IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT
Impact Factor: 6.252

Volume 2, Issue 3, May 2022

Real Time Direct Speech-to-Speech Translation

Sanchit Chaudhari, Aniket Shukla, Tanvi Gaware, Nehali Shinde
Students, Department of Information Technology[1,2,3]
Professor, Department of Information Technology[4]
Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India

Abstract: *Speech-to-Speech translation has been developed to ease and bridge the communication gap between people who speak different languages. Speech-to-Speech technology is effective because it allows speakers of languages from around the world to communicate with each other by minimizing the language. Gap in global commerce and cross-cultural communication. Recent studies have recognized speech-to-speech translation as one of the top technologies that will transform and reform the current language barriers in our world. Using Python and Spyder libraries development of the voice-to -voice translation system is possible. Use of Google's Translate API comes in handy during the whole development process.*

Keywords: Speech-to-Speech, Language Barrier, Translation, Python, Spyder, Voice-to-Voice, API

I. Introduction

The number of countries exchanging information continues to rise. International visitors for tourism, emigration, or overseas study are growing more diverse, necessitating the development of a system that allows people who speak different languages to communicate effectively. Automatic spoken-to-speech translation (S2ST) considerably reduces language barriers and bridges cross-cultural divides by allowing people to converse in their own languages. Over the last several decades, many researchers have been working on constructing an S2ST system.

Traditional S2ST approaches necessitate time and effort to build various components, such as automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis, all of which must be trained and calibrated separately. ASR processes and transforms voice into text in the source language, MT converts the source language text into comparable text in the target language, and TTS generates speech from the target language text.

II. Problem Statement

We propose a technique for training speech-to-speech translation tasks without the use of transcription or linguistic supervision in the suggested system. To identify and localize language, we utilized a machine learning technique and a speech-to-speech algorithm.

III. Motivation

The necessity for a speech-based translation system originates from the fact that standard Text-To-Text translation systems disregard paralinguistic factors such as prosody, voice, and emotion of the voice speaker. Our drive is focused on developing a system that can best embody the aforementioned characteristics while translating from one language to another. Furthermore, the cascaded three-stage machine translation systems have the potential to exacerbate the errors that occur with each successive phase. Speech-To-Speech systems, on the other hand, have a benefit over cascaded TT systems since they use a single step method that requires less computer power and has better inference latency.

IV. Objective

The main objectives of the system are: To create a login system for the users to logon into the program, to have a translator to translate the languages as per need of end user, to have the output of the translation in form of both voice as well as text.

V. Proposed Framework

The proposed real time speech-to-speech language translator consists of 4 modules i.e. Login, Input, Translation, Output.

- **Input Module:** This module consists of taking an input in form of voice and sending it into the system.

DOI: 10.48175/568

- **Translation Module:** This module does the hard work of translating the given voice input into the desired output by using the speech-to-speech translation technology by means of Google's Translate API.
- **Output Module:** This module provides the result of work performed by the system in the form of either text or voice

VI. System Overview

- **Operating System:** Windows 10+
- **Python:** It is a high level general purpose programming language which is significantly easy to understand and learn, especially for beginners.
- **VS Code:** It is a free to use development environment written in python and for python.
- **Anaconda Navigator:** Anaconda is a package and environment manager used for python. Anaconda Navigator is a GUI which comes pre-packed in Anaconda which allows us to easily launch applications and manage various packages
- **OpenCV:** The OpenCV provides a real time Computer Vision libraries and functions.
- **Pillow:** Pillow Library adds image processing functionality to our python interpreter.

VII. Literature Survey

- **Neural Machine Translation:** Dzmitry Bahdanau suggested that the neural translation machines aims towards constructing a single neural network which will be jointly tuned to increase the performance between translation. With this method we can gain a rapid translation which can be compared to the existing state of art phrase based system.
- **Tacotron:** Yuxuan Wang proposed that a text to speech translation system usually consists of multiple modules such as text analysis, acoustic model, audio synthesis model, etc. but building this requires expertise knowledge and domain expenses. Using Tacotron we display an end to end text to speech model that translates speech directly from any given text.
- **Speech Translation without Transcription:** Long Doung suggested that there are many low resource languages do not have any kind of orthography for transcription. Even if we try Phonetic transcription, the expense for it is high. By performing set of experiments using phone-to-word alignment there was upto 24% possibility of success over several transcription base lines.
- **Speech-to-Speech Translation for Non-transcribed Languages:** Andros Tjandra proposed a method for developing speech-to-speech translation without any kind of linguistic supervision. It consists of 2 steps – 1. Monitor and generate representation with unsupervised discovery with discrete auto encoder. 2. Execute a sequence-to-sequence model that directly matches the core language .
- **Language Identification for Multilingual Translation:** Arun Babhulgaonkar proposed a n-gram and machine learning based language identifiers to identify 3 specific Indian languages specifically Hindi, Marathi and Sanskrit which were present in a document for translation
- **Multilingual Corpus for Speech Translation:** Javier Iranzo-Sanchez described the corpus creation process and displayed a series of automatic speech recognition, machine translation and language translation.

VIII. Workflow

The central objective of Real Time Direct Speech-to-Speech Translation System is to provide an easy to use and accurate translation system which can present output in real time which previously was an hard task to be achieved due to lack of resources and transcriptions. The entire system consists of three main layers namely The User Interface, The Application and The Backend. Users will interact with the system by means of the user interface. The application acts as a bridge between the user interface and the backend database. Backend is the database that contains all the transcriptions of the languages that are supported in the translation.

7.1 Users

The two key intended users are the Admin and the end consumer.

1. Admin has the ability to edit and update the whole dataset that contains the resources used for translation process. Admin can also manage the registered users in the system.
2. End users can login to the system and use the Direct Speech-to-Speech Application for translating any language into their desired choice.

7.2 Admin Workflow

The admin generally manages the whole system's dataflow and registers the users who are going to implement the translation application. The admin can also manage the whole dataset in the backend database to edit or add new language dataset to further improve the functionality of the application. The workflow of an admin is as following steps.

1. Start
2. Register/Login
3. Monitor and check the status of the system
4. Add/Edit Language datasets
5. Logout
6. Stop

7.3 End Consumer Workflow

The end consumer does an initial registration by providing their details such as their Email and a Password for security purposes. After successfully registering into the application the end user can then login to the system and can make use of the Direct Speech-to-Speech translation to translate any kind of language into their desired one. Theworkflow of an end user compromises of the following steps.

1. Start
2. Register/login
3. Make use of the Speech-to-Speech Translator
4. Give input
5. Receive output
6. Logout
7. Stop

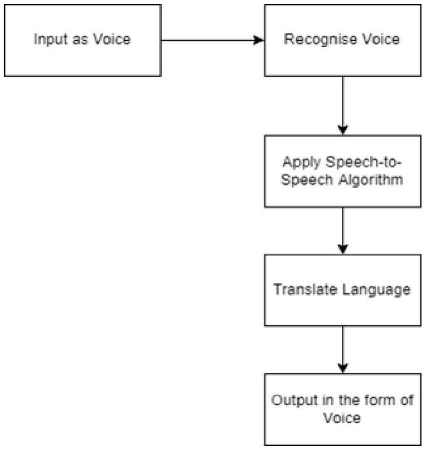

Figure 1: System Architecture for Speech-to-Speech Translation

IX. Algorithm

Speech-to-Speech: Speech recognition is a computer science and computational linguistics multidisciplinary topic that develops approaches and technology that allow computers to recognise and translate spoken language into text. Automatic speech recognition (ASR), computer voice recognition, and speech to text are some of the other names for it (STT). It includes computer science, linguistics, and computer engineering expertise and research. Some voice recognition systems necessitate "training" (also known as "enrolment"), in which a single speaker reads text or isolated vocabulary into the system. "Speaker dependent" systems are those that rely on training. The term "voice recognition" or "speaker identification" refers to recognising the speaker rather than the content of their speech. Recognizing the speaker can help systems that have been trained on a specific person's voice translates speech more quickly, or it can beused to authenticate or verify the speaker's identity as part of a security process. Speech recognition has a long history in terms of technology, with multiple waves of key advancements. Advances in deep learning and big data have recently improved the field. The advancements are proven not only by the increasing number of academic articles published in the subject, but also by the widespread industry adoption of a range of deep learning approaches in the design and deployment of voice recognition systems around the world.

X. Future Work

Since the technology has been under constant updates addition of Artificial Intelligence will boost the performance and make the flow of the system easier. Addition of multiple outputs will be a key factor in upcoming works. Maximizing the time of process and output can be a helpful upgrade.

XI. Conclusion

We propose a new approach for training a speech-to-speech translation between two languages without the need of transcription in this system. To begin, a discrete quantized auto encoder was trained to build a discrete representation from the target voice features. Second, given the source speech representation, we trained a sequence-to-sequence model to predict the codebook sequence. Because the target speech representations are learned and generated unsupervised, this method can be applied to any sort of language, with or without a written form. Our model can perform a direct speech-to-speech translation.

Acknowledgment

We wish to acknowledge our professor Prof. Nehali Shinde, who helped and guided us throughout the research process of the given entitlement “Real Time Direct Speech-to-Speech Translation”. We also acknowledge IJARSCT for publishing our paper on their portal.

References

[1]. Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 577–585.

[2]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” CoRR, vol. abs/1409.0473, 2014.

[3]. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., “Tacotron: Towards endto-end speech synthesis,” arXiv preprint arXiv:1703.10135, 2017.

[4]. Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn, “An attentional model for speech translation without transcription,” in NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, 2016, pp. 949–959.

[5]. Alexandre Berard, Olivier Pietquin, Christophe Servan, and ´ Laurent Besacier, “Listen and translate: A proof of concept for end-to-end speech-to-text translation,” CoRR, vol. abs/1612.01744, 2016. .

DOI: 10.48175/568