Towards Intelligent Libraries: Emotion-Aware Text-to-Speech for Adaptive User Engagement

Anita Gobinathan1, a), Nik Mohd Zarifie Hashim1, 2, b), Noor Ziela Abd Rahman3, c), Norazrai Daniel Afiq Razali1, d), Azahari Salleh1, 2, e) and Mohd Riduan Ahmad1, 2, f)

1Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia   
2Centre for Telecommunication Research and Innovation (CeTRI), Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

3Faculty Of Engineering & Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450 Bukit Beruang, Melaka, Malaysia

4Penerbit Universiti, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

*c) Corresponding author: ziela.abdrahman@mmu.edu.my*

a) b022110063@student.utem.edu.my

b) nikzarifie@utem.edu.my

*d) m122320012@student.utem.edu.my*

*e) azahari@utem.edu.my*

*f) riduan@utem.edu.my*

**Abstract.** The applications of smart technologies in contemporary libraries are changing the way users interact with them. This paper discusses the development of an Emotion-Aware Text-To-Speech (TTS) Synthesis that aims to improve adaptive user involvement in smart library and similar Context-aware environments. However, the proposed solution is different from traditional TTS systems with lifeless speech as it utilizes artificial intelligence (AI) and natural language processing (NLP) to understand textual content and user context and realize emotional feedback by modify tone, pitch and rate in the speech synthesis. Through the integration of deep learning models, the system detect emotions, including joy, sadness, or urgency, and modify speech outputs to provide users with a natural and engaging listening experience inside the library. This is a great feature for people that are blind, for language learners and for anyone that needs a little bit of help. We describe a prototype implementation of the model, report an experimental evaluation of it in the aspects of the naturalness of speech and the expression of emotion, and discuss its possible integration into digital library systems. Initial results show that the emotion aware TTS demonstrates better engagement and satisfaction compared to classic TTS. The results have implications in the dynamic AI-powered smart library development and can be used as a stepping stone to study human-centred AI in digital knowledge access.

# INTRODUCTION

The notion of smart libraries is becoming increasingly widespread to improve the user experience, turning environments into more interactive, attractive, and tailored ones. Here, intelligent technologies, such as speech and emotion recognition, are also in use. Utilizing progresses in artificial intelligence (AI) and natural language processing (NLP), these libraries purport to provide adaptable user experiences. The current work is within the affective computing perspective; to grasp and handle emotions in intelligent environments, such that more natural human-like experiences are achievable through technology. Because libraries are serving a broad spectrum of users, ranging from those who are visually impaired, learning a new language, or in need of some degree of personal assistance, emotion-aware technologies are crucial to make sure that everyone feels connected and welcomed in the digital environment. We present our work on developing an Emotion-Aware Text-to-Speech (TTS) system that aims to improve the adaptive user engagement for smart libraries [1]. Contrary to classical TTS systems that generate a robotic and emotionless voice, our system modulates speech tone, pitch and rhythm dynamically according to emotional state [2], textual content, and speaker. By employing deep learning-based emotion classification models (e.g., joy, sadness, urgency) the system has potential to deliver a qualitative auditory experience. The generation of emotional speech the emotivity of which can be perceived by humans is accomplished by using machine learned SER models which are superior with respect to data pre-processing, feature extraction and emotion classification aspects thereof. And more importantly, the system they designed employs state-of-the-art speech synthesis models, including WaveNet and Tacotron, which have been proved effective for synthesizing high-fidelity audio or spectrograms from text. The issue is how to integrate these speech synthesis models and emotion recognition to better obtain natural and emotionally intelligent speech output [3].

Although much progress has been made in speech synthesis and recognition, there remains an obvious gap to be filled in the integration of emotion-awareness into smart libraries. Conventional TTS applications cannot adapt to the emotions of the text, and the generated speeches may sound mechanical or not reflect the emotional color of text. Moreover, end-to-end TTS methods (e.g., Tacotron2) are accused of being brute-force and inconvenient (the output speech generation takes too long) when it comes for real-time applications. The combination of emotion recognition algorithms and WaveNet-based synthesis models has not been well studied, resulting in a lack of exploiting the potential to enhancing the user experience in intelligent environments [4]. In this paper, a hybrid model is presented that integrates emotion recognition and TTS synthesis for building a model that can provide a responsive and emotional resonant experience for users in smart library environment where limited resources are available.

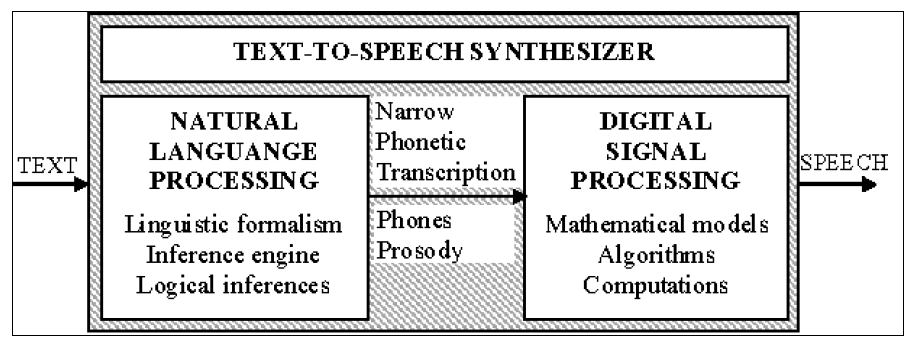
This paper makes a novel contribution to the AI-based smart library by introducing such an Emotion-Aware Text-to-Speech synthesis which has great potential in enhancing user engagement due to dynamical and emotionally adaptive speech it generates. It is an intensive analysis of the combination of emotion recognition algorithms and state-of-the-art speech synthesis methods, explaining deficiencies in TTS applications. Instead, the paper provides a number of case studies detailing prototype implementation and performance measurement of how emotion-aware systems in the digital library can support accessibility and engagement. The results itself add to an increasing knowledge of human-centered AI and are also a step towards novel developments in emotionally intelligent systems serving the respective needs of users in libraries.

# LITERATIRE REVIEW

## Text-to-Speech Synthesis (TTS)

Computers convert textual data to voice by way of Text-to-speech (TTS) technology. The TTS technology as a part of speech synthesis is one of the key technologies to improve human-machine communication. Generative model-based: For high-quality natural speech production, it is necessary to use model that can learn from audio datasets [3].

There are two major components for a TTS system, which are NLP and DSP. Speech synthesis through NLP The text conversion phase through NLP produces sound forms Sound forms are subjected to prosodic treatments (intended to avoid robotic speech) Following spoken code, the DSP transfers the processed data in native speech signals. The schematic diagram of TTS system is illustrated in Figure 1.



**FIGURE 1.** General functional diagram of TTS system

TTS is used in media applications such as radio and television as a common source of voiceover and dubbing news. A well-purposeful TTS system provides human-like speech and communicate clearly but also is ready to map the audio content in the movie industries, documentaries, and from other media forms through narration [5].

## Speech Emotion Recognition (SER)

Speech signals are among the most natural media in human communication, and they are also easy to measure in real-time. They are composed by paralinguistic information and linguistic message, and the former can carry the speaker’s affect. References to language variables have to do with human language and include message and settings. Conversely, para-linguistic components provide a quantitative measure of the differences in the rising tone of specific patterns in a language. The challenge in devising a speech emotion recognition system lies in spotting and extracting different speech features corresponding to diverse emotions. Feature extraction is crucial to the success of a classification and thus proper combination of appropriate audio features in SER task should be exploited. In literature, the following methodologies were used to detect emotion from speech and each of them used different speech features. LPCCs and MFCCs have been widely adopted for the recognition of speech emotion [6].

## Artificial Intelligence

An easy to use supporting artificial intelligence to automate both creation and categorization all image processing tasks and voice recognition. The application of the proper artificial intelligence technology to the smart library construction project in practice will lead to functional smart libraries and personalised proactive services. With the assistance of Text-to-Speech technology, artificial intelligence as a machines reads text and produces spoken language in human-like voices from complex algorithmic model which make use of authentic vocal inflections including stress and intonation and emotional elements to make speech more natural compared to previous TTS system [7].

## Natural Language Processing

The Text-to-Speech procedure heavily relies on Natural Language Processing, because with this kind of technology you are taking unprocessed text through a series of steps that will transform it into speech which makes sense, well performed and accurate. Language Processing techniques operate in much the same way as Automatic Speech Recognition systems to establish strong links between spoken and written language. The natural language processing works in two stages to bridge the gap between the machine comprehension and the way the humans communicate. Enhancement of TTS Tasks The performance of TTS tasks can be enhanced by using NLP, thus achieving accurate synthesis of text to speech and producing natural-sounding results [8 - 12].

# METHODOLOGY

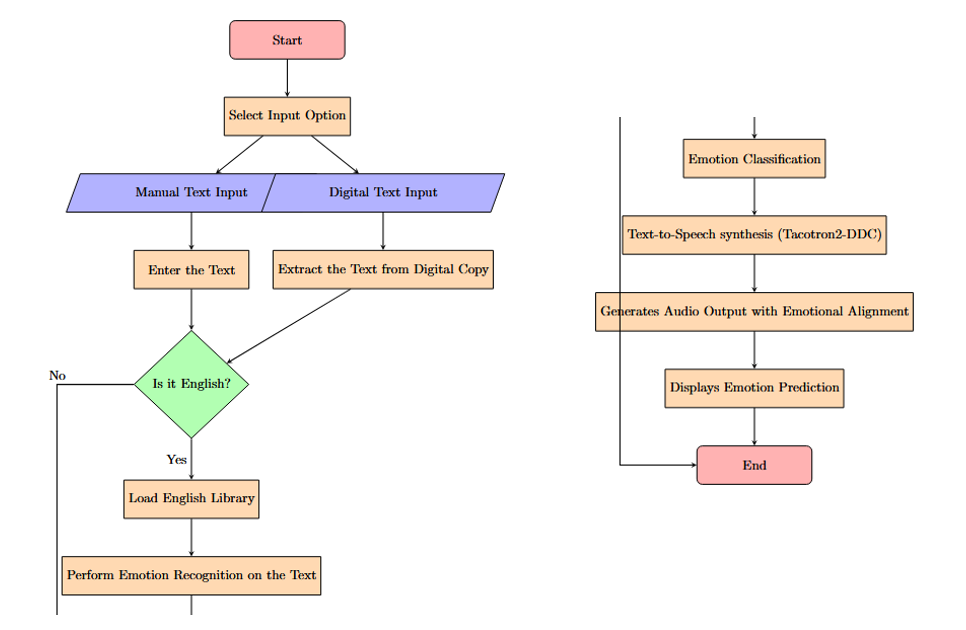
In this paper, we propose an emotion-aware TTS system that is built for capturing the text semantics and predicting the appropriate emotion of the user. In the context of smart library environment, such type of studies are faced with a great challenge. This approach leverages state-of-the-art emotion detection and expressive speech synthesis to provide users with a more engaging and personalized learning experience, as depicted in Figure 2. First, a lightweight version of BERT, DistilBERT is applied to analyse textual input for emotion detection. DistilBERT is fine-tuned on multiple emotion-labelled datasets such as the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) that covers a wide range of emotional affects such as happiness, anger, sadness and surprise. This serves to enable the system to effectively identify the emotional context of the input text. The emotion detection step determines the weight of given emotion emitted from the given sentence that gets transferred to the next stage, which is synthesizing voice.

As for the speech synthesis module, we use the Tacotron2-DDC model. This model, known for its high-quality output speech, is further trained to produce speech adapted to the detected emotions. The textual input is then passed through the Tacotron2-DDC system which manipulates parameters (pitch, tone, rhythm, prosody) according to the emotional type detected in the previous step. Integral parts of this proposal are that both emotion recognition and TTS synthesis are unified in a single processing pipeline. To allow user interaction, the web-based interface is built with Gradio for users to enter the text manually or upload documents. References The interface renders real-time emotion detection output and provides the playback and download of the synthetic emotional speech. This E2E system is developed to work in real-time, giving users a responsive solution to interact with smart library systems.

This system development is centred around the typical ML life cycle from problem formulation down to collecting data and preparing it. The emotion models and the synthesis system of speech are described and selected and put on the dipstick, with claims of facing with beneficial experiences. The system is further integrated with Flask web application framework, acting as the backend, for real-time processing. Finally, a comprehensive evaluation using performance measures and subjective ratings is performed to evaluate the emotion detection and speech synthesis adequacy. This complete approach paves the way for the development of emotion-aware technologies in intelligent environments e.g., smart libraries.

# RESULTs

The emotion-aware TTS system was carefully tested yielding promising results with the room for improvement. The emotional recognition module obtained a moderate correct rate of 43%. The system achieved the best results on part with “joy” for which it obtained very high precision and recall. Nevertheless, it struggled with the discrimination of subtle emotions, such as “disgust, “since only low training samples were available of these less-frequent emotions. Furthermore, the labels such as “fear” and “sadness” tended to be confused, as these emotions generally display similar language and tone, thus exhibit overlapping we have overlapping patterns in our training set. The performance of precision was inconsistent for the different emotions, and this is reflected in the low macro average precision of 0.25. These difficulties indicate that more advanced techniques in feature extraction, dataset balancing, and training model are required to enhance the system performance, especially in recognizing more challenging affective gestures.

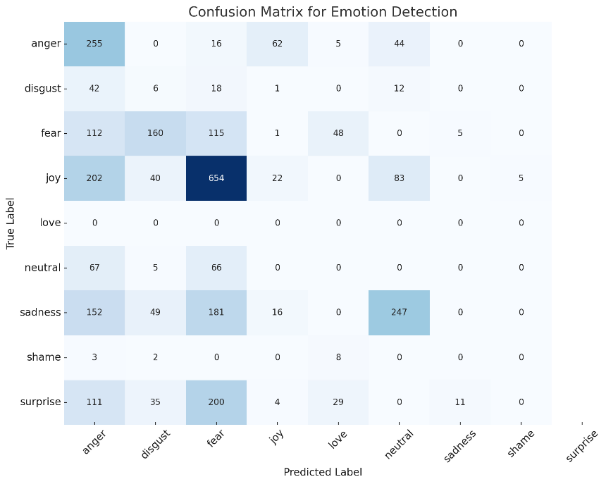


**FIGURE 2.** Overall proposed method

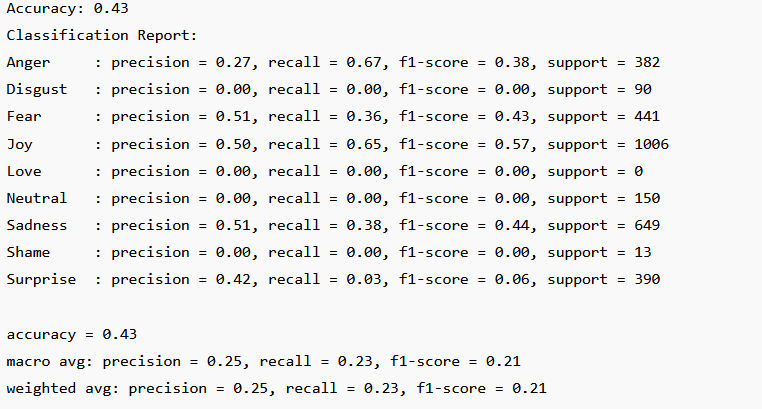
Regarding speech synthesis, the TTS module showed that it could generate high-quality speech that conveyed clear emotional expressions, particularly for emotions like “joy” and “anger.” The system adjusted pitch and tone effectively to match the emotional content of the text, making the speech more dynamic and expressive. However, softer emotions, such as “fear” and “neutral,” were less accurately represented, with the emotional intensity sometimes failing to align with the identified emotion. The prosodic adjustments made by the TTS model were sometimes too subtle, and the system struggled to convey the depth of these emotions. Despite this, the synthesized speech was still considered highly intelligible and natural. The user interface, developed with Gradio, was found to be intuitive and efficient, allowing users to input text easily and receive feedback quickly as in Figure 3. The system also offered flexibility in terms of document uploads, further enhancing its accessibility and usability. Across different emotions the precision varied widely which produced an overall low macro average precision score of 0.25, which is depicted in Figures 4 and 5. Better feature engineering methods coupled with improved dataset distribution strategies must be applied to increase system performance.



**FIGURE 3.** Proposed system’s interface

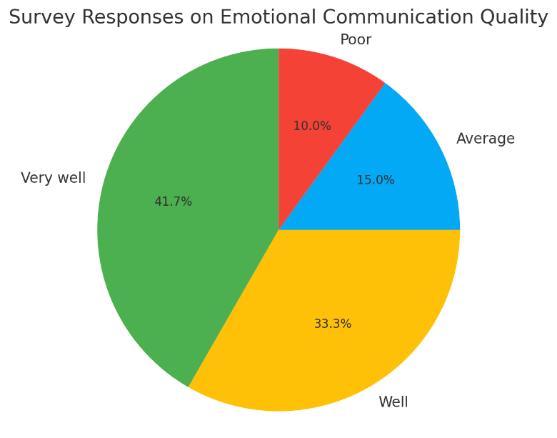


**FIGURE 4.** Confusion matrix



**FIGURE 5.** Performance metrics

User opinions were collected using survey, asking users to provide valuations of emotional communication and global properties of the system. In the Figure 6, 41.7% of the respondents rated the emotional delivery was “Very well,” and 33.3% was “Well. These ratings suggested that there is still scoped to improve subtlety and consistency of expression of emotion particularly of the more subtle emotions, If, on the one hand, the system is effective in conveying emotional speech on the other hand. Listeners also described infrequent artifacts like static/clip distortion in playback, but those were not considered significant artifacts. General feedback was good, and users perceived the design of the system to be user-friendly, with effective interaction between the emotion recognition and speech synthesis components.



**FIGURE 6.** Qualitative analysis study analysis

# CONCLUSION

In summary, the emotion-aware TTS system constructed in this paper shows very promising to improve user interaction in smart library. The addition of the emotional dimension to the textual user input text makes the information more personalized and facilitates a more entertaining and expressive speech synthesis. The emotion detection part well identified emotions like “joy” and “anger”, and we could express these emotions by changing pitch, tone, and rhythm of the synthesized speech. Nonetheless, difficulty remains in recognizing and portraying softer emotions such as “fear” and “neutral” and increasing the depth and regularity of expressing emotion during the generated speech. These constraints point out to the requirement of more studies regarding the improvement of the performance of EMODL, especially for the recognition of subtle and complex emotions. Notwithstanding the aforementioned issues, the interface and general system functionality were well-received, and the feedback suggested that the emotional communication was performed well, though there was a need to improve the realism of the representation of a broader variety of feelings. In the next stages, we plan to improve emotion recognition first, second to improve SSQ, and third to support other languages, and the final stage will be to run the real-time processing in large data and resource-limited devices. This report promises the way for the future development of emotionally intelligent technologies within smart libraries and other interactive spaces, and the applications could also be used in areas like healthcare, education and customer service where human computer interaction with emotion intelligence will be amped up. Such ongoing advances of emotion-aware TTS systems will further positively influence the evolution of the human-centered AI applications, which enable more authentic and sympathetic interactions between machines and human users.

# Acknowledgments

The author would like to express sincere gratitude to Universiti Teknikal Malaysia Melaka and Multimedia University providing the necessary facilities and support during the research process.

# References

1. Jonas Lea, “Text-to-Speech Market: Giving Voice to Technology and Advancements Solutions,” LinkedIn Article.
2. S. Madanian, A. Ahmadi, M. Assefi, S. K. Esfahani, and M. Rezvani, “Speech Emotion Recognition Using Machine Learning — A Systematic Review,” Intelligent Systems with Applications **20**, 200266 (2023).
3. J. Yu, Y. Yao, R. Feng, T. Liang, W. Wang, and J. Li, “A Review of the Text-to-Speech Synthesizer for Human-Robot Interaction for Patients with Alzheimer’s Disease.”
4. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural Speech Synthesis with Transformer Network,” Proceedings of the 33rd AAAI Conference on Artificial Intelligence, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pp. 6706–6713 (2019).
5. M. R. Hasanabadi, “An Overview of Text-to-Speech Systems and Media Applications,” Lecture Notes in Electrical Engineering **294**, 7–13 (2023).
6. S.-W. Byun and S.-P. Lee, “A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms,” Applied Sciences **11**(4), 1890 (2021).
7. K. Yu, R. Gong, L. Sun, and C. Jiang, “The Application of Artificial Intelligence in Smart Libraries,” Advances in Economics, Business and Management Research **100**, 708–713 (2019).
8. “What is Natural Language Processing (NLP)? Its Benefits, Challenges, Examples!” Shaip Blog.
9. U. R. Pol, “Hugging Face: Revolutionizing AI and NLP,” International Journal of Research in Applied Sciences and Engineering Technology **12**(8), 1121–1124 (2024).
10. Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face,” Advances in Neural Information Processing Systems **36**, 1–27 (2023).
11. “Emotion-Aware TTS — A Hugging Face Space by Anita-19,” Online Space.
12. Y. Lim, K.-W. Ng, P. Naveen, and S.-C. Haw, “Emotion Recognition by Facial Expression and Voice: Review and Analysis,” Journal of Informatics and Web Engineering **1**(2), 45–54 (2022).