International Journal of Advance Research in Science and Engineering
Volume No. 12, Issue No. 09, September 2023
www.ijarse.com

IJARSE
ISSN 2319 - 8354

Deep Learning-Driven Voice Casting for Content Localization

Anish Jain[1], Rajendra Patha[2], Rahul Pandit[3], Aditya Shah[4], AbhijeetCholke[5]

[1-5]*Computer Engineering, Trinity Academy of Engineering, Pune, India*

ABSTRACT

Content localization, the process of converting audio visual content from a source language to a target language, plays a crucial role in reaching global audiences. Traditionally, voice casting for dubbing content relies on manual selection of voice actors, but recent advances in deep learning offer an opportunity to automate and enhance this practice. Our research paper focuses on advancing a Text-to-Speech (TTS) model within a multilingual voice translation system. While leveraging existing libraries for Automatic Speech Recognition (ASR) and Machine Translation (MT), our TTS model plays a central role in lifelike speech synthesis. We explore technical intricacies, including TTS model architecture, fine-tuning strategies, and linguistic nuances."

Keywords: *Deep Learning, Text-to-Speech, Multilingual Voice Translation, Automatic Speech Recognition*

1. INTRODUCTION

In our highly connected world, the need for content localization has grown significantly. Voice casting, the process of selecting voice actors to dub content from one language into another, plays a crucial role in making content accessible to diverse global audiences. This process involves the meticulous transformation of original source language material into foreign languages, most notably achieved through dubbing. Central to this complex localization process is the critical element of voice casting. Voice casting is the art and science of selecting voice actors who will lend their vocal talents to dub the characters in the target languages. However, this process is often manual and subjective, which comes with its limitations. Recent advances in deep learning, specifically in technologies like recurrent neural networks and neural network embeddings, provide potential solutions to these challenges. These technologies can automate and improve voice casting by evaluating factors like language fluency, emotional expression, and cultural alignment. This research aims to explore how deep learning can simplify and enhance the voice casting process for content localization. Our main goal is to create a system that can smoothly translate voices from one language to another. By harnessing the power of deep learning, we seek to make localized content more accessible and engaging, contributing to bridging the linguistic and cultural gapsin the entertainment and media industries.

2. DESIGN METHODOLOGY

This system leverages its core capabilities to streamline the process of voice casting for multimedia content. Whether you're a filmmaker, content producer, or educator, our Automated Voice Casting System empowers you to bring your ideas to life in multiple languages with ease and efficiency. The Prototype will function as follows:

2.1 Data Collection:

Gather a diverse dataset of multilingual video content from the user, including source language videos and corresponding transcriptions, translations, and target language voiceovers.

2.2 Automatic Speech Recognition (ASR):

Implement a ASR model to transcribe source language speech in videos accurately.

2.3 Machine Translation (MT):

Develop and fine-tune a robust machine translation system for translating transcribed text into the desired targetlanguage.

2.4 Text-to-Speech (TTS):

Create a natural-sounding Text-to-Speech system capable of generating voiceovers in the target language.

2.5 Synchronization Algorithm:

Design and implement an algorithm to ensure precise synchronization between translated text and the originalaudio in videos.

2.6 User Interface:

Develop an intuitive, web-based user interface for video upload, language selection, and customization of voiceoptions.

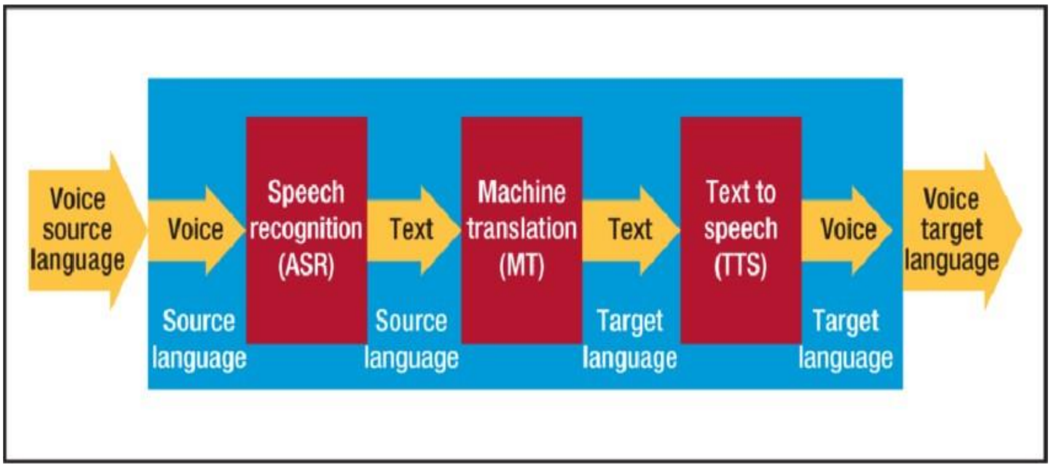

Fig. 1: Voice Translation and Synthesis System

3. USE CASE

3.1 Begin by uploading your video to our platform.

3.2 Next, select the source language, which is the language spoken in the video or audio.

3.3 Once the upload is complete, choose your desired target language.

3.4 Our automated system will then perform the following actions:

- It will transcribe the audio from the source language and translate it into your chosen target language.
- The transcript and the audio file will be processed by our Text to Speech model to generate a voice clone

with the provided content.

- This resulting output will be synchronized with the source video, producing a coherent final result.
- The completed video or audio, now translated and synchronized, will be presented to you.
   - Additionally, as a user, you have the option to download the translated video or audio for your convenience

4. RESULT (EXPECTED OUTCOME)

This System is designed to seamlessly convert spoken language into highly accurate written text using Automatic Speech Recognition (ASR). This transcribed text is then efficiently translated into multiple languages through Machine Translation (MT). However, our primary focus lies in our Text-to-Speech (TTS) technology, which transforms the translated text into a natural human voice, ensuring lifelike and engaging content. This emphasis on TTS significantly advances content localization, making information and communication more accessible and user-friendly across diverse linguistic backgrounds. With ASR facilitating accurate transcription and MT enabling multilingual adaptation, our platform simplifies global communication for promoting inclusive content localization.

CONCLUSION

The Multilingual Video Voice Translation System breaks language barriers, enhancing global accessibility to multimedia content. It streamlines the translation process, ensuring accuracy, cost-efficiency, and user engagement. By promoting cultural exchange and inclusivity, it stands as a powerful tool for content creators andaudiences worldwide.

ACKNOWLEDGEMENTS

The authors thank Trinity Academy of Engineering for their support and guidance.

REFERENCES

1) Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, “Voice conversion for persons with amyotrophic lateral sclerosis,” IEEE J. Biomed. Health Informat., vol. 24, no. 10, pp. 2942–2949, Oct. 2020.

2) J. Matoušek, Z. Hanzlíˇcek, D. Tihelka, and M. Méner, “Automatic dubbing of TV programmes for the hearing impaired,” in Proc. IEEE 10th Int. Conf. Signal Process., 2010, pp. 589–592.

3) N. Obin, A. Roebel, and G. Bachman, “On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification,” Proc. IEEE Int. Conf. Acoustics Speech Signal Process, (ICASSP), Florence, Italy, pp. 950–954, 2014, doi: 10.1109/ ICASSP.2014.6853737.

4) N. Obin and A. Roebel, “Similarity Search of Acted Voices for Automatic Voice Casting,” IEEE/ACM Trans. Audio Speech Language Process. 24(9):1642–1651, Sep. 2016, doi: 10.1109/ TASLP.2016.2580302.15) Ismail K. A. R and Celia V. A. G Rosolen, “Effects of the airfoil section, the chord and pitch distributions on the aerodynamic performance of the propeller”, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*,41:131,(2019).