Speech to Speech Translation System

Mr.Kalyan D Bamane[1], Utkarsh Mishra[2], Rohit Ippili[3], Bharghav Palaneer[4], Shravan Bhuyar[5]

[1]Mr. Kalyan D Bamane Assistant Professor, Dept. of Information Technology of D Y Patil College of Engineering, Akurdi, Maharashtra, India
[2]Utkarsh Mishra Dept. of Information Technology of D Y Patil College of Engineering, Akurdi Maharashtra, India
[3]Rohit Ippili Dept. of Information Technology of D Y Patil College of Engineering, Akurdi Maharashtra, India
[4]Bhargav Palaneer Dept. of Information Technology of D Y Patil College of Engineering, Akurdi Maharashtra, India
[5]Shravan Bhuyar Dept. of Information Technology of D Y Patil College of Engineering, Akurdi Maharashtra, India

---***---

Abstract -*Speech to Speech Translation (S2ST) is a pipe dream for human beings that enable communication between people speaking in different languages.*

Multilingual speech-to-speech translation devices are vital for breaking the language barrier, which is one of the most serious problems inherent in globalization.

To improve and expand the possible field of research extending the scope across languages increasing the overall efficiency and productivity of the aforementioned research

To help facilitate conversation among people who speak different languages using automatic transcription and subsequent translation of the input voice.

Without language barrier, people can communicate using the language they are comfortable with, which will in turn speed up a range of business processes.

Key Words: **Speech to Speech, Speech Recognition, Translation, Multilingual, Text to Speech**

1. INTRODUCTION

Since our world is becoming borderless day by day, the drastic increase in demand for translingual conversations, triggered by IT technologies such as the Internet and an expansion of borderless communities as seen in the increase in the number of EU countries, has boosted research activities on S2ST technology. Many research projects have addressed speech-to-speech translation technology, such as VERBMOBIL, C-STAR, NESPOLE! and BABYLON.

These projects mainly focused on the construction of prototype systems for several language pairs.

S2ST between Western languages and a non-Western language, such as English-from/to-Japanese, or English-from/to-Chinese, requires technologies to overcome the drastic differences.

2. LITERATURE SURVEY

A. *The ATR Multilingual Speech-to-Speech*

Translation System

ATR multilingual speech-to-speech translation (S2ST) system, which is mainly focused on translation between English and Asian languages (Japanese and Chinese).Problem faced were, the current translation system needs improvement in translating longer sentences often found in natural dialogs. It is also weak in translating variations often found in natural dialogs. Finally, a confidence measure for translation is also pursued.

B. Speech recognition with deep recurrent neural networks.

Investigation of deep recurrent neural networks, which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long range context that empowers RNNs. Some of the problems faced were the next step is to extend the system to large vocabulary speech recognition. Another interesting direction would be to combine frequency domain convolutional neural networks with deep LSTM.

C. Statistical parametric speech synthesis using deep neural networks.

Conventional approaches to statistical parametric speech synthesis typically use decision tree-clustered context-dependent hidden Markov models (HMMs) to represent probability densities of speech parameters given texts.

D. Wave Net: A Generative Model for Raw Audio

This paper introduces Wave Net, a deep neural network for generating raw audio waveforms. When applied to text-to-speech, it yields state-of-the-art performance.

3. SYSTEM ARCHITECTURE

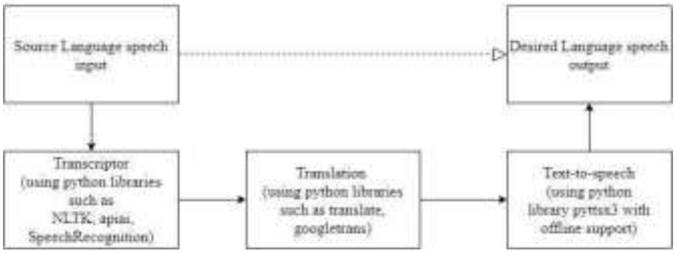
Source Language speech input
Desired Language speech output
Transcriptor (using python libraries such as NLTK, apiai, SpeechRecognition)
Translation (using python libraries such as translate, googletrans)
Text-to-speech (using python library pyttsx3 with offline support)

Fig -1: Architecture of proposed system

4. SYSTEM SPECIFICATIONS

- Hardware Requirement:
  - 2.2 GHZ CPU
  - 8 GB RAM
  - Speaker and Microphone.
  - Network Interface Card, good Internet Connection.
- Software Requirement:
  - 64 bit Operating System.
  - Python, CUDA, cuDNN.
  - Google Cloud SDK.
  - Visual Studio and Visual Studio Code.
  - Python Packages like NLTK, Translate, Wave, IBM Watson, gTTS

5. PROPOSED SYSTEM

Algorithms

We use the IBM Watson speech to text service which uses a combination of models like LSTM and ResNets [6]. In previous papers, the improvement in the English conversational telephone speech recognition using the combination of LSTM and ResNet models. It contains a combination of ResNet with two types of LSTM architectures i.e. SA-MTL and Feat. Fusion.

For translation purposes the following segmentation technique is one of the technique [7] used. For processing arbitrary words, words are broken into wordpieces by a trained wordpiece model. Special word boundary symbols are added before training of the model such that the original word sequence can be recovered from the wordpiece sequence without ambiguity. At decoding time, the model first produces a wordpiece sequence, which is then converted into the corresponding word sequence.

Modules

1. Speech Input : In this module , the input is given in either real time audio or saved audio file in desired langauge .(Implemented 100%)

2. Speech to Text Processing : We provide the audio input either in streams or in saved audio format as the input for this module. We use IBM Watson's Speech to Text API for this purpose and depending on the type of audio or it's format, we need to change the parameters accordingly.

(Implemented 100%)

3. Translation and Grammar Correction: This module takes the input from previous module, then checks for any grammatical errors and tries to rectify them so as to enable better understanding for the user of the output text.

(Implemented 50%)

4. Text to Speech Processing: In this module we use Google Text To Speech library or gTTS to convert the translated and conditioned transcript to speech in the language desired. (Unimplemented)

5. Speech Output: This module provides the output of the speech in the desired language and proving useful for the user. (Unimplemented)

6. CONCLUSION

We have successfully developed a system to facilitate translingual conversations among trading entities and improve the flexibility of research. The developed system is also capable of implementing the use of Cloud based Speech Recognition and TTS systems employing the mentioned methods of NNS. We understood the use of Machine Learning, Natural Language Processing, Object Character Recognition, Deep Neural Networks and Artificial Intelligence in the use of Speech to Speech Translation. Thereby we have managed to achieve our end goal of catering to customers and breaking the language barrier between them in a simple and robust manner.

7. FUTURE SCOPE

We have developed multilingual corpora and machine learning algorithms for speech recognition, translation, and speech synthesis. The results have convinced us that our strategy is a viable way to build a high-quality S2ST system. The current translation system needs improvement in translating longer sentences often found in natural dialogs; therefore, we are studying a method to split a longer sentence into shorter ones and translate them. It was also weak in translating variations often found in natural dialogs; therefore, we employed the RNN approach. Finally, a confidence measure for translation is now being pursued and it will be incorporated to reject erroneous translations.

The system initially consists of only one language as output. With time we can include other languages starting with the popular ones(German, Chinese, Japanese, Russian, etc) and building up the number of languages it can give the output as speech. It can be converted to a true offline application for translation, for reading documents or; over calls, by introduction of machine learning models for the transcription and subsequent translation process. The document translation can initially be supported for basic documents with extensions .txt, .docx, etc. while further it can be also used for translating pdf documents by introduction of image detection models. It will instigate more research as the overhead for translating research papers will get reduced and will save a lot of time for the purposes of reading or for citations.

8. ACKNOWLEDGEMENT

We would sincereley like to thank our head of department Dr Preeti Patil Mam and our assistant professor Mr Kalyan D Bamane for their constant guidance and motivation throughout the whole duration of the project . We could'nt have acheived such success in this project without their support. Lastly we are grateful for our seniors and colleagues for their support.

REFERENCES

1. W. Wahlster, Ed., Vermobil: Foundations of Speech-to-Speech Translations. Berlin, Germany: Springer-Verlag, 2000.

2. E. Costantini, S. Burger, and F. Pianesi, “NESPOLE!’s multi-lingual and multi-modal corpus,” in Proc. LREC, 2002, pp. 165–170.

3. K. Saino, A clustering technique for factor analysis-based eigen voice models, Master thesis, Nagoya Institute of Technology, 2008, (in Japanese).

4. H. Zen, M. Gales, Y. Nankaku, and K. Tokuda, “Product of experts for statistical parametric speech synthesis,” IEEE Trans. Audio Speech Lang. Process., vol. 20, no. 3, pp. 794–805, 2012.

5. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 14 –22, Jan. 2012.

6. Saon, George, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. English Conversational Telephone Speech Recognition by Humans and Machines. Proceedings of Interspeech 2017 (August 2017): pp. 132-136.

7. [Wuet al., 2016]Yonghui Wu, Mike Schuster, ZhifengChen, Quoc V. Le, Mohammad Norouzi, WolfgangMacherey, Maxim Krikun, Yuan Cao, Qin Gao, KlausMacherey, Jeff Klingner, Apurva Shah, Melvin John-son, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws,Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, KeithStevens, George Kurian, Nishant Patil, Wei Wang, CliffYoung, Jason Smith, Jason Riesa, Alex Rudnick, OriolVinyals, Greg Corrado, Macduff Hughes, and JeffreyDean. Google’s neural machine translation system: Bridg-ing the gap between human and machine translation.CoRR, abs/1609.08144, 2016.