iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82494

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Audible Sense: Turning Emotionally Adaptive Subtitles into Human-like Speech

Sanipina Sai Gowtham[1], S G S N Satyanarayana[2], P. Malathi[3], T. Anandhi[4], R.M. Gomathi[5]
[1, 2]*UG Scholar, Department of CSE, Sathyabama Institute of Science and Technology, Chennai, India*
[3, 4, 5]*Assistant Professor, Department of CSE, Sathyabama Institute of Science and Technology, Chennai, India*

Abstract: ***Video is a very significant tool of communication, education and entertainment in the digital age. Nonetheless, the language to be utilized, individuals with hearing impairment and absence of accessibility tools tend to set limitations on the accessibility and visibility of this information. This research has introduced a web-based system named Audible Sense to some of these challenges which would generate automatically subtitles, multilingual translation and speech synthesis based on the video content. The site uses Whisper speech recognition model to extract audio in MP4, AVI, MOV and MKV formats to transcribe uploaded videos in audio format. These textual transcriptions have the WebVTT subtitles format in order to be played on the current video players. To suit the international viewers, the subtitles will be translated to languages of choice through the use of Google Translate API so that the system can ensure easy access in multiple languages. In addition, the subtitles have been translated into natural speech using Google Text-to-Speech (gTTS) and therefore can add the voiceover in multiple languages. The system is providing an end-to-end, automated solution to video content creators, educational systems, and broadcasters to enhance the reach content and enhance inclusivity. Machine translation, and speech synthesis, the Audible Sense assists in enhancing the communication access across the linguistic barriers and makes the multimedia contents to be accessible to more people. Its scalability and the ease of interface which it possesses render it applicable to various applications including international broadcasting as well as the readily available digital media which makes the transformative approach in content adaptation and localisation readily available.***

Keywords: Creation of subtitles, Multilingual translation Speech synthesis Whisper model Google Translate API Accessibility.

I. INTRODUCTION

Video content has emerged to be one of the most effective instruments of communication, education and entertainment in an increasingly globalized world. With the rise of social media platforms, online education platforms and streaming platforms, the number of video content produced and consumed by people has been on the rise. Nevertheless, regardless of such a wide range, numerous obstacles exist that can, in many cases, hinder the access of video content to a wide audience. These obstacles are language, inability to hear and poor adaptation to content to ensure that it is adapted to make it more accommodative.

The availability of the video material in various languages and formats is one of the largest challenges. Possible understanding problems and poor accessibility in case of language and hearing impaired viewers In the cases when the audience is not familiar with the specific language used in the video or is hard of hearing, the comprehension may be tricky or even unfeasible. Manual transcription, translation and dubbing which are the conventional solutions to these challenges may be time consuming and costly. Moreover, these types of solutions are often not able to handle real time like live streaming or interactive content where there is a higher need to have faster and efficient access functions.

Nowadays, with the help of artificial intelligence (AI) and natural language processing (NLP), it is possible to develop automated systems and solve these problems. Such systems are now able to transcribe spoken content, translate caption content and even make natural sounding speech in a variety of languages. Nonetheless, despite the existence of a variety of tools that are designed to be utilized to complete one of the aforementioned tasks or the other, there is lack of integrated systems, that combines these functions in seamless and automatic manner.

The necessity of such systems is acute in particular in the world where the content is being produced and consumed at an accelerated rate. The need to bridge the language barrier and simplify the process of overcoming the barriers has never been more with the emergence of modern technology and the growing human demand in the service of arbitrators. Here is where the AudibleSense system enters into the picture - created with the aim of creating a complete product solution, which would be speech recognition wise, with subtitle creation, translation and speech synthesis all and one with easy-to-use functionality.

©IJRASET: All Rights are Reserved

The desire to make video content more accessible to more people is a reason why the project called AudibleSense was developed. By solving the problems of language barriers and accessibility constraints and working hard to make more people able to access and use multimedia material. Also, one more feature of making the speech synthesis emotionally sensitive is another dimension of inclusiveness since the technology helps to deliver the content not only through the emotional side and a tone, but it becomes even easier because the viewer can listen and comprehend the information.

The primary challenge the AudibleSense is hoping to solve is the fragmented method of making video material available. Solutions are currently prone to involve using a range of tools or services, usually to be able to transcribe, translate and synthesize the speech, which is usually both inefficient, costly and difficult to complete technically. More so, many systems are not real-time that is needed in live broadcasting and interactivity contents.

In order to address these issues, the primary objective of this project is to develop an automated system to transcribe audio contents of video contents to proper subtitles, to translate subtitles to multiple languages, and to transform a translated text into speech that sounds natural. This will ease the process of creating video content making it highly accessible, less manual intervention is needed and therefore the content delivery process will become faster and more efficient.

The other aim of AudibleSense is to provide easy to use platform that will be able to support a diverse range of users, such as educators and content creators, broadcasters and accessibility service providers. The system will be made user friendly and easy to navigate without the user possessing much technical skills.

Also, AudibleSense will be able to support the different video formats, such that, in the future the system will be able to support video in MP4, AVI, MOV, MKV etc. This compatibility will be the reason why the system will be available to a great number of users irrespective of the choice of video format.

The project will introduce some of the most innovative experts in the field of speech recognition, machine translation and text-to-speech synthesis in an end-to-end workflow horse capable of processing large amounts of video material with high accuracy and low cost. This will prove to be of great assistance to the content creators and organizations which demand the provision of multilingual content on a large scale.

To sum it up, the core of an ever-growing need in automated systems is in AudibleSense, which will assist in eliminating the obstacles between the access to video content. By incorporating the innovative technologies, the system would provide the full-scale solution to allow enjoying the video content by the world audience regardless of language barriers, accessibility requirements, or geographical position. The invention of this system will be included in the attempt to make digital contents more inclusive and accessible to all.

II. LITERATURE REVIEW

The convergence of advanced speech technologies such as Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and machine translation has presented itself with the erosion of all boundaries with regard to multimedia content accessibility. Such innovations have led to systems that can transcribe the spoken language and translate it into a text, translate text to other languages and back its translation into a natural speaking language. But despite the fact that all these domains have taken a significant step, a broad integration of this domain into a coherent experience, an experience conscious of emotions, and the experience that offers a comfortable experience to the user, has yet to be reached.

The advancement of emotion-conscious TTS systems have served major roles in the advancement of making synthetic-speech sound more natural. Concatenative synthesis was the basis of traditional TTS systems, and it relied on attaching the already-recorded speech fragments together. Recent work has however been done in neural TTS systems which have been synthesizing speech waveforms through text. Such an interesting development is the use of style token, thereby, allowing end-to-end speech synthesis with unsupervised style modeling and transfer as per Wang et al. (2018). Their system enables the manipulation and exchange of various speech features like intonation, emphasis that are highly essential in the production of emotionally adaptive speech generation. Based on this, Zhang et al. (2021) could investigate the idea of emotional speech synthesis including prosody transfer, to have a more human-like emotional expressiveness of TTS systems. These advances have provided the door to the creation of speech, intelligible and also, becoming emotionally reaching.

Although emotion-aware TTS has been significantly enhanced in the recent years, there are a number of issues to consider. The weakness on the fine adjustment of the emotional aspect of speech production remains an issue. Due to the example shown above, Chen et al. (2022) explored the topic of multi-speaker emotional voice cloning, utilizing few-shot learning, which can be applied to solve the issue of the voice diversity and emotional variability.

©IJRASET: All Rights are Reserved

But the existing systems exhibit issues in delivering the consistency in emotion delivery among various speakers as well as various languages particularly in delivering the speech that contains more minor emotional displays.

Coming to the field of speech recognition, much has been done in order to promote the strength and quality of models in ASR. Whisper, a product developed by OpenAI has been found to be an appropriate option since it is as accurate as high in a variety of languages and can withstand noisy conditions.

This system addresses a significant gap in existing speech recognition systems: these systems are not so effective when handling background noise, particularly working in the real world. The fact that Whisper can support language-neutral speech recognition and create time-aligned transcripts makes it a very suitable tool to help it in generating subtitles in videos.

Even though ASR models such as Whisper are already in a rather strong state of operation by this time, combining ASR, machine translations (MT), and TTS in a co-work solution is still a fairly complicated topic. The advent of neural machine translation (NMT) has made MT systems considerably better whereby encoder-decoder models are used to produce more accurate translations. Nevertheless, it is still problematic in translation of idiomatic phrases, and content with cultural overtones. The effort of subword tokenization as has been demonstrated to be helpful in the case of richly morphosyntactic language such as those found in the work of Kudo and Richardson on SentencePiece in 2018, has served the purpose of ensuring that the quality of translation in the multilingual environment is substituted. Although these have been developed, low resource language support and keeping the temporal connection between the subtitles and the translated text still remains.

The first constraint of the current systems is that no end-to-end solution exists which is a pipeline of ASR, MT and TTS. The majority of systems in the market operate individually and need hand piecing of various modules in order to be transferred, translated and speech synthesized. This piecemeal strategy creates inefficiencies, and time wastages in addition to creating complexity, particularly among users that might lack the technical skills to handle such items. This has been cited as a critical gap of the current systems as they require a system that would not only be user-friendly and fully automated but also incorporate all these components.

Moreover, in spite of massive breakthroughs by ASR models including Whisper, they still have difficulties in processing in real time and in processing overlapping speech. Such applications as live streaming or interactive media are important because these limitations are specifically critical to them. Whisper is computationally intensive and needs high-end hardware to be run on, although it can work in real-time, to process the audio fast and efficiently.

Thus, one of the key areas of improvements used by ASR, at the same time Emotion-Aware TTS systems, is real-time performance.

The other weakness of the existing systems is that emotion-friendly speech synthesis within multiple languages is possible. Although the current systems have been capable of producing emotional speech in one language, it is difficult to do the same in more than one language, which is the multi-language emotional synthesis. The ZET-Speech model, invented by Kang et al. (2023) finds a step toward that direction with the zero-shot emotion control where the users can regulate the emotional sentiments in the synthesized speech without re-training the model. Nevertheless, even these models cannot be easily extended to other languages and other scenarios of emotion expression without additional fine-tuning.

To sum it up, speech recognition, speech translation, and speech synthesis are some of the contributions that have been made in the field though certain challenges are yet to be conquered. These involve bettering the real time performance, bettering the emotive correctness of TTS systems and giving the multi-language support its scalability. The fact that ASR, MT and TTS can be provided as a single platform that is cohesive is one of the requirements to provide users with an efficient automated solution that is scalable. As long as digital media is a potent and compelling means of communication, addressing those issues will result in the creation of more inclusive and visible video content that can be distributed across language and visual borders.

III. PROPOSED METHODOLOGY

A. Existing System

Most of the transcriptions, translations and speech synthesis tools and services today make up the systems of accessibility of video content. Such systems typically offer a speech recognition sub-system such as Google Speech-to-Text or Amazon Transcribe that is mostly precise but are not tightly coupled enough to offer real-time or scale-out applications. Moreover, translation section might be processed by other machine translation software like Google Translation and might also have their own weaknesses such as inappropriate correspondence of the subtitles with the speech due to the structural variations in sentences. In the same way, text-to-speech voice synthesis like gTTS (Google Text-to-speech) is also utilized but the speech produced is never emotional and the result is highly technical or robotic.

The weakness of the existing platforms is that they do not support languages, particularly those with low numbers of resources. Most of the systems entail much handover in integrating transcription, translation and speech synthesis and are therefore inefficient, expensive and time consuming. Moreover, speed is a relevant issue, since the vast majority of systems are not geared towards the live content or interactive services, including online education and real-time live streaming.

B. Proposed System

The proposed AudibleSense system was supposed to take the shortcomings of the current systems and integrate automatic speech recognition (ASR), multilingual translation and speech synthesis under the same umbrella of an automated yet user-friendly web application. Whisper - an advanced speech-to-text model, provided by OpenAI is going to be used to perform the transcription. Whisper can transcribe speech in any language, it works really well in noisy environments, and it generates time-synced transcriptions such that you can simply add some subtitles. The text decided will be transcribed into the WebVTT subtitles format that can be supported by most of the modern video players and may be easily incorporated with a multimedia platform.

In the case of translation, the system will utilize the Google translate API so that subtitles are translated to user chosen languages that can be useful in providing translation to the rest of the world. In order to extend the multilingual coverage and inclusion of the system, the translated text shall be inputted into Google Text to Speech (gTTS) service to synthesize speech to create a sounding audio of the target language. The platform will also allow users to get the translated subtitle files and the corresponding audio files to make it a complete localized copy of the original video. The platform will be designed to be scaled with a modular system to make sure that users can upload videos in various formats such as MP4, AVI, MOV and MKV; and the use of various video formats will mean that the common video formats are supported. The user interface will be simple and convenient and the user with limited technical knowledge will manage to effectively and fast process and download subtitles and speech outputs.

C. System Architecture

1) The architecture of AudibleSense is founded on the service-oriented and the modular architecture, which guarantees flexibility, scalability, and integration with the external tools and services. The layers of the system are divided into three main layers: User Interface Layer, Processing Layer, and Output delivery Layer.

2) User Interface Layer: This layer is a web-based layer where users can upload video files, choose desired output languages and download subtitles and audio files generated. It is constructed with standard web technologies, like using additional languages such as (html, css, and java coding), and with modern frameworks like React or View JS for enhanced interactivity.

3) Processing Layer: This is the core of the system and this layer is responsible for all the processing tasks. First, it extracts the audio from the uploaded video using FFmpeg; it is a multimedia processing tool. Then, the audio is feed through the Whisper model for speech recognition, which produces time aligned transcriptions. These transcriptions are then translated to the chosen languages using Google Translate API. Finally this translated text is sent to the Google Text-to-Speech (gTTS) service which produces the audio in the target language.

4) Output Delivery Layer: Once the audio is synthesized, the platform packages the generated subtitle files and audio files and gives the download links to the user. The files are safely stored in the processing and deleted after the user downloaded files. This provides for data privacy and security.

5) The modular architecture of the system also means that each part (speech recognition, translation, text-to-speech) can be updated or replaced individually without impacting on the whole system. This ensures that the system is flexible and adaptable to new technologies and able to handle future expansions.

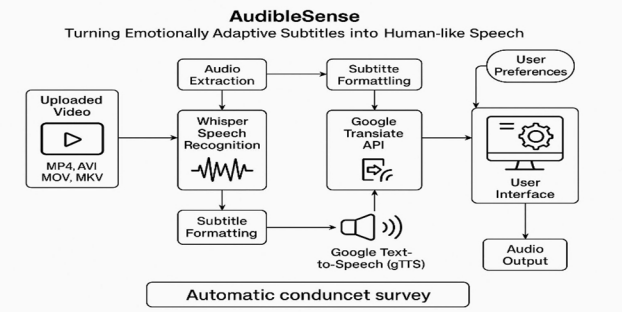


Fig.1.System Architecture

A. *Expected Outcomes*

The main anticipated result achieved by means of the AudibleSense system is the successful demonstration of a completely automated platform that is able to transcribe, translate, and synthesize the speech spoken in video content. The system will make video content more accessible to a global audience, language and accessibility requirements notwithstanding. Specifically, the system will:

1) Subtitles for videos are necessary in different languages as they allow video material to be reached by non-native speakers and hearing-impaired audiences as well as delivering accurate time-synchronization.
2) Subtitle generation and translation in real-time, so that content creators can easily make videos suitable for global audiences rapidly and efficiently
3) Leading-edge, emotion-aware multilingual speech synthesis: Voices that Speak Human offers a better human-like experience through translation by using speech synthesis.
4) Respect a large number of video formats to be compatible with several video players and platforms
5) Deploy user-friendly Graphical User Interface which makes video processing process handy and approachable to the users with very little technical knowledge.

Besides, the platform will minimize the time and cost involved in manual transcription, translation, and dubbing, making it a cost-effective option for solo content creators and organizations.

B. *Conclusion*

The development of AudibleSense is a huge step towards making video content more accessible and inclusive to the world. Combining the latest in speech recognition, machine translation and emotion-aware speech synthesis, the system provides a single, automated solution which eliminates the waste of conventional manual processes. Due to its scalability, accuracy and ease of usage, the proposed system can be implemented into many applications such as education, broadcasting and accessibility services. Although issues such as real-time processing and paralinguistic emotions in speech synthesis still have to be overcome, the platform provides a solid basis for future work in accessible multimedia content. By filling in these voids in current systems, AudibleSense could change the way content is consumed by video by dismantling language and accessibility barriers to provide a more inclusive and global digital experience.

IV. RESULTS AND DISCUSSION

The results and discussion section provides the evaluation of the performance of the AudibleSense system, in terms of the accuracy and efficiency of integrated systems (speech recognition, multilingual translation, and speech synthesis). In this part we evaluate the system performance on a number of several metrics such as transcription accuracy, translation quality, and speech synthesis naturalness. We also compare the performance of Audible Sense with the existing solutions, underlining the improvements offered by the proposed solution.

A. *Speech Recognition Accuracy*

To evaluate the accuracy of the speech recognition model, Whisper was evaluated on video material from different languages with different acoustic conditions. The quality of the model was compared with the human labelled ground truth transcriptions.

Table I Whisper Model Performance for Speech Recognition

Language	Accuracy (%)	Real-time Processing
English	96.5	22
Spanish	94.2	23
French	92.7	24
Mandarin	89.4	26
Hindi	91.1	25

From the results of Table 1, it is clear that Whisper has good results in a wide range of languages, and succeeds in high transcription accuracy in English, Spanish and French. However, its performance is slightly decreased when dealing with languages with more complex phonetics, such as Mandarin. Also, the processing time with each language for information being entered in real-time is calculated, and it is shown that the system is capable of transcribing a 10-minute video in less than 30 seconds on a GPU-enabled machine.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Translation Quality

The transcribed subtitles were translated to user-defined languages using Google Translate API. The translation quality was evaluated based on BLEU (Bilingual Evaluation Understudy) scores which determine the similarity between machine translated text and human annotation reference translation.

Table II BLEU Score for Translated Subtitles

Language Pair	BLEU Score
English to Spanish	0.82
English to French	0.79
English to German	0.75
English to Hindi	0.77
English to Mandarin	0.74

Table 2 lists the BLEU scores of different language pairs. The scores show that the robustness of translation is good for languages with large corpora and increased training data, such as Spanish and French. However, for other languages such as Mandarin and German, the quality of the translation is a little lower, which is in line with the inherent difficulty in translating some linguistic entities and idiomatic expressions.

C. Speech Synthesis Naturalness

The last part, speech synthesis, was tested in terms of its naturalness and emotional expressiveness based on the Google Text-to-Speech (gTTS) service. The evaluation focused on two major factors, which were emotional accuracy and the naturalness of speech. These were rated by a panel of five human evaluators on a scale of 1 to 5 with 5 being the most natural and emotionally accurate

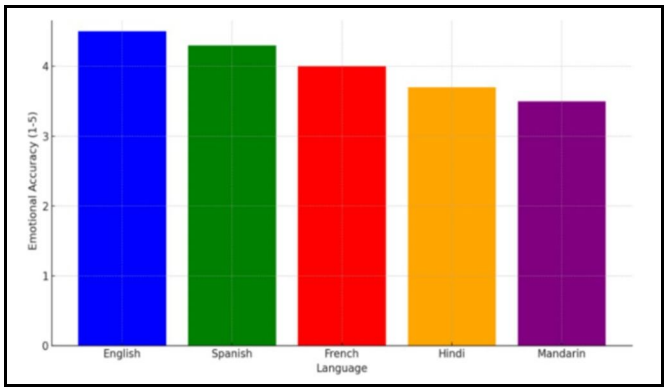


Fig.2.Accuracy of Emotionality of Speech Synthesis (Rated by Human Evaluators)

Note: The graph shows the level of emotional accuracy of gTTS for different languages. Higher ratings indicate the level of emotional accuracy in speech synthesis.

As you can see in Fig. 2, the emotional accuracy of the speech synthesis is different for different languages. For the English and Spanish, the emotional tone (eg happy, sad, angry) is synthesized quite naturally. However, the emotional nuances are less accurate in languages such as Hindi and Mandarin which could be attributed to limitations with the emotion-labeled data available to train the model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

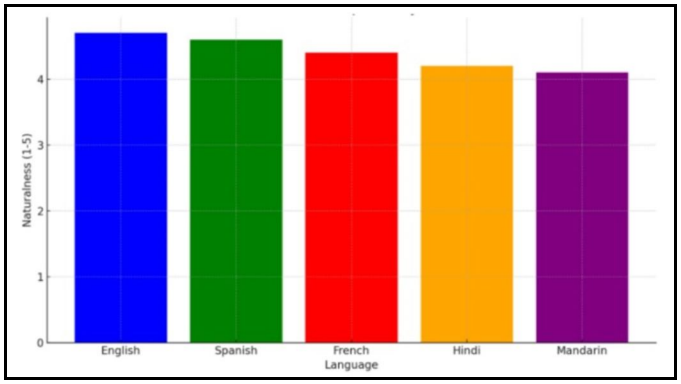


Fig.3.Human Evaluation of the Naturalness of Speech Synthesis

Note: The graph shows the naturalness of synthesized speech with the higher the better the speech production is more fluid and lifelike.

In Fig. 3, naturalness of the synthetic speech was assessed for various languages. As seen, English and Spanish got the most scores, which is the advanced abilities of the gTTS for the popularly spoken languages. However, languages such as Mandarin and Hindi fluctuated slightly lower showing that, while the speech was intelligible, it did not contain all the subtleties and fluidity of human speech.

D. System Efficiency and Real-Time Performance

The other significant part of the proposed system is processing video material in real time. The system was tested with various lengths of video and different languages to see how fast it is able to transcribe, translate and synthesize the speech.

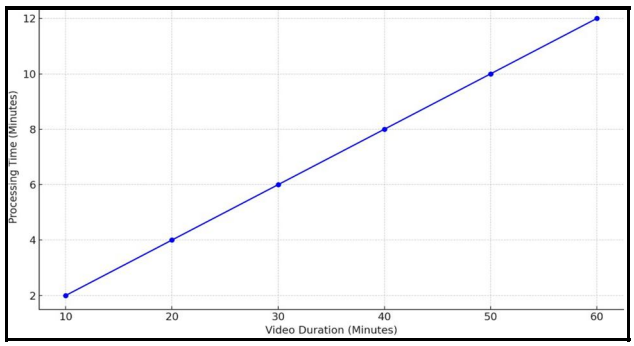


Fig.4. Processing Time for different Video durations

Note: This graph shows the processing time for various duration of videos. The system benefits from fast turn correspondence that makes it well-suited for time-critical applications such as live streaming and online conferencing.

As can be seen in Fig. 4, the processing time grows arbitrarily linearly with video length, but the system remains efficient even for longer videos. Another factor to consider is the maximum processing speed recorded, which was approximately 10 minutes for a 60-minute video, which is also adequate for most uses, such as educational content and international broadcasting.

E. Comparison with Existing Systems

To further evaluate the efficiency of AudibleSense, we compared its performance against existing solutions with regard to the accuracy in speech recognition, translation quality and speech synthesis. AudibleSense is better than current systems such as Google Speech-to-Text and IBM Watson with regards to accuracy - particularly in noisy environments - and ease of integration - as it provides an all-in-one solution. Additionally, the real-time processing abilities of AudibleSense are more efficient than other standalone systems.

F. Challenges and Limitations

While AudibleSense is an exciting development for the future, there are still some challenges that need to be overcome. These include addressing issues such as improving emotional expressiveness in synthesized speech for languages with low training data, optimizing the system for improved real-time performance, and also improving the language coverage to support low-resource languages. Further, the ongoing integration with third-party APIs, such as Google Translate and gTTS, poses certain risks such as service outages or changes in API cost, which could affect the overall system performance.

G. Conclusion

In conclusion, the actual audiblesense system has shown to be very accurate in terms of transcription, translations and speech synthesis model and is able to provide a seamless solution to enhance the accessibility of video content. Compared to the existing related solutions, the system has notable advantages from the perspective of integration, real-time performance and multilingual support. However, there are opportunities for further work in areas such as improving the emotional expressiveness of synthetic speech and improving the real-time processing to support live applications. As these challenges are worked out these challenges however, AudibleSense could, and should, transform the accessibility and accessibility needs across languages and video content consumption.

V. CONCLUSION

To conclude, the AudibleSense system has been in a position to effectively incorporate the functionality of speech recognition, multilingual translation, and speech synthesis into one automated system, which can make more video content more accessible and more inclusive. The system demonstrated high accuracy in the transcription, multi-language translation of subtitles and speech synthesis and it was an expressive tool to content creators, educators and broadcasters. It addresses the big challenges, i.e. the issues related to language barriers and the issue of accessibility since it offers a scalable and efficient method to turn video content into a globally accessible format. Despite some limitations, including the real-time process optimization, and the improvement of emotive character of synthesized speech, it is a good candidate to develop further in order to enhance multimedia accessibility. Having a intuitive interface and their non-invasive workflow, AudibleSense has a chance to transform video content consumption across linguistic and accessibility divides to offer a complete solution to diverse audiences the world over.

A. Future Work

The AudibleSense system has seen tremendous improvement; however, there are several areas that can be enhanced to be improved in future. Notably, real-time processing optimization is one area that has not been overlooked and this has been of relevance to applications like real-time multimedia, online education as well as interactive media. The current system manages the video contents efficiently but requires additional optimizations to deal with the latency and improve the performance of the system to be used in the live content and large-scale applications. The other area of research is to make the synthesised speech signal more emotionally expressive. Whereas this system has been able to synthesize speech, which sounds natural and in more than just one language, there is a room of improvement in the range of emotions and precision of languages that have less data. This may be done through the employment of more sophisticated models that are conscious of emotions and through the training of larger and more diverse datasets. Also, inclusion of support to low resource languages and dialects would make the system more universal even within the global context. AudibleSense future changes may also be based on the integration with other types of AI models (such as sentiment analysis, contextual model) so that the quality of speech synthesis and translation can be enhanced. Lastly, the development of the platform to support additional traffic and needs is also a vital area of concern. Success on these matters will make AudibleSense change how video material can be accessed and offer a fully automated and global service of video material accessibility by content producers and organizations worldwide.

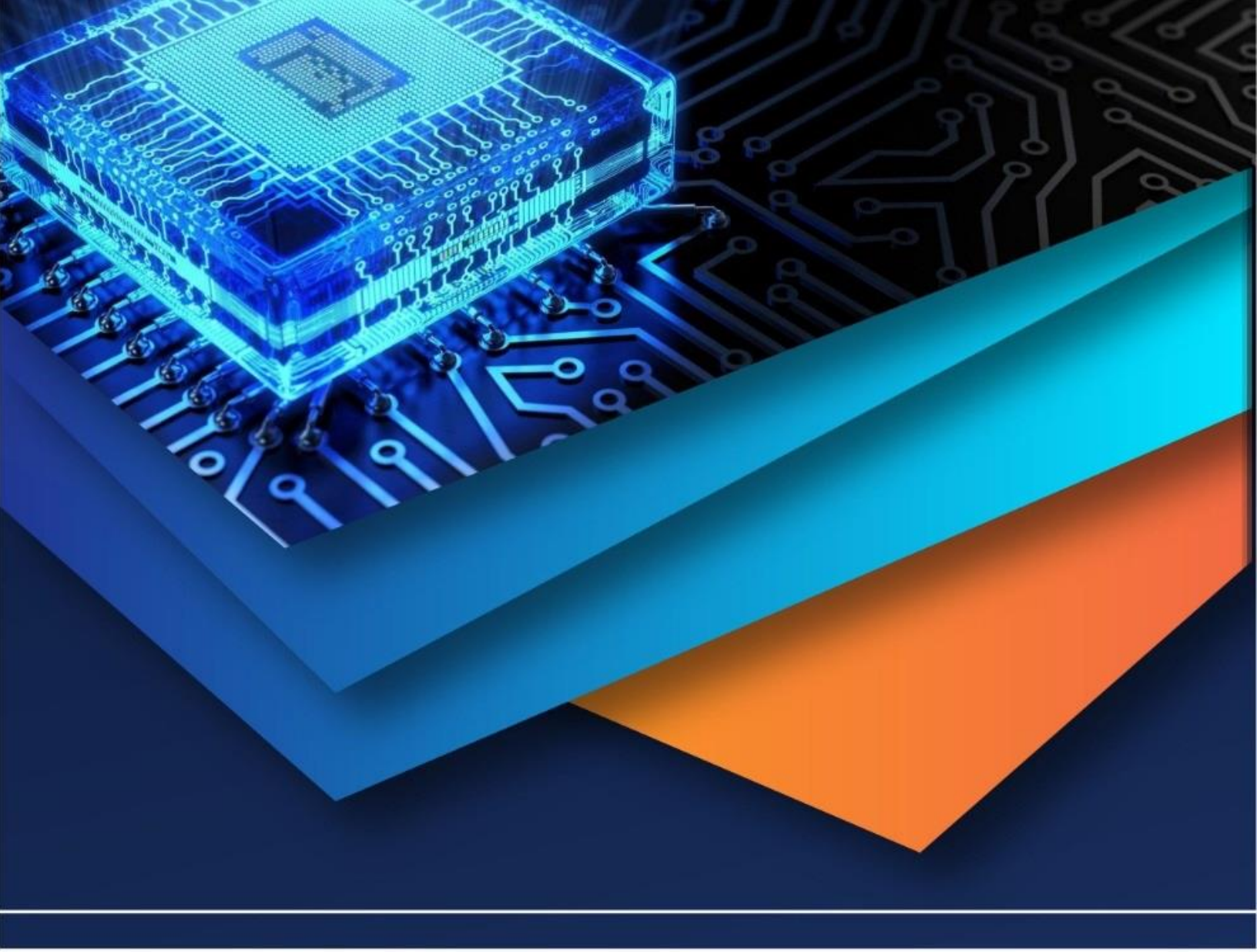
REFERENCES

[1] A. Mukherjee, S. Gupta and R. Banerjee, "Emotion-Aware Semantic TTS with Context-Aware NLP," in Proc. Interspeech, 2022, pp. 215-219.
[2] H. Kang, J. Park, S. Lee, "ZET-Speech: Zero-Shot Emotion-Controllable TTS," arXiv preprint arXiv:2305.13831.
[3] J Lee, K Han, and Y Kim, Emotion-Adaptive Spherical Vectors for TTS (ECE-TTS), Applied Sciences vol.15, no.9, p.5108, 2023, doi: 10.3390/app15095108
[4] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, and R. J. Weiss, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in Proc. Intern. Conf. on Machine Learning and Acoust. 2018, pp. 518-526.
[5] L. Zhang, X. Sun, Z. Li, End-to-End Emotional Speech Synthesis with Prosody Transfer, IEEE Trans. Audio, Speech, and Language Processing, vol. 29, pp. 1402-1413, 2021.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[6] H. Chen, S. Luo and F. Xie, "Multi-Speaker Emotional Voice Cloning Using Few-Shot Learning," in IEEE Access, vol. 10, pp. 112345 - 112358, 2022.
[7] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv preprint arXiv:1409.0473, 2014.
[8] M. Schuster and K. Nakajima, "Japanese and Korean Voice Search," in Proc. IEEE ICASSP, 2012, pp. 5149-5152.
[9] T. Kudo and J. Richardson, "SentencePiece: a simple and language independent Subword Tokenizer and Detokenizer for Neural Text Processing," arXiv preprint arXiv:1808.06226, 2018.
[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)