Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

Real-Time Text & Speech Translation Using Sequence To Sequence Approach

Dikshita Patel
B.Tech Student
Dept. of Information Technology
S.P.I.T, Andheri
Mumbai, India
dikshita.patel@spit.ac.in

Minakshi Kudalkar
B.Tech Student
Dept. of Information Technology
S.P.I.T, Andheri
Mumbai, India
minakshi.kudalkar@spit.ac.in

Shashank Gupta
B.Tech Student
Dept. of Information Technology
S.P.I.T, Andheri
Mumbai, India
shashankkumar.gupta@spit.ac.in

Renuka Pawar
Prof.IT Dept
Dept. of Information Technology
S.P.I.T, Andheri
Mumbai, India
renuka_pawar@spit.ac.in (Prof)

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544509

***Abstract* -The following paper discusses the process of research and development of a real-time application that helps people from different linguistics to interact with each other fluently without any language barriers. We aim to improve the current traditional 3-tiered convolutional architecture of translation and propose a sequence-to-sequence improvisation without relying on the intermediate text representation length. We propose an application where there is availability for chat/audio/video translations for the users.**

Keywords - Speech to Text conversion, text translation, text to speech conversion, Parallelization, Convolutional architecture.

I. INTRODUCTION

Individuals and groups around the world are facing language barriers while learning and communicating with each other, thus reducing efficiency and the potential reach of ideas. This gives the need for a translation system across chat-messengers, conference calls, etc. with minimum latency, and maximum accuracy.

The current translation systems work on the traditional 3-tier architecture converting speech to text, performing machine translation of this, and converting the translated text back to speech. This system has the main disadvantage of high latency.

Developing a real-time application will help in bridging the barriers in communication and promote cross-language and cross-border interactions having an impact on the world economy as well.

The model makes use of an improved 3 tier architecture, dividing the translations into multiple mini-batches for quicker processing.

This is coupled with fine-tuning, cleaning & processing to reduce the temporal dimension of the audio input. Further, a vast corpus for each of the languages is maintained on which the model is regularly trained and updated to maintain maximum accuracy possible. Finally, a convolutional neural network (CNN) is implemented to predict the translations. This results in higher accuracy than the normally used RNN sequence to sequence model.

II. LITERATURE SURVEY

A system where the user speaks in his/her own language of comfort. The audio is converted to text. Real-time translation of the text into the language opted by the other party. Technologies used for developing the model include Automatic Speech Recognition for human voices, TTS, also MT i.e., Machine Translation system. Future work aims to enhance and upgrade the ASR and MT system,
along with synthesizing standards for Text-to-Speech translation. Recent work provides support for Malayalam, English & Hindi translations. Accuracy is 70% for Malayalam.[1]

A plugin free and platform-independent,

multilingual video chat application which allows user speak/text in different languages of their comfort. The paper is similar to a multilingual chat system, where the chat messages during the call are translated into the user's preferred language. Technologies used for their system include Web Real-Time Communication for peer-to-peer communication, Transliterate API from Google, Translator provided by Microsoft, also Google Web Speech API.[2]

The paper provides a speech translation model along with a friendly user interface to satisfy the requirements of users. Future work includes that if the speech-to-speech translation log records are collected for performance enhancement of the translation, it can result in an increased inaccuracy. Present accuracy is 85% for text-based machine translation. [3]

The paper provides us with the security features of the recently widely used messaging applications, which also states the reason to improve the security of those systems. Many messaging apps are not secured which are widely used for message transfer. For example, WeChat, Viber is prone to Man-In-The-Middle attacks as they do not provide end-to-end encryption of the messages.[4]

This paper proposes an extension "Skype-SeVid", that uses a selective encryption technique for video calling, via Skype. The host performs data encryption and thus has less control over the processing related to the video data stream. TechStack: Skype client (C++ programming language) from the Skypekit API. Drawbacks include that it cannot be run on proxy servers, also other messaging features are not provided.
Accuracy:
Encryption 0.015534 seconds/frame,
Decryption 0.015902 seconds/frame [5]

Advocate a framework that helps in developing complex and real-time software applications with the use of component processing algorithms.
It was designed for use in speech-to-speech translation systems that require real-time processing and Advocate gives this along with reduced latency. The only delay will be on the first service of 5 seconds. The average throughput will be 12 requests/min. Less throughput due to delays in transmission over the network. Throughput is also lowered when there are very large data transmitted as this caused overhead in data transfer but on the other hand, we have low processing time.[6]

[8]A real-time translation method for mobile devices is a paper that explains captured images that can be converted to text extracted. This makes use of the GPS (Global positioning system) databases i.e., Native languages and encourages swift tourism. The characters in the image captured are recognized with a language used in the location of the mobile device and can be used for translation. This proposal presents a new method of translation. Instead of upcoming artificial intelligence models, they translate between languages using a translation database according to the GPS location of the country/region.

[9] The next patent is 'System and method for assisting language learning'. It is essential to study the mechanism behind the teaching assistant as we plan to add it to the application. This proposal logs speech data. This is recorded for analysis over a period of time. The voice input from every language assisting session may be reckoned to find out the overall score and performance of the user. This provides us with results and instructions for learning and exploring a new language.

[10] In-Call Translation is the most recent and modern approach to machine translation applications. It follows the standard market-ready 3-tiered architecture. Users have to select their choice of the source language and all their conversations on the text will be converted to that language. The output of this is given to the next block of text translation and finally, this text is converted to speech in the choice of language of the other user. Synthetic Voice Addition plus visual avatar for actions. Appropriate notifications are given to the users while the translation is in process.

[11] The patent claims a model to provide an end-to-end method for the Android platform. It uses a

protocol called ERTP or Encrypted Real-Time Transfer Protocol to encrypt the sender's and receiver's messages using a public key - private key combination. The voice messages are signed with a DSA (Digital Signature Algorithm) key which is then encrypted with the AES algorithm to ensure there is no loss of Confidentiality and Integrity.

[12] The patent claims to provide real-time multi-platform support across the Internet as well as Application Development Operating Systems like Android, Windows, Linux, etc. It deploys special

prebuilt dictionaries with continuous improvisations and can be used to translate instant messages, chat, SMS messages, web pages, Internet search results, and so on.

[13] Real-time conversational analytics facility. This patent describes the procedure of the speech to text and machine translation stage. It provides the score and sentiment behind the source language voice input. It also discusses retaining the voice characteristics of the language spoken by the user 1. It assigns scores, sentiment, and category to make outputs more closely to human interaction for better knowledge transfer and learning.

III. ARCHITECTURE

The phases in the speech translation are given in the below image Figure 2. The user who wishes to use the feature of speech translation can use our web application. The following are the steps of speech translation.

1. The user selects the input language.
2. Speech input in the above-selected language using the microphone,
3. The user selects the output language in which he /she wants the translation.
4. Select the Language translation button to get translation in speech as well as text.

Our application will first convert user Speech to text, then apply language translation on the text using our translation model, finally convert the converted text to speech.

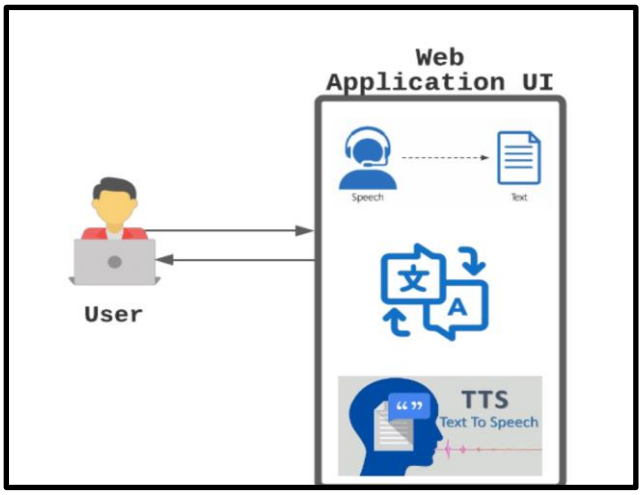

Figure 1 Speech Translator

Our application will first convert user Speech to text, then apply language translation on the text using our translation model, finally convert the converted text to speech.

The chat application is built on Django for frontend and backend and SQLite as the database. We record the user's id and password for user authentication and store chats in a different table with attributes like user id (as a foreign key), messages as text, and timestamp of the chat. We have built this application in the form of a chatroom where any number of users can come together to have discussions in the language of their choice.

The further architecture of the machine learning model is as follows:

We make use of a convolutional approach to achieve the same accuracy as recurrent models but with less latency. Our model outperforms the recurrent models because we can fully parallelize the inputs making our model accuracy and latency independent of the input length. We have achieved a BLEU score of 0.26 in our model training. We have successfully implemented the system with 8 global languages using this system architecture.

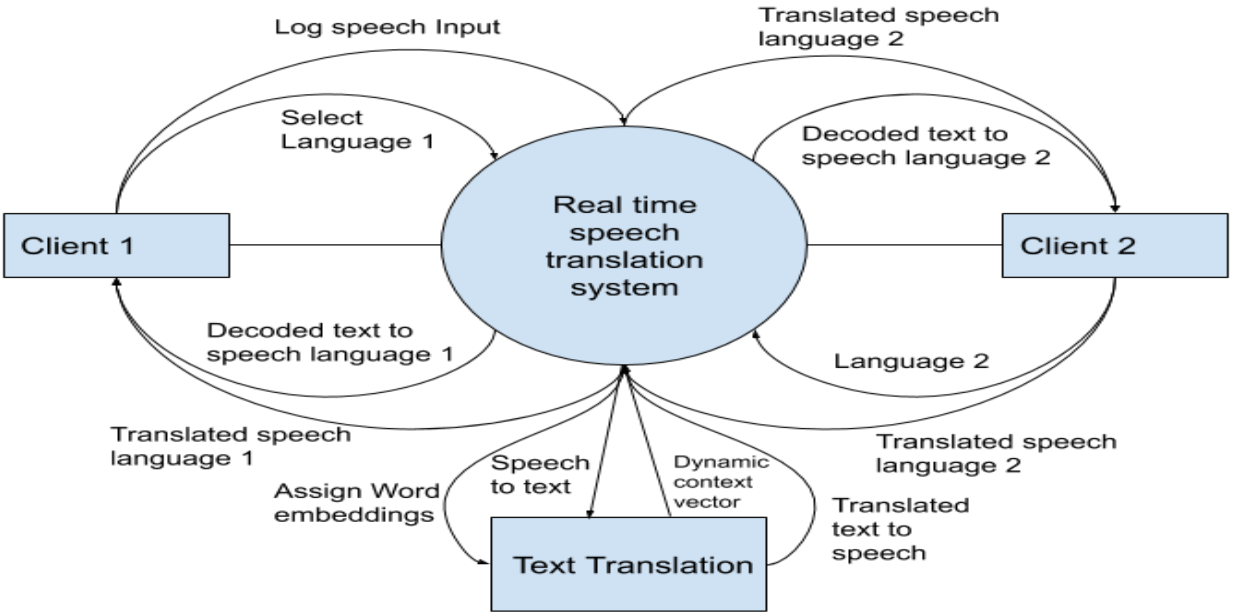

Figure 2 Level 0 Data Flow Diagram: Proposed Methodology

IV. OUR PROPOSED SYSTEM

1. Fine-Tuning Data

We propose a sequence-to-sequence model for machine translation using convolutional architecture. The data set we used was [7]. A Multilingual Speech Translation Corpus is a translation corpus whose size and quality will facilitate the training of end-to-end systems for SLT from English into 8 languages.

For Hindi to English, we used the NLTK Word2vec embedding library [14]. Word2vec is one algorithm for learning a word embedding from a text corpus.

There are various problems one needs to tackle to reach the final efficient and swift machine learning model. Some of these problems in our problem statement is:

1. Language may not follow the same order of words to form the same sentence as the source language

2. Language may use a different number of words to represent the source sentence

To address this issue, we make use of a context vector. Each word in the input sentence (source language) is mapped to a word embedding which gives content to the word in that sentence. We make use of these embeddings to form this context vector and implement an encoder-decoder model to tackle the translation. There might be cases where the input to the encoder is too large to be combined as one context vector, in this case, we use dynamic hidden vectors that are created and fed into the model for faster/parallel training.

2. Sequence to Sequence Approach

We have an attention module that improves the simple recurrent approach. It not only helps in dynamic translations but also alignment. While encoding the input transcript into a single context vector, the attention model creates context vectors that are scanned and filtered for each output step. Thus, we achieve a stage where our input sentence can be as long as possible, it wouldn't affect the speed or accuracy as much.

Models	Epoch	Total Time	Loss at final epoch	Accuracy at final epoch	Time Adjusted for Epoch
Simple RNN	10	23	2.66%	0.5192	92
Embedding RNN	40	120	0.62%	0.8432	120
Convolutional	40	396	1.44%	0.7454	396
Combining Embeddings and Convolutional	40	481	0.03%	0.9927	481

Table 1 Comparison of different algorithms for translation

While training, compared to the recurrent neural networks model, contextualization and calculations over all modules can be fully parallelized. Achieving optimization is easier with this approach as the number of non-linear features will remain fixed and independent of the input length to the machine translation stage.

We will utilize gated linear units that ease gradient

propagation. Will provide each decoding layer with a completely separate attention module.

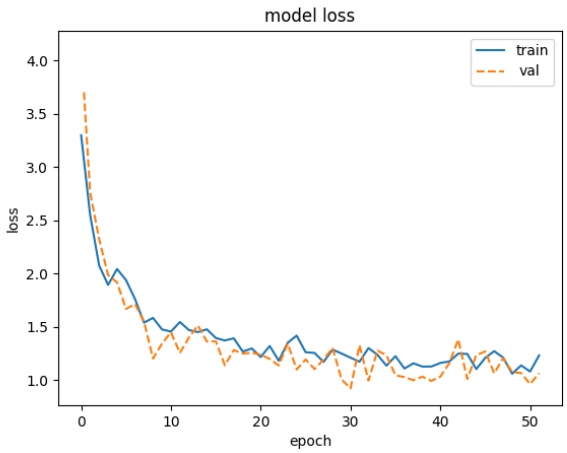

Figure 3 Train test curve

Unlike recent RNN networks, this convolutional approach allows the discovery of diverse structures in the sequences more easily and swiftly.

3. Results and Observations

We have trained convolutional models for 8 global languages including English, German, French, Spanish, Japanese, Brazilian, Hindi, and Marathi. The average BLEU score we achieved was 0.26 for translation. We tested our model with long documents like one with 4 pages containing 10098 characters and observed that the time for translation of this document was as similar as translation time for a single sentence. This was the advantage of using a convolutional approach along with word embeddings instead of a recurrent approach.

Table 2 shows the comparison of results for a word document containing 10098 characters and long sentences consisting of the most popular English sentences. The first three bars, Simple RNN, Embedding RNN, and Convolutional are the test results for our methodology broken down into its basic algorithms. The final bar denotes the highest accuracy which is '0.9927/1' for the proposed model.

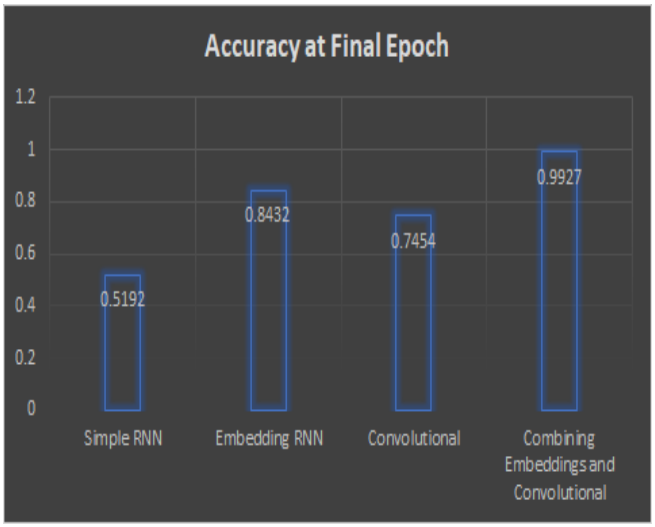

Table 2 Results & Comparison Of Accuracy

V. LIMITATIONS

Now, although we get the appointments scheduled, there will be some cases where the agents take a lot of time to come on topic and agree to the same goals. Here there is a possibility of agents fall into an infinite loop of communicating with each other without producing necessary results.

For ensuring the agents remain aligned with the target, we can use dynamic timely checks for the same.

VI. CONCLUSION

When a user ventures online to acquire a certain service, in a global market where there exist 7117 languages, this can be a hindrance to businesses or personal communications.

The proposed implementation of real-time speech translation provides an accuracy of

99.27% which is a new perspective to the traditional 3-tiered recurrent neural network architecture replacing it with convolutional structure increasing the overall accuracy of the system and can help with communication across languages. Compared to the traditional model there is a definite increase in accuracy and performance vis a vis speed.

VII. FUTURE SCOPE

Real-time speech translation systems are one of the most trending research areas in machine learning. The future scope of this research is to make use of this to build fully furnished applications to make our world more connected, breaking language barriers and communicating freely.

The convolutional model serves as the basis for future applications which can leverage this technology.

VIII. REFERENCES

1. J. Andhale, C. Dadi and Z. Fei, "A Multilingual Video Chat System Based on the Service-Oriented Architecture," 2017 IEEE Symposium on Service-Oriented System Engineering (SOSE), San Francisco, CA, 2017, pp. 126-131, doi: 10.1109/SOSE.2017.17.
2. A. Gopi, Shobana Devi P, Sajini T, J. Stephen, and Bhadhran VK, "Multilingual speech to speech MT-based chat system," 2015 International Conference on Computing and Network Communications (CoCoNet), Trivandrum, 2015, pp. 771-776, doi: 10.1109/CoCoNet.2015.7411277.
3. P. K. Aggarwal, P. S. Grover, and L. Ahuja, "Security Aspect in Instant Mobile Messaging Applications," 2018 Recent Advances on Engineering, Technology, and Computational Sciences (RAETCS), Allahabad, 2018, pp. 1-5, doi: 10.1109/RAETCS.2018.8443844.
4. S. Yun, Y. Lee and S. Kim, "Multilingual speech-to-speech translation system for mobile consumer devices," in IEEE Transactions on Consumer Electronics, vol. 60, no. 3, pp. 508-516, Aug. 2014, doi: 10.1109/TCE.2014.6937337.
5. A. A. Ramdan and R. Munir, "Selective encryption algorithm implementation for video call on Skype client," 2012 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA), Bali, 2012, pp. 120-124, doi: 10.1109/TSSA.2012.6366035.
6. A. Ryan Aminzadeh And Wade Shen, “Advocate: A Distributed Architecture For Speech-To-Speech Translation”, Volume 18, Number 1, 2009 N Lincoln Laboratory Journal
7. Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, Marco Turchi, "MuST-C: a Multilingual Speech Translation Corpus" Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1

Patents

8. Inventec Corp, 2012, ‘*Real Time Translation Method For Mobile Device*’ Us2012130704a1.
9. Lue Julia, Jerome Dubreull, Jehen Bing, 2017, ‘*System and Method for Assisting Language Learning’*, California, USA, US9786199B2.
10. Anthony Aue, Arul A. Menezes, Jonas Nils Lindblom, Frederik Furesjo, Pierre P. N. Greborio, 2017, ‘*In-Call Translation’*, Redmond, Washington, USA, US9614969B2.
11. Changzhou Academe Southeast University, 2018, ‘*Voice End-To-End Encryption Method Aiming At The Mobile Terminal With Android System’* Cn103974241b
12. Robert E. Levin, 2006, ‘*Language Translation System and Method Using Specialized Dictionaries’*, New York City, USA, US6996520B2
13. Dwyer Michael C, Gallino Jeffrey A, Kendrick Scott A,Salinas Frank, Strand Erik A, Wolf Scott R, Xue Shaoyu, 2015, ‘*Real-Time Conversational Analytics Facility*’, Massachusetts, Usa, Us2015195406a1.
14. API used: http://www.nltk.org/nltk_data/, Used Word2vec library from the NLTK has built-in support for dozens of corpora and trained models.
15. The tool used for creating diagrams are, https://creately.com/ b. https://app.diagrams.net