International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Real-Time Transcription Language Translation Using WebRTC

Kartikeiya Rai[1], Riya Rai[2], Kasturi Dive[3], Amrit Kadbhane[4], Rajeshwari Mahesh Goudar[5]

[1,2,3,4,5]School of Computer EngineeringMIT Academy of Engineering Pune, India

Abstract

Real-time communication (RTC) enables web browsing to support services like social media, video confer- encing, and unified communication. Users can interact with video and audio content using cloud infrastructures that ensure high-quality service. However, implementing multipoint video conferencing is challenging due to the lack of interoperability between proprietary protocols and codecs. WebRTC (Web Real- Time Communication) is a cutting-edge, open technology that allows real-time audio, video, and data communication in web browsers using JavaScript APIs, without the need for plugins. This paper introduces a peer-to-peer (P2P) video conferenc- ing system based on WebRTC. The proposed system uses a WebRTC-based communication channel, along with HTML5 and a Node.js server, for high-speed data transmission. Our experiments demonstrate WebRTC's scalability for live videoconferencing in browsers. Additionally, we explore the challengesin language translation.

Keywords: Real-time video translation, Speech-to-text tran- scription, Machine translation, Multilingual communication, We- bRTC, Natural language processing (NLP)

INTRODUCTION

In recent years, rapid advancements in deep learning and neural network technologies have driven significant progress across multiple domains, including speech processing, nat- ural language processing (NLP), real-time communication, machine translation, and video captioning. These technolog- ical breakthroughs have reshaped how we communicate and process language, particularly in real-time applications. This paper provides a comprehensive overview of recent research in these areas, highlighting key contributions and advancements. In the field of NLP, significant progress has been made in addressing the challenges posed by small text datasets. By combining techniques such as Word2Vec, FastText, and Transfer Learning, researchers have developed algorithms that deliver remarkable accuracy in NLP tasks, even with limited data availability. These advancements offer new opportunities

Identify applicable funding agency here. If none, delete this. for NLP applications across domains where data scarcity is a major constraint. In terms of real-time communication, protocols like WebRTC have been extensively explored, fo- cusing on message scheduling and timing behaviors in time- critical applications. Researchers have provided valuable in- sights into ensuring reliable communication, particularly in browser-to-browser video conferencing systems. The design and evaluation of WebRTC have facilitated scalable, efficient video communication applications, responding to the growing demand

for seamless real-time interaction over the web. Ma- chine translation has also seen substantial progress, with the development of multilingual speech translation systems that allow for real-time translation across different languages. By integrating robust speech recognition, machine translation, and speech synthesis technologies, handheld translation devices have been created to break down language barriers, offering solutions in areas like tourism, international business, and cross-cultural communication. The motivation for this project stemmed from a personal experience during a conversation with a friend who doesn't understand Hindi. It was difficult for us to communicate effectively, and that challenge inspired the idea of developing a system that could provide real- time transcription and translation during a video call. This system aims to bridge the language gap by enabling one user to speak in their native language while the recipient receives a transcription in their preferred language, making communication seamless across language barriers. Further- more, advancements in video captioning using Transformer Networks show potential for generating natural language descriptions for videos. These models, built using attention mechanisms and sequence processing, open up possibilities for content-based recommendation systems, human-robot in- teraction, and accessibility services. This paper explores recent trends and advancements in speech processing, NLP, real-time communication, machine translation, and video captioning. By reviewing key contributions in these fields, we gain insights into the emerging technologies and identify potential areas for future research and development. The intersection of these domains holds promise for creating intelligent, inclusive, and adaptable systems that can transform the way we communicateand interact in a multilingual world.

OBJECTIVE

The primary objective of this project is to develop a real-time language translation system that enables seamless communication between users speaking different languages. Specifically, the project involves developing a user-friendly interface that supports real-time voice input, processing the audio to convert it into text, translating the text into the target language, and displaying the translated text to the recipient. This will be achieved through the integration of WebRTC for real-time audio communication and Natural Language Processing (NLP) for accurate language translation and tran- scription. Future extensions may include support for additional languages and advanced features such as contextual translationimprovements and enhanced user interaction capabilities.

RELATED WORK

In recent years, various deep learning models such as CNNs, RNNs, LSTMs, GRUs, and RecNNs have been applied to a wide range of NLP tasks, including text classification, sentiment analysis, and machine translation. CNNs, originally designed for image processing, have proven effective in cap- turing local dependencies in text, making them well-suited for tasks like text classification. On the other hand, RNNs and their variants like LSTMs and GRUs excel in sequence prediction tasks due to their ability to manage long-term dependencies, which is critical for tasks such as language modeling and translation. RecNNs, known for their ability to handle hierarchical structures, have shown effectiveness in tasks like syntactic parsing. The mathematical foundations of these models are explained in detail, linking them to specific NLP applications. This technical exploration is set within a historical context, highlighting how deep learning has transformed NLP by enabling automatic representation learning from large datasets, moving beyond traditional rule- based and statistical methods [1]. Combined with advance- ments in cloud-based systems, such as the WebRTC-based audio/video conferencing system

implemented on virtualized clouds[2], these deep learning approaches are paving the way for more adaptive, scalable, and cost-efficient solutions in fields like real-time communication, machine translation, and other collaborative applications. Together, these advancements underscore the evolving landscape of NLP and real-time systems, offering new possibilities for innovation in various domains.

A real-time machine translation system that focuses on trans- lating English audio to Indian languages using Neural Machine Translation (NMT) and Long Short-Term Memory (LSTM) networks. The system is designed to address India's linguistic diversity by providing automatic translations without the need for human interpreters. The proposed system captures English audio, processes it through an LSTM-based encoder-decoder model, and outputs translated text and audio in an Indian language. OpenNMT, an open-source toolkit, is used to build the NMT system, which improves on previous translation methods like Rule-Based and Statistical Machine Translation. The authors claim their approach offers better accuracy and efficiency compared to older models, with results stored in a text file and played as audio to the listener. The system's architecture consists of a two-layer LSTM with 500 hidden units for both encoding and decoding, supported by a parallel corpus of English and Hindi sentences for training and valida- tion. The system's performance is evaluated based on transla- tion accuracy, with promising results in terms of meaningful translation and improved accuracy over previous models. The paper highlights the challenges of handling large datasets and suggests future improvements in this area. Overall, the paper provides a solid foundation for real-time speech-to- speech translation in multilingual environments, with a focus on leveraging advanced neural models to overcome limitations of earlier machine translation approaches.[3]

Talking about the approach to reducing latency in video conferencing systems. The primary focus is on minimizing the end-to-end delay in the video signal chain, specifically within the video codec. By leveraging sub-frame based data flow instead of full-frame processing, the authors propose a design that reduces the overall latency from typical industry standards of 33 milliseconds per frame to as low as 2 milliseconds. The hardware implementation is based on Texas Instruments' DaVinci series (DM816x) devices, where the introduction of novel mechanisms such as pre-fetching video data, entropy engine optimizations, and sub-frame level packet transmission helps reduce processing latency and improve video call quality. The key components in video conferencing, such as video capture, encoding, network transmission, decoding, and dis- play, each contributing to overall latency. In typical settings, a latency of up to 150 milliseconds is required for a high-quality user experience. Their design mitigates the limitations of con- ventional hardware, which assumes full-frame availability, by breaking the video stream into smaller, manageable sub-parts (sub-frames), allowing for faster processing and transmission. The proposed solution demonstrates a significant improvement in video conferencing latency, enhancing the user experience by reducing delay, thus achieving smoother and more natural face-to-face communication.[4]

The speech-to-text-to-speech (STS) pipeline presents a promis- ing alternative to traditional voice codecs, particularly in scenarios requiring high data compression and resilience to packet loss. By converting speech into text for transmission and then synthesizing the text back into speech at the receiver's end, the STS approach significantly reduces bandwidth usage while maintaining acceptable levels of intelligibility. In tests comparing the STS pipeline with standard PCM codecs under varying packet loss conditions, the STS method demonstrated comparable intelligibility, achieving word error rates (WER) close to those of uncompressed audio, even in challenging network environments. This indicates the feasibility of using automated speech transcription and synthesis for communi- cation in bandwidth-

constrained or high-packet-loss settings, offering substantial savings in data transmission without critically degrading the quality of communication.[5]

The evolution of Natural Language Processing (NLP) in Requirements Engineering (RE), from early efforts to generate models from requirements documents to current applications like defect detection, traceability, and classification. They em- phasize the growing relevance of NLP due to the rise of deep learning techniques such as BERT, which has proven useful for various RE tasks, including classification and named-entity recognition. Despite advancements, there is a gap between research and industry adoption, as only a small percentage of companies utilize automation in RE processes. The paper outlines the need for a more holistic understanding of how NLP can enhance RE, drawing from a mapping study of 404 papers to provide insights into available techniques, tools, and future directions. It advocates for greater integration between NLP advancements and RE practices, suggesting the potential for technologies like BERT to transform RE tasks where data is scarce. The briefing targets both researchers and practitioners, aiming to bridge the gap between research innovations and practical applications in the RE field.[6] This paper presents a detailed explanation and implementation of a Text-to-Speech (TTS) synthesizer that utilizes Natural Language Processing (NLP) and Digital Signal Processing (DSP) technologies. The TTS system converts input text into synthesized speech and allows users to save the speech output as an mp3 file. Developed using C# and the .NET framework, the system is divided into two modules: the main application module, which handles the graphical user interface (GUI), and the conversion engine, which processes text into speech. The interface is user-friendly, enabling text input, file imports, and various controls for speech output such as volume, speed, and the ability to save the output in audio format. A key strength of the TTS Gramaty system is its simplicity and usability, making it accessible for users with varying technical backgrounds. The paper outlines the complete text processing pipeline, in- cluding text-to-phoneme conversion, prosody generation, and waveform synthesis. By integrating features like adjustable speech rate, volume control, and the ability to load external text files, the system enhances user interaction and flexibility. However, the use of American English as the default language engine limits its application to non-English users, though the authors acknowledge this and suggest future developments formultilingual support.

While the system shows promise, the paper could benefit from quantitative metrics or user feedback to validate its perfor- mance in real-world scenarios. Although the authors describe the system as performing excellently, data to support this claim would make their findings more robust. Additionally, while the paper introduces a basic TTS system, more advanced features, such as machine learning-based prosody improvements or natural-sounding speech synthesis, were not explored. Future work could include extending the system to support multiple languages and real-time web-based synthesis, which would greatly expand its usability and impact.[7]

The proposed Hindi-English speech-to-speech translation (S2ST) system demonstrates a comprehensive approach by integrating automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis. The ASR system, using HMMs, effectively converts speech into text, with performance evaluated through word error rate (WER), showing a high accuracy in recognizing travel-related expressions. The MT subsystem employs statistical methods to translate between languages, utilizing IBM alignment mod- els and statistical tables to gauge translation accuracy via translated edit rate (TER), highlighting challenges such as alignment errors and the need for more precise modeling. The TTS system, based on HMMs, synthesizes speech from the translated text, with performance

measured by mean opinion score (MOS), indicating room for improvement in generating natural-sounding speech. Overall, while the system performs well in recognizing and translating basic travel expressions, enhancements in alignment accuracy, translation modeling, and synthesis quality are necessary to achieve more fluent andaccurate speech translation.[9]

A thorough examination of WebRTC technology and its prac- tical implementation for video conferencing between browsers over various network types, including LAN and WAN. It highlights the core components of WebRTC—such as Me- diaStream, RTCPeerConnection, and RTCDataChannel—and addresses the challenges associated with signaling and peer- to-peer communication. By utilizing the WebSocket protocol through Node.js, the study effectively demonstrates the estab- lishment of real-time video communication, with performance evaluations covering CPU usage, bandwidth consumption, and Quality of Experience (QoE). The findings reveal that CPU usage ranges from 13% to 17%, and bandwidth consumption remains efficient at 48-54 kbit/s, reflecting good quality in audio and video streaming. The paper underscores the practical benefits of WebRTC in enabling browser-based video con- ferencing without additional plugins while identifying areas for future research, such as developing scalable signaling mechanisms and comparing WebRTC with traditional VoIP protocols.[10]

an innovative approach to addressing the challenges of train- ing Natural Language Processing (NLP) models on limited datasets, particularly focusing on generating high-accuracy results despite sparse training data. It combines techniques such as Word2Vec and FastText for word embedding, along- side transfer learning, to create a custom algorithm capable of processing and understanding English language structures. The methodology involves preprocessing input data to extract key sentiments and synonyms, enabling the model to map sentences to trained classes effectively. By leveraging a large unlabeled dataset for pre-training and a manually curated labeled dataset for specific tasks, the approach aims to en- hance NLP model performance without the need for extensive training data, making it particularly relevant for applications like chatbots and digital assistants.[14]

While analyzing the impact of machine translation (MT)
and speech synthesis on speech-to-speech translation (S2ST) systems, which consist of speech recognition, MT, and speech synthesis components. It highlights that while various inte- gration techniques for speech recognition and MT have been proposed, the role of speech synthesis is often overlooked. The authors conducted a subjective evaluation to assess how the fluency and adequacy of translated sentences affect the naturalness and intelligibility of synthesized speech. Results indicated a strong correlation between MT fluency and both the quality of speech synthesis and the intelligibility of the synthesized output, suggesting that improved fluency in trans- lation enhances the overall performance of S2ST systems. The study emphasizes the importance of considering speech synthesis in the development of more effective S2ST systems, advocating for a holistic approach to integrating all three components.[16]

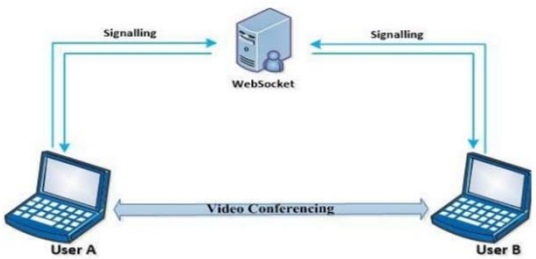

Fig. 1. WebSocket Server Architecture

LANGUAGE COMPONENT AND NLP APPLICATION

Natural Language Processing (NLP) is a pivotal branch of artificial intelligence that bridges the gap between human communication and computer understanding. By enabling machines to interpret, generate, and respond to human language, NLP facilitates a myriad of applications that enhance interactive experiences and streamline information processing. As shown in Fig.2 NLP encompasses several linguistic components that collectively enable machines to comprehend and manipulate human language.

LANGUAGE COMPONENTS:

1. Phonetics: Phonetics studies the sounds of speech and their physical production, essential for accurate speech recognition.
2. Lexis: Lexis refers to the vocabulary of a language, including individual words and expressions used incommunication.
3. Grammar: Grammar provides the structural rules for forming sentences, ensuring clarity and correctness in language processing.
4. Semantics: Semantics focuses on understanding the meaning behind words and sentences in context.
5. Discourse: Discourse deals with how sentences are structured in conversation or text to form coherent Communication.

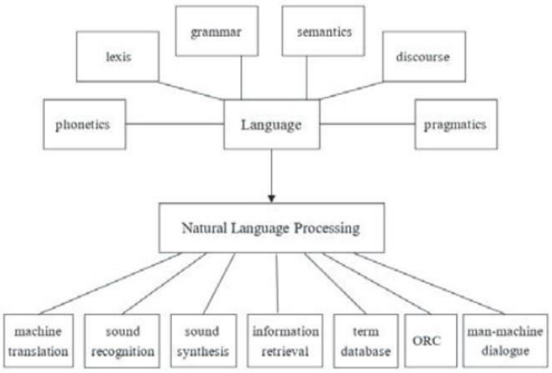

Fig. 2. Language Component and Application of NLP

6. Pragmatics: Pragmatics interprets language in context, considering speaker intent and situational factors.

APPLICATION OF NLP:

1. Machine Translation: Automatically translates text orspeech between languages in real time.
2. Speech Recognition: Converts spoken language into writtentext for processing.
3. Speech Synthesis: Converts written text into natural-sounding speech output.
4. Information Retrieval: Finds and retrieves relevant information based on user queries from large datasets.
5. Term Database: Stores specialized terms for consistent usein translation and technical fields.
6. Optical Character Recognition (OCR): Converts images oftext into machine-readable, editable text.
7. Man-Machine Dialogue: Enables interactive communication between humans and machines through natural language.

LIVE SPEECH WITH ITS TRANSCRIPTION The challenge of synchronizing live speech with its cor-
responding transcription at a syllabic level in real-time. The motivation stems from the need to generate audio books in unstructured languages like Thai, where traditional synchro- nization at the utterance level is insufficient due to the lack of clear word boundaries and punctuation in written text.[8]

A. Problem Statement and Motivation

- Unstructured Language Challenge: Thai language does not use spaces or punctuation marks consistently to denote word or sentence boundaries, making it difficult to synchronize audio and text at the utterance level.
- Application Need: Generating DAISY format audio books requires precise synchronization to allow users to navigate to specific words or phrases within the text.
- Existing Methods: Previous research has primarily fo- cused on synchronization at the utterance level, which is inadequate for unstructured languages where utterances can vary greatly in length.

B. Proposed Solution

- Syllable-Level Synchronization: The authors propose an algorithm that synchronizes live speech with its transcrip- tion at the syllabic unit, providing a finer granularity necessary for languages like Thai.
- GReal-Time Processing: The algorithm is designed to work in real-time, processing live speech input without significant delays.

C. Future Work

- Improved Landmark Detection: Incorporating additional features or refining the detection algorithm could reduce errors related to lax vowels and sonorant consonants.
- Applicability to Other Languages: While focused on Thai, the method is adaptable to other languages that require fine-grained synchronization.
- Robustness Testing: Evaluate the algorithm under vari- ous conditions, including different speakers, accents, and background noises.

TRANSLATION CATEGORIES

As discussed in the Introduction, representing characters (numbers, symbols, and letters) with numbers is called char- acter encoding. Multiple languages around the world require different character representations. Fortunately, all characters can be encoded into UTF(Unicode Transformation Format)-8 Unicode. UTF-8 is a variable byte sized encoding scheme that can represent up to 4 bytes or 4,294,967,296 characters andis the most widely used encoding scheme for Web pages.

TABLE I TRANSLATION EXAMPLE

Translation Source	***Description Text***
Spanish Source Text	”Abuela, ¿por que tienes los ojos tan grandes?” Caperucita Roja pregunto. ”Para que yo pueda ver mejor,” Dijo la abuela. ”¡Oh, abuelita, ¿por que tienes la boca tan grande?” ”Para poder comerte mejor!” Entonces, la abuela salta de la cama.
Correct English Transla-	”Grandma, why do you have such big eyes?” Little Red Riding Hood asked. ”So that I can see better,” the grandma said.

tion	”Oh, Grandma, why do you have such a big mouth?” ”So I can eat better!” Then, the grandma jumps out of the bed.
Google Translate Result	”Grandma, why are your eyes so big?” Little Red Riding Hood said. ”So I can see better,” said the grandmother. ”Oh, Grandma, why have the big mouth?” ”To eat better!” Then the grandmother jumps out of bed
Microsoft Translator Result	”Grandmother, why have such large eyes?” Little Red Riding Hood asked. ”So I can see better,” said the grandmother. ”Oh, grandmother, why have the big mouth?!” ”To be able to eat better!” Then, Grandma jumps out of the bed.
Yahoo Babelfish Result	”Grandmother, why you have the so great eyes?” Red Caperucita asked. ”So that I can see better,” the grandmother said. ”Oh, grandma, why you have the so great mouth?” ”In order to be able comerte better” Then, the grandmother jumps of the bed.

In Table I, we present example source texts used to evaluate various translation services, along with their corresponding translated descriptions for comparison.

TABLE II TRANSLATION CATEGORIES

Category	***Description***
Exact	Translated sentence is exactly the same as the correct translation.
Alternate	Translated sentence conveys the same meaning but with different words or order.
Different	Legitimate translation but does not convey the same meaning.
Wrong	Forms a sentence but cannot be interpreted as the correct translation.
Ungrammatical	Grammatically deficient translation.

Table II categorizes translation quality into five types: Exact, Alternate, Different, Wrong, and Ungrammatical, each with specific descriptions of their accuracy and grammaticality. This classification provides a framework for evaluating the effectiveness and reliability of translations. Using these categories, Table III summarizes the performance of three machine translation services—Google Translate, Microsoft Translator, and Yahoo Babelfish—across these five categories. It provides a count of how many translations from each service fall into these categories based on six tested sentences, allowing for a direct comparison of the accuracy and grammatical quality of the translations produced by each service.

TABLE III MT PERFORMANCE FOR EXAMPLE

Category	**Google Trans- late**	**Mi- crosoft Trans-lator**	**Yahoo**
Exact	0	1	0
Alternate	5	3	2
Different	0	0	0

Wrong	0	0	2
Ungrammatical	1	2	2

DISCUSSION

The experiments demonstrated that real-time language trans- lation in video conferencing significantly enhances communi- cation between speakers of different languages, making it an effective tool for multilingual interactions. The high translation accuracy achieved shows the potential of integrating NLP with video conferencing to bridge language gaps. However, the observed latency highlights the need for further optimization, particularly in wireless and WAN environments, to improve real-time performance.

The user experience feedback underscores the practical ben- efits of the system, although it also points to areas for improvement, such as handling idiomatic expressions and reducing latency. Future developments should focus on re- fining translation algorithms to handle more complex lan- guage constructs and optimizing system performance under varying network conditions. Overall, the integration of real- time translation into video conferencing systems holds promise for facilitating global communication and collaboration, with ongoing advancements expected to enhance its effectiveness and reliability.

CONCLUSION

Integrating real-time language translation into video con- ferencing marks a major leap in multilingual communication, crucial for business, education, healthcare, and diplomacy. This advancement enables participants to communicate seam- lessly in their preferred languages, fostering greater accessi- bility and collaboration across diverse linguistic and cultural backgrounds. While progress has been significant, challenges such as translation accuracy, latency, and cultural nuances per- sist. Ongoing improvements in machine translation and video conferencing technologies are essential to overcoming these issues and enhancing global connectivity and understanding.

REFERENCES

1. Megha Gupta, Shailesh Kumar Verma, Priyanshu Jain “Detailed Study of Deep Learning Models for Natural Language Processing”
2. Sunghyun Yoon, Taeheum Na, and Ho-Yong Ryu “An Implementation of Web-RTC based Audio/Video Conferencing System on Virtualized Cloud”
3. Raj Vyas, Kirti Joshi, Hitesh Sutar, Tatwadarshi P. Nagarhalli “Real Time Machine Translation System for English to Indian language”
4. Mihir Mody, Pramod Swami and Pavan Shastry “Ultra-Low Latency Video Codec for Video Conferencing ”
5. Rafael Dantas, Dr Chris Exton, Dr Andrew Le Gear “Communications using a speech-to-text-to-speech pipeline”
6. Alessio Ferrari, Liping Zhao, Waad Alhoshan, ‘NLP for Requirements Engineering: Tasks, Techniques, Tools, and Technologies”
7. Partha Mukherjee, Soumen Santra, Subhajit Bhowmick, Ananya Paul, Pubali Chatterjee, Arpan Deyasi “Development of GUI for Text-to-Speech Recognition using Natural Language Processing”
8. Nat Lertwongkhanakool, Proadpran Punyabukkana, Atiwong Suchato “Real-time Synchronization

of Live speech with Its Transcription"
9. Mrinalini K, Vijayalakshmi P "Hindi-English Speech-to-Speech Trans- lation System/or Travel Expressions"
10. Naktal Moaid Edan, Ali Al-Sherbaz, Scott Turner "Design and Evalua- tion of Browser-to-Browser Video Conferencing in WebRTC"
11. Satoshi Nakamura "Development and Application of Multilingual Speech Translation"
12. Kai Jiang, Xi Lu "Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review"
13. Jaideep Rao, Neil Daftary, Aditya Desai, and Prof. Reeta Koshy "Algo- rithm for using NLP with extremely small text datasets"
14. Jaideep Rao, Neil Daftary, Aditya Desai, and Prof. Reeta Koshy "Algo- rithm for using NLP with extremely small text datasets"
15. Jan Chorowski, Ron J. Weiss, Rif A. Saurous, Samy Bengio "On using backpropagation for speech texture generation and voice conversion"
16. Kei Hashimoto, Junichi Yamagishi, William Byrne, Simon King, Keiichi Tokuda "An analysis of machine translation and speech synthesis in speech-to-speech translation system"
17. Zhaorong Zong, Changchun Hong "On Application of Natural Language Processing in Machine Translation"
18. Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita, Yoshinori Sagisaka "A Purely Monotonic Approach to Machine Translation for Similar Languages"
19. Kiran Kumar Guduru, Sachin Dev "WebRTC Implementation Analysis and Impact of Bundle Feature"
20. Aditya Jain, Gandhar Kulkarni, Vraj Shah "Natural Language Process- ing"