also require figuring out a metaphor or idiom that is used in the text (e.g., when "bear" refers to a big, hairy person rather than an animal). Sentiment analysis looks for intangible elements in the text, such as attitudes, feelings, sarcasm, bewilderment, and mistrust. Natural language generation is the process of converting structured data into human language; it is frequently referred to as the opposite of voice recognition or speech-to-text.

II. LITERATURE REVIEW

The field of computer science known as "natural language processing" (NLP) is more particularly the field of "artificial intelligence" (AI) that is concerned with providing computers with the capacity to comprehend written and spoken words in a manner similar to that of humans. It is extremely challenging to create software that reliably ascertains the intended meaning of text or voice data since human language is rife with ambiguity. Homonyms, homophones, sarcasm, idioms, metaphors, exceptions to the rules of grammar and usage, and changes in sentence structure are just a few examples of the irregularities in human language that take humans years to learn but that programmer must teach natural language-driven applications to recognise and understand accurately from the beginning if those applications are to be useful. In order to help the computer understand the text and speech data it is absorbing, several NLP activities deconstruct human text and voice data. These are only a few of these jobs: 3 The process of accurately translating voice data into text is known as speech recognition, commonly referred to as speech-to-text. Any programme that responds to voice commands or questions must use speech recognition. The way individuals speak—quickly, slurring words together, with varied emphasis and intonation, in various dialects, and frequently using improper grammar—makes speech recognition particularly difficult. The act of identifying a word's part of speech based on its use and context is known as part of speech tagging, also known as grammatical tagging. In the sentences "I can create a paper plane" and "What make of car do you own?," the word "make" is classified as a verb and a noun, respectively. Word sense disambiguation is the act of choosing a word's meaning from among its possible meanings using semantic analysis to discover which word makes the most sense in the context at hand. Word sense disambiguation, for instance, clarifies the difference between the meanings of the verbs "make" and "make the grade" (achieve) and "make a bet" (place). Words or phrases are recognised as useful entities using named entity recognition, or NEM. NEM identifies "Kentucky" as a place or "Fred" as the name of a guy. The task of determining whether and when two words refer to the same item is known as co-reference resolution. The most typical example is figuring out who or what a certain pronoun refers to (e.g., "she" = "Mary"), but it can also require figuring out a metaphor or idiom that is used in the text (e.g., when "bear" refers to a big, hairy person rather than an animal). Sentiment analysis looks for intangible elements in the text, such as attitudes, feelings, sarcasm, bewilderment, and mistrust. Natural language generation is the process of converting structured data into human language; it is frequently referred to as the opposite of voice recognition or speech-to-text.

The purpose of this paper is to present a learning method for creating language translation rules from multilingual text samples. The languages involved are controlled languages. They are domain-specific sublanguages that are disambiguated by restricting vocabulary and syntax. The learning method presented here allows for supervised, human-assisted learning of generalized translation rules, making it faster and easier to adapt machine translation systems to new languages. Sentence segmentation deals with the problem of finding the end of a sentence. That is, separate the period that marks the end of a sentence from periods in other functions (abbreviations, numbers, etc.). Sentence comparison calculates the matching probability of sentences in the source and target languages. Optimization finds the best sentence correspondence from a matrix of match probabilities.User interface translations are a crucial component of creating software for the World Wide Web. There are three ways to translate web pages: by the developer as part of the regular life cycle, by the community with a stake in the field, as well as through machine translation. An informal assessment of the quality generated by these three methodologies is provided in this study. We explore the fallacious belief that language translations may be assessed by performing additional translations and assessing the final product. The translation evaluation criteria are offered. We frequently take language into account when designing user interfaces as a means of user communication. When creating interfaces for worldwide audiences after the globalisation and localization procedures, we pay particular attention to language. The usage of symbolism, colours, currency denominations, date, hour, and number formatting on other surface difficulties are frequently referred to as cultural issues in interface design.

Language translation has always involved entering the source as text or audio and waiting for the system to produce the required translated output. The Acoustic Dialect Decoder (ADD), a voice-to-voice earpiece translation system, is described in this work. We introduce and review the most recent developments in speech engineering for use in attention deficit disorder (ADD), paying special attention to the three main processing processes of recognition, translation, and synthesis. The earpiece's speech recognition unit will capture ambient speech, and once one sentence has been successfully read, translation will begin. Our goal is to provide translated output as and when input is being read in this way. HMM-based Tool-Kit (HTK), RNNs with LSTM cells, and HMM-based Speech Synthesis System (HTS) will be used in the recognition and synthesis units, respectively. The initial purpose of this system will be as an English to Tamil translator.

Language translation has always involved entering the source as text or audio and waiting for the system to produce the required translated output. The Acoustic Dialect Decoder (ADD), a voice-to-voice earpiece translation system, is described in this work. We introduce and review the most recent developments in speech engineering for use in attention deficit disorder (ADD), paying special attention to the three main processing processes of recognition, translation, and synthesis. The earpiece's speech recognition unit will capture ambient speech, and once one sentence has been successfully read, translation will begin. Our goal is to provide translated output as and when input is being read in this way. HMM-based Tool-Kit (HTK), RNNs with LSTM cells, and HMM-based Speech Synthesis System (HTS) will be used in the recognition and synthesis units, respectively. The initial purpose of this system will be as an English-to-Tamil translator

India is a multilingual nation; nevertheless, not all Indians are polyglots. There are eleven well-known scripts and 18 official languages. Since the majority of Indians, especially those living in distant villages, cannot read, write, or understand English, an effective language translator must be used. Without regard to language barriers, machine translation systems that translate text from one language to another would advance Indian civilization. We suggest an English-to-Hindi machine translation system based on declension rules because English is a worldwide language and Hindi is spoken by the majority of Indians. This essay also discusses the various machine translation methodologies

This system design can be extended to handle complex compound sentences for translations. This system can also be used to develop other Indic multilingual translation systems. The system can continue to build better word options using its large database of bilingual dictionaries

Traditional speech translation systems use a cascade manner that concatenates speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis to translate speech from one language to another language in a step-by-step manner. Unfortunately, since those components are trained separately, MT often struggles to handle ASR errors, resulting in unnatural translation results. Recently, one work attempted to construct direct speech translation in a single model. However, that work is only evaluated in Spanish, and English language pairs with similar syntax and word order.This paper describes the machine learning approach developed at VTT for the discovery of translation rules for translating from one controlled language (CL) to another. Our Webtran machine translation software implements supervised machine learning to help adaptation to new language pairs . The rules we are dealing with are rewritten rules consisting of feature constructs.

This paper describes the machine learning approach developed at VTT for the discovery of translation rules for translating from a controlled language (CL) to another. Our Webtran machine translation software implements supervised machine learning to help adaptation to new language pairs . The rules we are dealing with are rewrite rules consisting of feature constructs.This paper describes the machine learning approach developed at VTT for the discovery of translation rules for translating from a controlled language (CL) to another. Our Webtran machine translation software implements supervised machine learning to help adaptation to new language pairs . The rules we are dealing with are rewrite rules consisting of feature constructs.The translation software maintains much of the grammatical structure of the original language when translated to another language. This results in a low score in the translation. The result of translating back into the original language is that the resulting structure is appropriate because it was the originating structure to begin with.

III. EXISTING TECHNOLOGY

India is a multilingual country; different states have different territorial languages but not all Indians are polyglots. There are 18 constitutional languages and ten prominent scripts. The majority of the Indians, especially the remote villagers, do not understand, read or write English, therefore implementing an efficient language translator is needed. Machine translation systems, that translate text form one language to another, will enhance the knowledgeable society of Indians without any language barrier.

3.1GOOGLE TRANSLATOR

Google Translator converts written words between different languages. Ninety languages are supported. The number of paragraphs and technical terminology that Google Translate can translate is limited. While being translated, some nouns, such as person characters, stay unchanged. Google Translate doesn't need an introduction. The translation tool, which was introduced in 2006, formerly produced translations word-for-word using SMT. However, Google has now stopped using SMT in favour of the more precise NMT, leading to 14 consistently higher translation quality. Recurrent neural networks are used in Google's proprietary machine translation system, known as Google Neural Machine Translation (GNMT), to translate complete sentences while preserving as much of their context as possible.

3.2DEEPL TRANSLATOR

The German company Linguee GmbH (now known as DeepL GmbH), which focuses on creating deep learning machine translation technology, created the NMT service DeepL Translator. Since its 2017 release, DeepL Translator has studied and learned as much as it can about the most effective translation possibilities from reputable linguistic sources. In contrast to its competitors, DeepL Translator uses artificial intelligence to provide translations that are more accurate and sophisticated. It might actually live up to its own claim to be "the most accurate translator in the world.

3.3BING MICROSOFT TRANSLATOR

The foundation of Bing Microsoft Translator is Microsoft's own machine translation engine, which utilises cutting-edge NMT technology. Microsoft has concentrated its research efforts on creating smarter machine translations that correspond to natural language use, like the majority of machine translation software vendors. For instance, Bing Microsoft Translator uses an attention algorithm to assess the words that should be translated in what order to produce the most accurate results.

3.4SYSTRAN TRANSLATE SYSTRAN is an established brand. It was established in 1968 and is arguably the first company to provide paid machine translation services. And this industry pioneer hasn't rested on its laurels: this machine translation supplier frequently adds cutting-edge machine translation technology and capabilities to its SYSTRAN Translate product. In order to produce extremely accurate translations, for instance, SYSTRAN's most recent pure neural machine translation (PNMT) engine simulates the entire machine translation process using an artificial neural network.

3.5 AMAZON TRANSLATE For those in the know, Amazon is more than just a top retail site; with its Amazon Translate feature, it also offers machine translation. According to the Amazon Translate website, this NMT service was developed to offer economical, quick, and high-quality language translations. Regardless of the length of the source text, Amazon consistently improves its datasets to provide the finest translations. Additionally, users can add their own translation data to tailor translations to their tastes using Amazon Translate's Active Custom Translation functioncomprehensibility, the model is pre-trained on vast amounts of audio-visual data and fine-tuned on particular datasets. The study's evaluation of the model's applicability to the reconstruction of Mandarin speech yields encouraging findings. The researchers also demonstrate that a pre-trained speech recognition system can achieve cutting-edge performance on benchmark datasets in both English and Mandarin by fine-tuning the produced audios.

3.1.1 DEMERITS OF MACHINE TRANSLATION SYSTEMS

- Level of accuracy can be very low
- Accuracy is also very inconsistent across different languages.
- Machines can't translate context.
- Mistakes are sometimes costly.

Inconsistency in the level of accuracy. A complex project that required multiple languages, would need to hire numerous professional native-speaking translators. While this enables you to reach a much broader customer base, it lowers the level of accuracy and can easily distort the true meaning of your message. Having multiple translators interpret your messages opens the door to the translator's nemesis: inconsistency. For example, the word “expression” may be repeated throughout the source text, but could end up being translated differently by various translators, leading to varied interpretations and inconsistencies. The lack of uniformity becomes even riskier when you're dealing with a multilingual project. 16 Limited Use Since it cannot offer you accuracy and precision, machine translations have very limited use. Certain critical documents that involve technicalities and intricacies should preferably be translated through professional translators rather than machine translations. Poor Quality Quality of the translated text is one of the biggest disadvantages that come with machine translations. Machine translations fail to translate the text in the appropriate context. Translation of cultural references, idiomatic expressions, industry jargons and other details happen better when you rely on human translations. Costly Mistakes Since nobody's perfect, your translator is likely to make mistakes at times, like giving you faulty translations or accidently altering the final message in minor ways. Such errors can be very costly, especially if you rely on the translation to make serious decisions. And because you probably don't understand the target language, you'll only realize the mistakes when the damage is already done. Remember that language translation involves reproducing the actual meaning of a message in the source language, as accurately as possible. This is a delicate exercise because you must be sure that the words in the translation are the most acceptable rendition of the original text.

IV. MOTIVATION

There can be several motivations for reporting on speech to speech translation, including:

1. Advancements in technology: Speech to speech translation technology is constantly evolving and improving, so reporting on the latest developments can help keep people informed about the state of the art.
2. Accessibility: Speech to speech translation can help break down language barriers, making communication easier and more accessible for people who speak different languages.
3. Cultural exchange: Speech to speech translation can facilitate cross-cultural communication and understanding, allowing people from different parts of the world to learn about each other's cultures and traditions.
4. Business and trade: Speech to speech translation can be beneficial for companies and organizations that do business with partners or clients from different parts of the world, making it easier to communicate and negotiate deals.
5. Travel: Speech to speech translation can be a useful tool for travelers who visit countries where they don't speak the local language, allowing them to communicate with locals and navigate unfamiliar environments more easily.

Overall, reporting on speech to speech translation can help raise awareness of the technology and its potential benefits, as well as its limitations and challenges.

V. PROPOSED SOLUTION

OVERVIEW

The project aims to develop a real-time language translation system using the Python programming language. Instead of dubbing in movies and videos, our project works at converting one language to another using Natural Language Processing without human intervention. Initially the audio from video is extracted using Moviepy library of Python. The source audio is converted to text. The source text is converted to destination language text using GoogleTrans API. Then, the GTTS Google Text to Speech Translation is used to convert the text to destination language audio. The audio is then merged with the original video to give the final output.

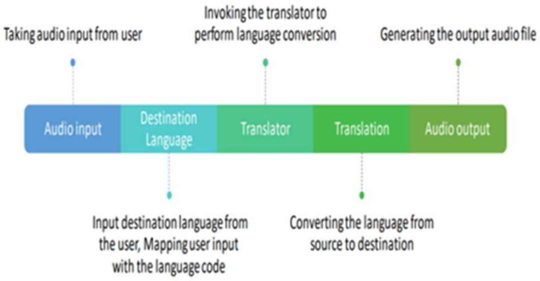
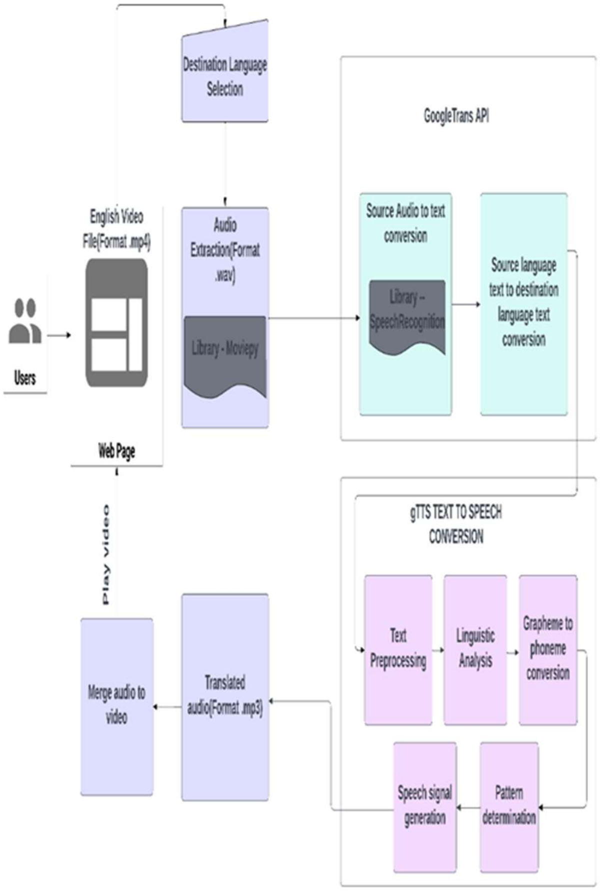

Taking audio input from user
Invoking the translator to perform language conversion
Generating the output audio file
Audio input
Destination Language
Translator
Translation
Audio output
Input destination language from the user, Mapping user input with the language code
Converting the language from source to destination

Fig. 4.1 Work flow of Machine Translation

For those who have trouble speaking, this technology offers a way to translate lip motions into speech. The device can reliably and quickly transcribe speech in real-time by analyzing the movement of the lips and facial expressions using lip-reading techniques and sophisticated machine learning algorithms. For those who have trouble speaking, this technology can significantly enhance communication and make it easier for them to take part in real-time interactions. The technique enables voice to text translation for persons with cognitive impairments and may be incorporated into a variety of devices, including teleconferencing equipment, speaking aids, and speech recognition software. This has the potential to transform how people with speaking challenges connect with the outside world by taking use of the inherent co-occurrence of audio and visual streams in videos.


Destination Language Selection
GoogleTrans API
English Video File(Format .mp4)
Audio Extraction(Format .wav)
Source Audio to text conversion
Library - SpeechRecognition
Source language text to destination language text conversion
Library - Moviepy
Users
Web Page
Play video
gTTS TEXT TO SPEECH CONVERSION
Text Preprocessing
Linguistic Analysis
Grapheme to phoneme conversion
Merge audio to video
Translated audio(Format .mp3)
Speech signal generation
Pattern determination

VI. IMPLEMENTATION

6.1 GOOGLE TRANSLATE

To translate text, documents, and websites from one language into another, Google developed Google Translate, a multilingual neural machine translation tool. It provides a website interface, an Android and iOS mobile app, and an API that aids developers in creating software apps and browser extensions. Google Translate covers 133 languages at various levels as of January 2023, and as of April 2016, it claimed to have over 500 million overall users, translating more than 100 billion words every day. In May 2013, the firm had said that it serviced more than 200 million people every day. It was introduced in April 2006 as a statistical machine translation service, gathering linguistic information from documents and transcripts from the European Parliament and the United Nations. In most of the language pairings it suggests in its grid, with a few exceptions, such as Catalan-Spanish, it initially translates text to English before pivoting to the target language. When translating, it searches through millions of papers for patterns that can be used to guide the selection of words and their placement in the target language. Its accuracy, which has received criticism on numerous times, has been found to be very inconsistent between languages. Google Neural Machine Translation (GNMT), which translates using neural machine translation, was introduced as the replacement for Google Translate in November 2016 "Instead of only one word at a time, use entire phrases. It determines the most pertinent translation using this larger context, which it then rearranges and tweaks to sound more like a human speaking with good grammar."

Translation Methodologies

Google Translate launched in April 2006 with a statistical machine translation engine. Since Google Translate's algorithms are based on statistical or pattern analysis rather than on conventional rule-based analysis, it does not follow grammatical rules. Franz Josef Och, the original developer of the system, has argued that statistical methods are more effective than rule-based algorithms in many situations. The first iterations of Google Translate were built on a technique known as statistical machine translation, and more particularly on work by Och, who won the 2003 DARPA contest for speed machine translation. Och oversaw Google's machine translation division until he left in July 2014 to work for Human Longevity, Inc. From one language to another (L1 to L2), Google Translate does not translate. Instead, it frequently translates from the original language (L1) to English and then from there to the target language (L2). However, this might lead to translation problems because English, like all human languages, is ambiguous and reliant on context. For instance, the Russian equivalent of the French word vous is you т OR В/в. It would be vous you В/в OR tu thou т if Google were acting as the intermediary and utilising an unambiguous artificial language. Such word suffixing clarifies the many meanings of the terms. Therefore, depending on the target language, publishing in English, using unambiguous phrases, offering context, and using idioms like "you all" may or may not make a better one-step translation. The following languages lack an English to or from them translation on Google. In addition to using English, the following languages are translated using the stated intermediate language (which is typically closely related to the target language but is more generally spoken) Belarusian (be ↔ ru ↔ en ↔ other); Catalan (ca ↔ es ↔ en ↔ other); 24 Galician (gl ↔ pt ↔ en ↔ other); Haitian Creole (ht ↔ fr ↔ en ↔ other); Korean (ko ↔ ja ↔ en ↔ other); Slovak (sk ↔ cs ↔

en ↔ other); Ukrainian (uk ↔ ru ↔ en ↔ other); Urdu (ur ↔ hi ↔ en ↔ other). Och asserts that a bilingual text corpus (orparallel collection) of more than 150–200 million words and two monolingual corpora, each of more than a billion words, would provide a strong foundation for the creation of a workable statistical machine translation system for a new pair of languages. The translation between those languages is subsequently performed using statistical models created from these data. Google used documents and transcripts from the European Parliament and the United Nations to gather this enormous amount of linguistic data. The UN regularly publishes documents in each of the six official UN languages, creating a fairly sizable corpus of the six official UN languages. Representatives of Google have participated in domestic conferences in Japan where they have asked scholars for multilingual data. In order to select the best translation, Google Translate analyses patterns in hundreds of millions of pages as it makes translation proposals. Google Translate generates educated estimates (AI) about what a proper translation should be by looking for patterns in texts that have already been translated by humans. Before October 2007, Google Translate was built on SYSTRAN, a software engine that is still used by a number of other online translation services like Babel Fish, for languages other than Arabic, Chinese, and Russian (now defunct). Instead, starting in October 2007, Google Translate made use of in-house, proprietary technology built on statistical machine translation until switching to neural machine translation. 25 Statistical machine translation There are languages that continue to employ the conventional translation technique known as statistical machine translation, despite Google's deployment of a new system called neural machine translation for higher quality translation. Using predictive algorithms, it is a rule-based translation technique that makes educated guesses about how to translate texts into other languages. Instead of translating individual words, it looks for overlapping sentences to translate. To create a statistical model that translates texts from one language to another, it also examines bilingual text corpora. Google Neural Machine Translation The Google Neural Machine Translation system (GNMT) was developed by a research team at Google to improve the fluency and accuracy of Google Translate. In November, it was reported that Google Translate would transition to GNMT. The massive end-to-end artificial neural network used by Google Translate's neural machine translation system aims to do deep learning, namely long short-term memory networks. Because it employs the example-based machine translation (EBMT) technique, in which the system "learns from millions of examples," GNMT sometimes provides better translation quality than SMT. Google researchers claim that it translates "Instead of just one word at a time, use entire sentences. It determines the most pertinent translation using this larger context, which it then rearranges and tweaks to sound more like a human speaking with good grammar ". Google Translate has incorporated the "proposed architecture" of "system learning" from GNMT in more than a hundred languages. With the end-to-end architecture, Google claims but does not show that "the system learns over time to provide better, more natural translations" for the majority of languages. The GNMT network tries interlingual machine translation, which "encodes the semantics of the sentence rather than just memorising phrase-to-phrase translations," and the system "uses the commonality found in between multiple languages, rather than inventing its own global language." Eight languages were initially supported by GNMT, including 26 English, Chinese, French, German, Japanese, Korean, Portuguese, Spanish, and Turkish. Hindi, Russian, and Vietnamese were the first languages to have it enabled in March. Bengali, Gujarati, Indonesian, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu were added in April. Googletrans API for Python Googletrans is a free and unlimited python library that implemented Google Translate API. Python googletrans is a module to translate text. It uses the Google Translate Ajax API to detect languages and translate text. Features Fast and reliable - it uses the same servers that translate.google.com uses Auto language detection Bulk translations Customizable service URL Connection pooling (the advantage of using requests.Session) HTTP/2 support 4.3.4 TEXT TO SPEECH The process of turning words into vocal audio is known as text-to-speech (TTS). The tool, programme, or software takes a user-provided text as input, understands the linguistics of the language being used, and applies logical inference to the text using techniques from Natural Language Processing. The next block performs digital signal processing on the processed text after passing it there. This processed text is finally transformed using numerous algorithms and transformations into a spoken format. Speech synthesizing plays a role in every step of this process. 27 Fig. 4.3 Speech Synthesizer Google Text To Speech gTTS is a very easy to use tool which converts the text entered, into audio which can be saved as a mp3 file. The gTTS API supports several languages including English, Hindi, Tamil, French, German and many more. The speech can be delivered in any one of the two available audio speeds, fast or slow. However, as of the latest update, it is not possible to change the voice of the generated audio. Grapheme to Phoneme Conversion Grapheme-to-phoneme (G2P) conversion is the process of generating pronunciation for words based on their written form. It has a highly essential role for natural language processing, text-to-speech synthesis and automatic speech recognition systems. Grapheme-to-phoneme conversion (G2P) is the task of transducing graphemes (i.e., orthographic symbols) to phonemes (i.e., units of the sound system of a language). For example, for International_Phonetic_Alphabet IPA): "Swifts, flushed from chimneys …" → "ˈswɪfts, ˈfləʃt ˈfɹəm ˈtʃɪmniz …". 28 Modern text-to-speech (TTS) synthesis models can learn pronunciations from raw text input and its corresponding audio data, but by relying on grapheme input during training, such models fail to provide a reliable way of correcting wrong pronunciations. As a result, many TTS systems use phonetic input during training to directly access and correct pronunciations at inference time. G2P systems allow users to enforce the desired pronunciation by providing a phonetic transcript of the input.G2P models convert out-ofvocabulary words (OOV), e.g. proper names and loaner words, as well as heteronyms in their phonetic form to improve the quality of the synthesized text. Heteronyms represent words that have the same spelling but different pronunciations, e.g., “read” in “I will read the book.” vs. “She read her project last week.” A single model that can handle OOVs and heteronyms and replace dictionary lookups can significantly simplify and improve the quality of synthesized speech

MODULES USED

MOVIEPY

MoviePy is a Python module for video editing, which can be used for basic operations on videos and GIF’s. Video is formed by the frames, combination of frames creates a video each frame is an individual image. An audio file format is a file format for storing digital audio data on a computer system. The bit layout of the audio data is called the audio coding format and can be uncompressed, or compressed to reduce the file size, often using lossy compression. We can load the audio file with the help of AudioFileClip method.

SPEECH RECOGNITION

A program's capacity to convert spoken language into written language is known as speech recognition, also known as Automatic Speech Recognition (ASR), computer voice recognition, or speech-to-text. Despite being sometimes confused with voice recognition, speech recognition focuses on converting speech from a verbal to a text format whereas voice recognition only aims to distinguish the voice of a certain person. Working of Speech Recognition Python's speech recognition uses algorithms that model speech in terms of both language and sound. In order to extract the more important parts of speech, such as words and sentences, acoustic modelling is utilised to distinguish the phenones and phonetics in our speech. With the aid of a microphone, speech recognition first converts the sound energy supplied by the speaker into electrical energy. This electrical energy is subsequently transformed from analogue to digital and eventually to text. It separates the audio data into sounds and then uses algorithms to analyse the sounds to determine which word is most likely to fit the audio. Natural Language Processing and 20 Neural Networks are used for all of this. The accuracy of voice recognition can be increased by identifying temporal patterns using hidden Markov models.

Features of Speech Recognition

There are several voice recognition software and hardware options, but the more sophisticated ones make use of AI and machine learning. To comprehend and process human speech, they integrate the grammar, syntax, structure, and composition of audio and voice signals. They ought to develop their reactions as they go along, learning from each engagement. The finest solutions also let businesses alter and modify the technology to suit their particular needs, from brand recognition to language and speech peculiarities. For instance: Language weighting: Increase accuracy by giving extra weight to certain phrases that are commonly used in speech (such as brand names or industry jargon). Speaker labelling: Produce a transcription of a multi-participant conversation that references or tags each speaker's contributions. Training in acoustics: Focus on the acoustical aspect of the enterprise. Train the system to adjust to different speaker types and acoustic environments, such as the background noise in a contact centre (like voice pitch, volume and pace). Filters can be used to clean voice output by identifying specific words or phrases that are considered profane.

Speech Recognition Algorithms

Different computer methods and algorithms are utilised to recognise voice into text and enhance transcription accuracy. Some of the most popular techniques are briefly explained below

Natural Language Processing

Although there isn't necessarily a single algorithm employed in speech recognition, natural language processing (NLP) is the branch of artificial intelligence that focuses on communication between humans and machines using speech and text. Many mobile devices have speech recognition built into their operating systems to enable voice search (like Siri) and to increase messaging accessibility. Hidden Markov Model The Markov chain model, which argues that the probability of a given state depends on its present state and not its past states, is the foundation for

Hidden Markov Models (HMMs)

A hidden Markov model enables us to include hidden events, such as part-of-speech tags, into a probabilistic model, whereas a Markov chain model is appropriate for observable events, such as text inputs. They are used as sequence models in speech recognition, labelling each item in the sequence (words, syllables, phrases, etc.). These labels build a mapping with the input given, enabling it to choose the best label order.

N-Grams The simplest kind of language model (LM), known as n-grams, assigns probabilities to individual sentences or phrases. A series of N words make up an N-gram. For instance, the words "order the pizza" and "please order the pizza" each have a three-gram or trigram length. The use of grammar and the likelihood of particular word combinations helps to increase recognition and precision.

Neural Networks Neural networks handle training data by simulating the connectivity of the human brain using layers of nodes, which is mostly used for deep learning algorithms. Each node consists of an output, a bias (or threshold), weights, and inputs. This "fires" or activates the node, sending data to the following layer in the network, if the output value exceeds a predetermined threshold. This mapping function is 22 learned by neural networks through supervised learning, with gradient descent adjustments made in response to the loss function. Although neural networks are more accurate and have a larger data set than classic language models, this comes at the expense of performance efficiency.

Speaker Diarization (SD) Algorithms for speaker identification and voice segmentation. This makes it easier for programmes to tell apart people in a conversation and is widely used at call centres to tell apart customers and sales personnel

CONCLUSION

Implementing real-time speech translation with Python requires the use of multiple libraries and technologies such as SpeechRecognition, Googletrans, and gTTS. The process involves capturing the audio input, recognizing thelanguage, translating it into the desired language, and then synthesizing the translated text into speech. The quality of the translation depends on the accuracy of the speech recognition model and the machine translation model used. It is important to note that in real-time language translation, it is crucial to use efficient algorithms and optimize code for performance. In addition, to improve the accuracy of the translation, it is recommended to use pre-trained models and optimize them with a specific dataset. Overall, real-time language translation requires a good understanding of the underlying technologies and careful implementation to achieve an accurate result

REFERENCES

[1] S. Bakhshaei, S. Khadivi, N. Riahi, “Farsi-German Statistical Machine Translation Through Bridge Language”, IEEE 5th International Symposium on Telecommunications, pp. 557-561, 2010.

[2] C. Yang, “Cross-Language Instant Messaging With Automatic Translation”, IEEE Computer Society, Fourth International Conference on Ubi-Media Computing, pp. 222-226, 2011.

[3] H. Yu, F. Ren, D. Huang, and L. Li, “Designing Effective Web MinningBased Techniques for OOV Translation”, IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-8, 2010.

[4] K. Macherey, O. Bender, and H. Ney, “Applications Of Statistical Machine Translation Approaches to Spoken Language Understanding”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 4, pp. 803- 818, May 2009.

[5] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura, “End-to-end speech translation with transcoding by multi-task learning for distant language pairs,” IEEE ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 1342–1355, 2020.

[6] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, “Sequence-to-sequence models can directly translate foreign speech,” in Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, 2017, pp. 2625–2629

[7] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, “Neural lattice-to-sequence models for uncertain inputs,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2017, pp. 1380–1389.

[8] H. Ney, S. Nieben, F. Josef Och, H. Sawaf, C. Tillmann, and S. Vogel, “Algorithms for Statistical Translation of Spoken Language”, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, January 2000. 39

[9] J. Tenni, A. Lehtola, C. Bounsaythip, and K. Jaaranen, “Machine Learning Of Language Translation Rules”, 1999 IEEE International Conference On Systems, Man, and Cybernetics, Vol. 5, pp. 171-177, 1999

[10] Gen-ichiroKikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita, “Comparative study on corpora for speech translation,” IEEE TransactionAudio, Speech & Language Processing, vol. 14, no. 5, pp. 1674–1682, 2006

[11] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, “Attention-passing models for robust and data-efficient end-to-end speech translation,” Transactions of the Association for Computational Linguistics,vol. 7,pp. 313–325, 2019.

[12] Kaho Osamura, Takatomo Kano, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura, “Using spoken word posterior features in neural machine translation,” in Proceedings of the 15th International Conference on Spoken Language Translation, IWSLT, vol. 21, p. 22, 2018

[13] S. E. Bou-Ghazale and J. H. L. Hansen, “A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress,” IEEE Trans. Speech Audio Process., vol. 8, no. 4, pp. 429–442, 2000

[14] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” IEEE Trans. Audio, Speech Lang. Process., vol. 17, no. 1, pp. 66–83, 2009.

[15] Sanjib Das, “Speech Recognition Technique: A Review”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue 3, May-Jun 2012.