L.MOUNIKA, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 10, Issue 8, (Series-III) August 2020, pp. 57-61

RESEARCH ARTICLE OPEN ACCESS

Speech to Speech Translating (S2ST) System using ASR and TTS Techniques for Indian languages (English to Telugu or Tamil)

L.MOUNIKA*, Prof. B.V.S.S.N.RAJU**
**(M.Tech, Department of E.C.E, SRKR ENGINEERING COLLEGE (Autonomous), Bhimavaram, A.P-India*
*** (Professor, Head of Dept, Department of E.C.E, SRKR ENGINEERING COLLEGE (Autonomous), Bhimavaram, A.P-India.*

ABSTRACT
Speech to Speech Translation (S2ST) is a subfield of computational linguistics that investigates the use of software to translate text or speech from one natural language to another. S2ST can be especially useful for performing tasks that involve understanding and speaking with people who don't speak the same language. Manual translation has been limited to important official documents, news items and some award winning literary works. There exists a huge backlog of materials that needs to be translated for administration, education, commerce, tourism etc. Technological support in the form of machine aids for translation is of great importance. In this paper, we proposed a deep learning based Speech to Speech Translating System using MFCC, Automatic Speech Recognition (ASR) and Text To Speech(TTS)Technologies.

Keywords-Automatic Speech Recognition (ASR)**,** Text To Speech (TTS), Machine Translator, Neural Networks, Deep Learning.

Date of Submission: 06-08-2020 Date of Acceptance: 20-08-2020

I. INTRODUCTION

The global scenario adds the demand of communication among speakers of different languages is actually one of the great challenges before the Information Technologists to overcome language barriers across the global community, and enable them to express themselves in real time. Further, the task of bridging the Digital-divide can never be accomplished in real sense without breaking the language barriers with an intelligent system or machines. Speech to Speech translation technology is being able to speak and have one's words translated automatically into the other person's language has long been a dream of humankind. In the recent predictions, speech to speech translation has been placed as one of the ten technologies that will change the world.Speech-translation technology is significant because it enables speakers of different languages from around the world to communicate, erasing the language divide in global business and cross-cultural exchange. Achieving speech translation would have tremendous scientific, cultural, and economic value for humankind.

Speech to speech translation is one such system that can play important role by facilitating communication between persons speaking different languages. Worldwide efforts are being made to achieve this goal and implement it practically for use by common man.

II. RELATED WORK

Speech translation was first noticed during 1983 ITU Telecom World (Telecom'83), when NEC Corporation made a demonstration of speech translation as a proof of concept. In 1993, an experiment in speech translation was conducted linking three sites around the world: the ATR, Carnegie Melon University (CMU) and Siemens. Germany launched the Verbmobil project; the European Union the Nespole! and TC-Star projects; and the United States launched the TransTac and GALE projects.

Research and development has gradually progressed from relatively simple to more advanced translation, progressing from scheduling meetings, to hotel reservations, to travel conversation. Moving forward, however, there is a need to further expand the supported fields to include a wide range of everyday conversation and sophisticated business conversation.

The current frameworks for ASR utilize complex factual models. Shrouded Markov Models have been effective. These are factual models that yield a grouping of images or amounts. GMM-HMMs are utilized in discourse acknowledgment in light of the fact that a discourse sign can be seen

as a piecewise fixed sign or a brief timeframe fixed sign. Another motivation behind why GMM-HMMs are main stream is on the grounds that they can be prepared naturally and are basic and computationally practical to utilize. In any case, GMM-HMMs make different presumptions about the discourse and accordingly neglect to sum up.
The disadvantages are:

- It is costly, both as far as memory and register time.
- GMMs are measurably wasteful for demonstrating information that lie on or almost a nonlinear complex in the information space.
- The HMM should be prepared on a lot of seed groupings and by and large requires a bigger seed.
- For a given arrangement of seed successions, there are numerous conceivable HMMs, and picking one can be troublesome.

III. PROPOSED METHOD

Speech-to-speech translation has three components: Automatic speech recognition (ASR), MT, and voice synthesis (or text to speech; TTS). As shown in Figure 1, the ASR component processes the voice in its original language, creating a text version of what the speaker said. This text in the original language goes through the MT component, which translates it to the target language. Finally, this translated text goes through the TTS component, which "speaks" the text using a synthesized voice in the target language. For each step of this process, many other technologies could be used in the future to improve the quality of the overall speech-to-speech translation.

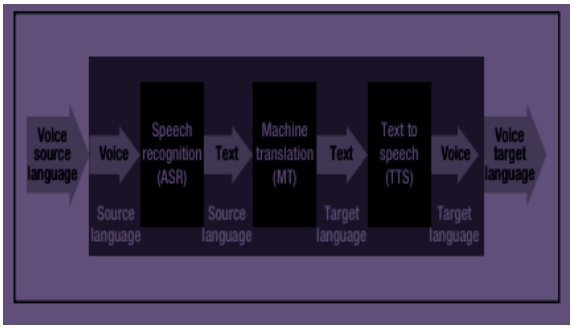

Fig1: **Speech to Speech translation components.**

Speech recognition is the capacity of a machine or program to recognize words and expressions in communicated in language and convert them to a machine-comprehensible arrangement.Numerous discourse acknowledgment applications, for example, voice dialing, basic information passage and discourse to-text are in presence today. Programmed discourse acknowledgment frameworks include various separate parts drawn from a wide range of orders, for example, measurable example acknowledgment, correspondence hypothesis, signal handling, combinatorial arithmetic, and semantics. Discourse acknowledgment is an option in contrast to customary strategies for connecting with a PC, for example, printed contribution through a console. A viable framework can supplant, or decrease the unwavering quality on, standard console input Attempts to fabricate programmed discourse acknowledgment (ASR) frameworks were first made during the 1950s. These early discourse acknowledgment frameworks attempted to apply a lot of linguistic and grammatical standards to recognize discourse. On the off chance that the expressed words clung to a specific standard set, the framework could perceive the words. Be that as it may, human language has various special cases to its own principles. The manner in which words and expressions are verbally expressed can be unfathomably adjusted by accents, lingos and idiosyncrasies. Accordingly, to accomplish ASR we utilize Deep Learning Algorithm.

MT is challenging because translation requires a huge amount of human knowledge to be encoded in machine-processable form. In addition, natural languages are highly ambiguous: two languages seldom express the same content in the same way. Google Translate is an example of a text-based MT system that applies statistical learning techniques to build language and translation models from a large number of texts. Other tools offer cross language chat services, such as IBM Lotus Translation Services for Sametime and VoxOx.

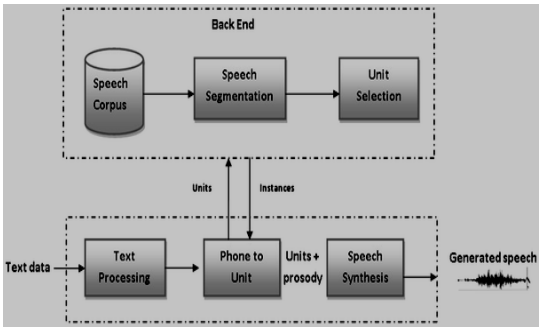

Fig 2: A Functional TTS.

The primary concern to comprehend about discourse is that the sounds produced by a human are separated by the state of the vocal tract including tongue, teeth and so forth. This shape figures out what sound comes out. On the off chance that we can decide the shape precisely, this

should give us an exact portrayal of the phoneme being created. The state of the vocal tract shows itself in the envelope of the brief timeframe power range, and the activity of MFCCs is to precisely speak to this envelope.
The highlights are removed as follows:
• Frame the sign into short edges.
• Apply hamming window to make the sign periodic.
• Calculate the periodogram gauge of the force range.
Apply the mel filterbank to the force spectra, whole the vitality in each channel.
• Take the logarithm of all filterbank energies.
• Take the DCT of the log filterbank energies.
• Keep DCT coefficients 2-13, dispose of the rest.
• Create a setting window of nearby casings to catch the phoneme setting, further took care of to neural system.

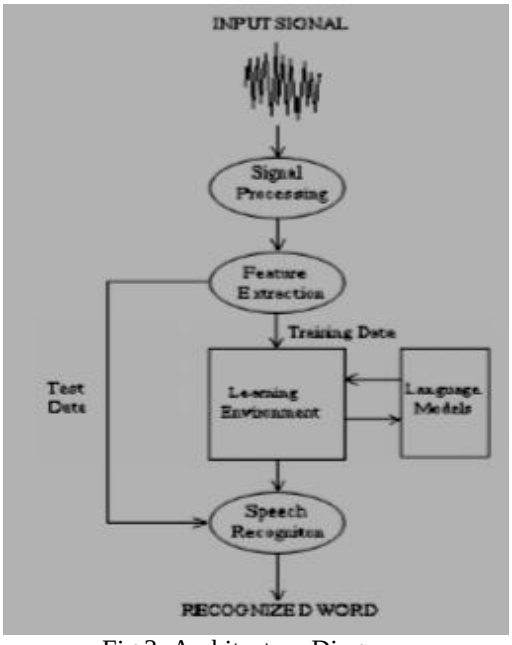

Fig 3: Architecture Diagram

A. Preprocessing:

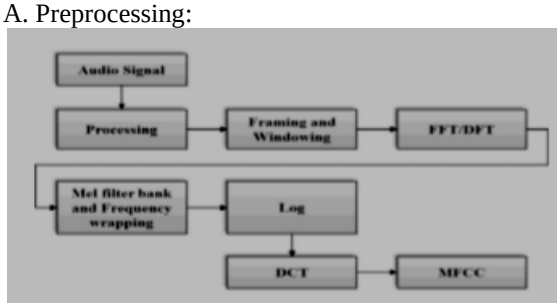

Fig 4: Feature Extraction

IV. DESIGN & IMPLEMENTATION

(a) Automatic Speech Recognition

A technical definition of ASR is the building of system for mapping acoustic signals to a string of words. The general problem of automatic transcription of speech by any speaker in any environment is still far from solved. But recent years have seen ASR technology mature to the point where it is viable in certain limited domains.
Five main sub-components of an ASR system are
a. Acoustic Model (p(O|W)
b. Language Model (P(W))
c. Lexicon/Pronunciation Model (HMM)
d. Feature Extraction
e. Decoder

For building Acoustic model, we used audio data of 8567 sentences that accounted for more than 60 h of recording. These sentences were recorded in a clean noise free environment, by the speakers uniformly distributed over all age groups from 17 to 60 years. Prototype models for 61 phonemes are built using flat start approach. The hmm models are over 61 context independent phonemes. Julius recognition engine is used for decoding the utterances which is a two pass stack decoder. The performance of LVCSR is measured in terms of recognition rate.

(b) Machine Translation

The approach for implementation is primarily using statistical machine translation (SMT).The advantage of SMT is that one does not require deeper syntactic understanding of Source and Target languages. The base models can be quickly built, as soon as we have the parallel corpus of the language pair with us. In the proposed scenario it is very difficult to have the man-power who is having experience and knowledge of multiple languages. And the languages of the consortia members are not even remotely related. Hence, SMT was the obvious choice for development. English has been chosen as the linking language around which each consortia member is developing their own language corpus. Then the respective translation models are to be developed among different language pairs directly without intervening English in between them. Each consortia member is tuning the system by supplementing linguistic information, such as transliterations, part-of-speech and chunk information.

(c) Text to speech synthesis (Telugu)

The Festival speech synthesis system is primarily designed for phoneme and di-phone units, and we adapted it to work for syllable units.

For data preparation, the text prompts were recorded in an anechoic chamber by a professional speaker. The recorded prompts were manually labeled at phoneme level using EMU speech tool. The prompts were further labelled to higher levels like syllables and words. The prosodic phrasings were also introduced in the database. The text processing module breaks the incoming text sentence to a syllable sequence. The unit selection module selects the best unit realization sequence from the many possible unit realization sequences for the given syllable sequence. The prosody prediction module predicts energy, pitch etc. Finally, in the concatenation module, the units are modified according to the predicted prosody before concatenation.

V. RESULTS ANALYSIS

This section gives results of proposed Speech to Speech Translator System using Matlab Software. The following figures show the results of Speech Training , Feature extraction, Speech Recognition and Speech to Speech (English word pronounced is taken as input and produced Telugu or Tamil Translating word is displayed as output and also language sounds pronounced as output reply to corresponding input English Sentences) Translator Results.

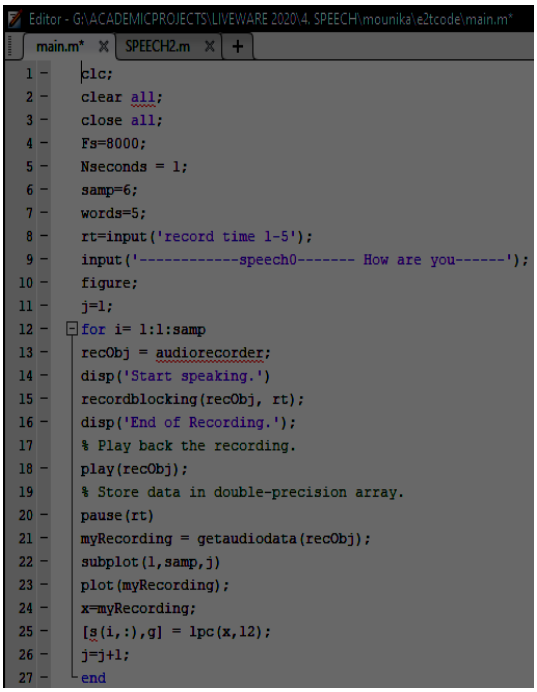

Fig 5: Training speeches results

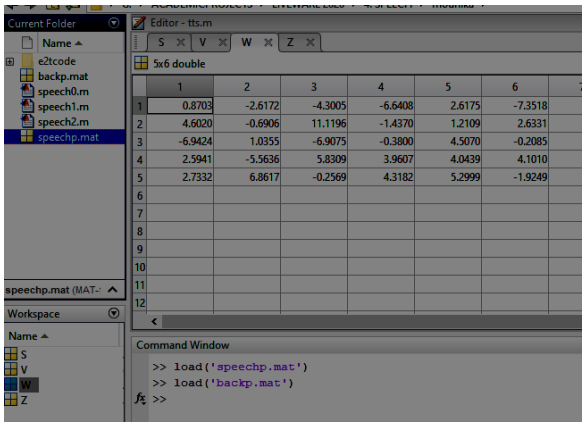

Fig 6: Speech Features extraction Results

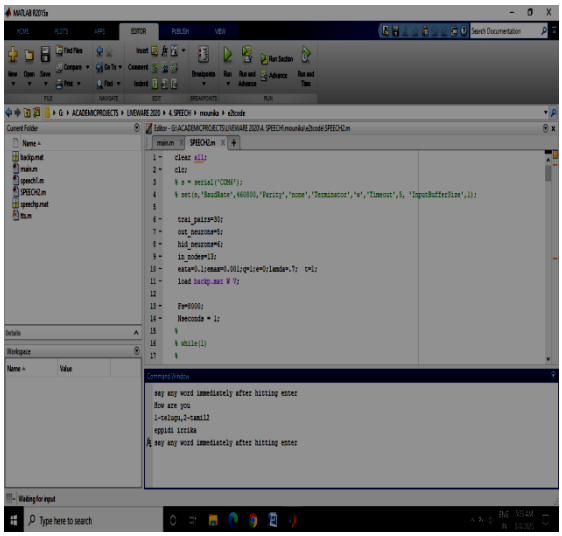

Fig 7: Speech Recognization Results

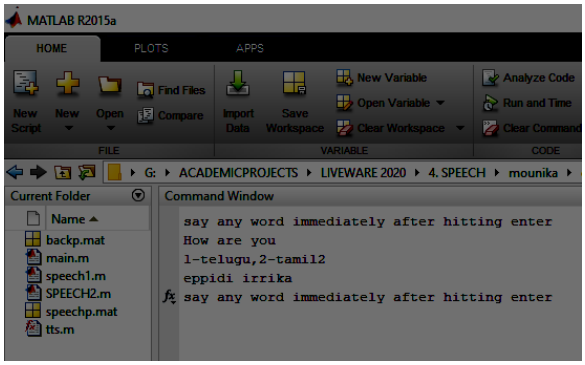

Fig 8: Speech to Speech (English→Tamil) Results

VI. CONCLUSION

Automatic speech recognition, translating of spoken words into text, is still a challenging task due to the high viability in speech signals. Deep learning is becoming a mainstream technology for speech recognition and has successfully replaced Gaussian mixtures for speech recognition and feature coding at an increasingly larger scale.Future tasks include improving upon and optimizing

current technology. Collaborating with more and more countries and their language to form a global consortium, share research activities among the multilingual communities and making it accessible and usable for the target audience are the next milestones for this consortium.

REFERENCES

[1]. Yan Zhang, Andrew Ng, Speech Recognition Using Deep Learning Algorithms,2017.
[2]. Dong Yu, Li Deng, Speech Recognition - A Deep Learning Approach, Microsoft Research, ISBN 978-1-4471-5779-3.
[3]. Graves, Alex, A-R. Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
[4]. G. E. Hinton, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". Signal Processing Magazine, IEEE, 29(6):8297, 2012.
[5]. Mohamed, A., Dahl, G. and Hinton, G. "Acoustic modeling using deep belief networks", IEEE Trans. Audio, Speech, & Language Proc. Vol. 20 (1), January 2012.
[6]. Riccardi, G. Hakkani-Tur, D. Active learning: Theory and applications to automatic speech recognition. IEEE Trans. Speech Andaudio Process. 2005, 13, 504–511.
[7]. Xiong, W.; Wu, L.; Alleva, F.; Droppo, J.; Huang, X.; Stolcke, A. The Microsoft 2017 Conversational Speech Recognition System. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
[8]. Wessel, F.; Ney, H. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. IEEE Trans. Speech Audio Process. 2005, 13, 23–31.
[9]. Windmann, Haeb-Umbach, R. Approaches to Iterative Speech Feature Enhancement and Recognition. IEEE Trans. Audio Speech Lang. Process. 2009, 17, 974–984.
[10]. H.C. Wang, S. R. Fussel, and D. Cosley, "Machine Translation vs. Common Language: Effects on Idea Exchange in Cross-Lingual Groups," Proc. ComputerSupported Cooperative Work (CSCW 13), ACM, 2013, pp. 935–937.