PRE-EMPTIVE EMOTIONAL PROTECTION FOR EMOTIONAL LABORERS THROUGH REAL-TIME VOICE ANALYSIS AND MACHINE LEARNING

CHANG-MOOK OH

HelpU, Inc., Seoul, South Korea

ABSTRACT

This paper proposes and develops 'LINCALL,' a real-time voice processing and machine learning system designed to provide preemptive emotional protection for emotional laborers working in environments such as call centers. LINCALL analyzes voice data in real-time conversations with customers to understand their emotional states, converting this data into text for analysis. The system identifies keywords and offers appropriate questions and answers, aiding emotional laborers in handling customer interactions more effectively. It also includes a feature for responding appropriately to the customer's emotional state, such as voice blocking. Additionally, the automatic voice-to-text conversion feature reduces fatigue for emotional laborers, enhances the efficiency of consultations, and aids in preemptive emotional protection for emotional laborers.

KEYWORDS

Emotional laborers, Emotion analysis, Speech recognition, Pre-emptive emotional protection, Machine learning

1. INTRODUCTION

Call center agents typically experience high-stress levels due to emotional labor, work intensity, and burden. This pressure is further amplified as they must comprehend and respond appropriately to customers' emotions and demands in real-time. Agents face verbal abuse and sexual harassment from customers, resulting in accumulated fatigue and stress. Therefore, a service to protect and assist these agents is required [1].

Agents must understand customers' complex emotional states and demands in real-time and develop appropriate response strategies. A systematic approach is needed to accurately understand the customer's emotions and support and protect the agents.

The advancement of AI and machine learning is bringing innovative changes to customer service environments by combining real-time voice processing and emotion recognition technologies. The importance of emotion recognition in customer service has been extensively surveyed and highlights the growing need for sophisticated tools. Recent research has sought to understand emotions more accurately using machine learning algorithms and natural language processing technologies. Real-time voice processing technology allows applying these emotion analysis techniques to real-time conversations, enabling an understanding and appropriate response to the customer's emotional state.

David C. Wyld et al. (Eds): SIPR, NCWC, BIBC, SOFEA, DSML, ARIA, NLP, CSEN -2023
pp. 97-107, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131908

In this paper, we design and implement an emotion analysis-based support system for preemptive emotional protection of agents using these technologies. This system enhances interactions between customers and agents while reducing the emotional burden on the agents and improving the quality of customer service.

2. System Concept

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

2.1. System Concept and Configuration

Figure 1 shows the concept diagram of the LINCALL system. This system performs bidirectional emotion recognition through voice analysis and voice-to-text conversion. Depending on the severity of the emotions, it either issues a warning to the agent or automatically activates LINCALL to perform preemptive protection functions on behalf of the agent.

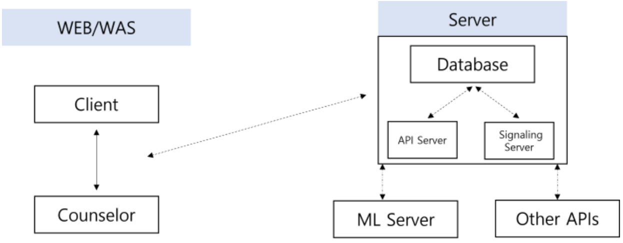


Figure 1. System Concept Diagram of LINCALL

The system comprises a voice analyzer, text converter, WebRTC-based conversation channel, and a media server.

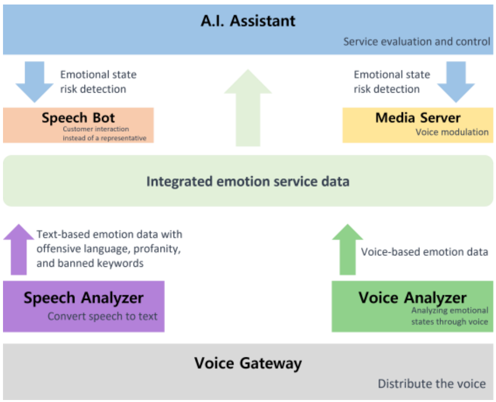

A.I. Assistant
Service evaluation and control
Emotional state risk detection
Emotional state risk detection
Speech Bot
Customer interaction instead of a representative
Media Server
Voice modulation
Integrated emotion service data
Text-based emotion data with offensive language, profanity, and banned keywords
Voice-based emotion data
Speech Analyzer
Convert speech to text
Voice Analyzer
Analyzing emotional states through voice
Voice Gateway
Distribute the voice

Figure 2. System Process Flowchart

When a customer initiates a consultation, an agent is selected, and mutual consultation begins. The agent and the customer converse via webRTC (web RealTime Communication)—both the agent and the customer record their voices, which are transmitted to a voice processing server. The server uses Speech-to-Text (STT) to convert the recorded voices into text and generate a conversation transcript. During the consultation process, the content of the conversation is transmitted to both the agent and the customer.

The system analyzes customer emotions based on voice and also through the voice converted into text. This allows the creation and refinement of emotion categories that can describe the state of emotional laborers, capable of expressing virtual emotional states such as happiness, anger, and calmness.

Through the agent monitoring feature, agents can see or remember the content of the consultation through text conversation records. When a customer's emotion changes and emotional consumption occurs, the voice is automatically converted into text. Auxiliary consultation support can be received through contents such as automatic questions and answers. These features can DB-ify a large amount of consultation information, train it, and provide optimal consultation techniques suitable for various situations. If additional features such as keyword extraction, word cloud generation, and emotion analysis are needed, it requests the results from the machine learning server and provides them to the user. Transcripts during the consultation process, emotional states, and keyword information are stored in the database so that the customer and agent can access the desired information even after the consultation is finished.

2.2. Main Features

2.2.1. Real-Time Transcript

The customer and the agent communicate via voice. The voice is converted into text to create a conversation record to quickly grasp any missed conversation during the consultation and support various other functions assisting the agent's talk. The STT (speech-to-text) technology stores the voice as text.

2.2.2. Voice Emotion Analysis

When the recorded file of the user's voice arrives at the server, audio and text emotion analysis are performed simultaneously. The weight of the audio emotion analysis is set at 40%, and the text emotion analysis is at 60%. The audio emotion analysis is developed using the Inception-v3 model for learning and training.

2.2.3. Text Emotion Analysis

Text emotion analysis was done using a library for Korean natural language processing tasks. The library uses state-of-the-art NLP models such as BERT and Transformer implemented in PyTorch to perform various NLP tasks, including emotion analysis. This library is used to perform emotion analysis on the Korean natural language, and the logit values of the output vectors before passing through the model's classification layer are used to generate the audio emotion analysis results. The final emotions of happiness, anger, and calmness are extracted through the weighted average of the audio and text emotion analysis results.

2.2.4. Keyword Generation

The keyword generation process uses an AI-based keyword extraction model and Python's WordCloud library. This field of Natural Language Processing (NLP) uses sentiment analysis features to summarize long sentences and extract general nouns from the results. The extracted nouns are classified into positive or negative keywords, which are transmitted to the server and stored in the database. The reserved keywords are sent to the word cloud generation API to generate the desired word cloud results.

2.2.5. Real-Time Call

The real-time call feature is implemented using WebRTC. WebRTC is a P2P communication protocol allowing web applications and sites to capture and stream audio and video media between browsers without an intermediary. This technology allows real-time calls between the agent and the customer. For real-time calls, signaling servers, session servers, and media servers are developed and built, and the coTurn server is constructed and used for data transmission in firewall or private network environments.

2.2.6. Voice Blocking

If a customer continuously or several times says negative sentences during a consultation, the service automatically blocks that voice. This uses voice recognition and emotion analysis technology. The blocked voice phrase is converted into text and displayed to the agent, and if necessary, the agent can unblock the voice-blocking function from the system. Blocked voices are stored in the machine learning server and used as learning data.

2.2.7. Providing the point when the customer became angry

It is easy for an agent to understand a customer's current emotional state, but it can be difficult to remember why the customer started to get angry during the consultation. This system analyzes the customer's emotions in real time to identify when the customer starts to get angry. This allows the agent to immediately understand the point and reason why the customer started to get angry and respond appropriately [2].

2.2.8. Intelligent Q&A

The intelligent Q&A feature includes question recommendation and automatic answer recommendation functions. During a consultation, the customer's speech is analyzed to provide recommended questions and answers according to the customer's speech. Recommended questions and answers are pre-registered in the database and continue to be updated. This function analyzes the customer's speech using the Universal Sentence Encoder. This allows the calculation of sentence-level semantic similarity and improves the performance of downstream classification tasks for untrained data.

2.2.9. Agent Monitoring

The agent monitoring feature calculates the agent's voice volume, the number of syllables per minute, etc., to regulate the agent's speech rate and size. If the voice suddenly gets louder or the number of syllables per minute increases, the system automatically notifies the agent to control their emotion and continue the consultation. This uses voice recognition technology and real-time data processing technology.

3. SYSTEM IMPLEMENTATION

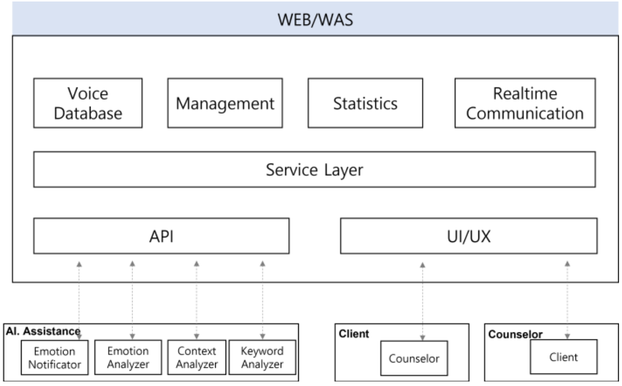


Figure 3. Detailed System Module Structure

Figure 3 shows the detailed module structure of the implemented system. Voice processing is divided into preprocessing and post-processing stages, including voice input, text conversion, conversation processing, voice blocking and conversion, and voice listening. For details on the implemented modules, please refer to Figure 4.

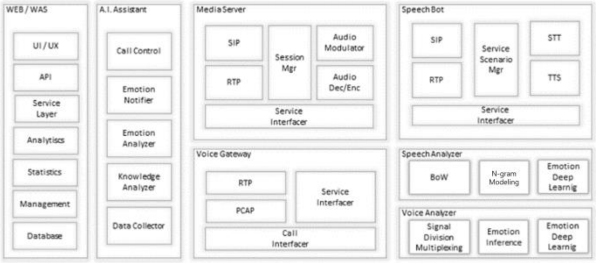


Figure 4. User Interface Processing Module Structure

The image below shows the implemented user interface. We provided and configured scenarios for the emotional labor protection service and managed them. Various features have been implemented, including the main screen seen by the customer and the consultant, the screen during customer consultation, the screen for checking customer consultation records, the consultant's main screen, and the consultation screen.

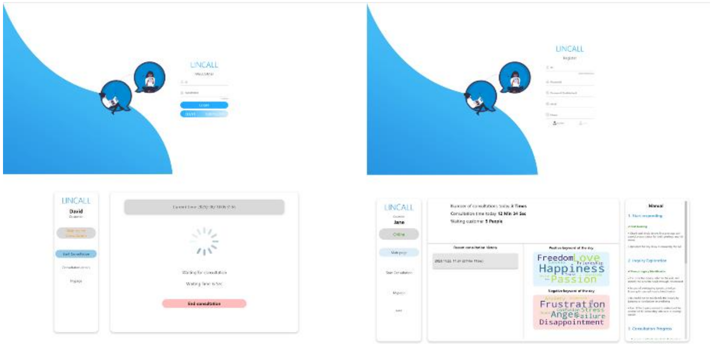
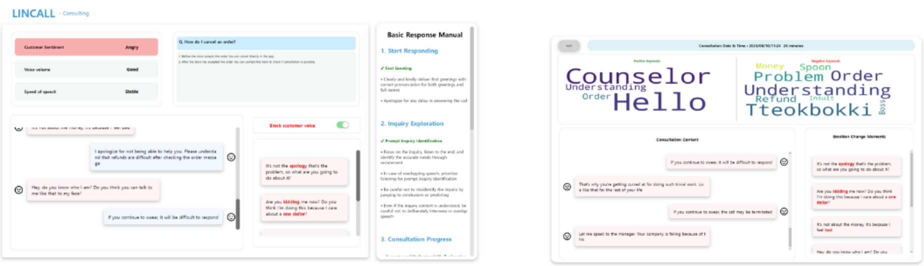


Figure 5. Overall Implementation Screen



LINCALL - Consulting

Customer Sentiment	Angry
Voice volume	Good
Speed of speech	Stable

Q. How do I cancel an order?

1. Before the store accepts the order You can cancel directly in the app.
2. After the store has accepted the order You can contact the store to check if cancellation is possible.

It's not about the money, it's because I feel bad

I apologize for not being able to help you. Please understand that refunds are difficult after checking the order message

Hey, do you know who I am? Do you think you can talk to me like that to my face!

If you continue to swear, it will be difficult to respond

Block customer voice

It's not the apology that's the problem, so what are you going to do about it?

Are you kidding me now? Do you think I'm doing this because I care about a one dollar?

Basic Response Manual

1. Start Responding

✓ First Greeting

- Clearly and kindly deliver first greetings with correct pronunciation for both greetings and full names
- Apologize for any delay in answering the call

2. Inquiry Exploration

✓ Prompt Inquiry Identification

- Focus on the inquiry, listen to the end, and identify the accurate needs through restatement
- In case of overlapping speech, prioritize listening for prompt inquiry identification
- Be careful not to misidentify the inquiry by jumping to conclusions or predicting
- Even if the inquiry content is understood, be careful not to deliberately intervene or overlap speech

3. Consultation Progress

Consultation Date & Time : 2023/08/10/11:24 24 minutes

Consultation Content

If you continue to swear, it will be difficult to respond

That's why you're getting cursed at for doing such trivial work. Live like that for the rest of your life

If you continue to swear, the call may be terminated

Let me speak to the manager. Your company is failing because of this

Emotion Change Moments

It's not the apology that's the problem, so what are you going to do about it?

Are you kidding me now? Do you think I'm doing this because I care about a one dollar?

It's not about the money, it's because I feel bad

Hey, do you know who I am? Do you

Figure 6. Screen during and after consultation

3.1. Attention Mechanism in Conversation Processing

The proposed system employs an attention mechanism to enhance the analysis and understanding of customers' speech in conversation processing. This technique allows the model to focus on specific input parts when producing the output, capturing the nuances of human conversation and contributing to the system's efficiency in real-time dialogue handling. The implementation follows the principles outlined in the seminal work 'Attention is all you need' [3].

4. System Structure and Technology

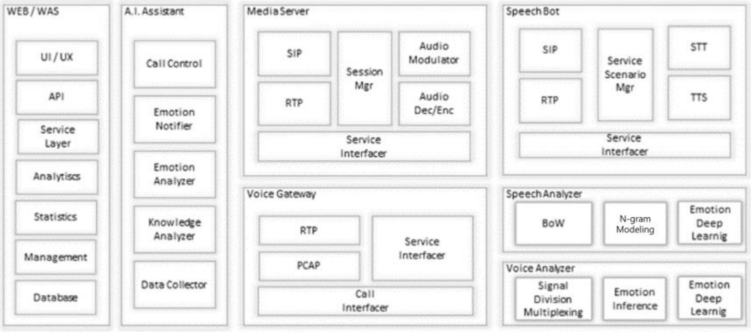


Figure 7. System Structure

4.1. System Structure Details

4.1.1. A.I. Assistant

This module uses the final emotion/voice analysis and text data to determine and control the necessary work efficiency and speaker protection services. Machine Learning (ML) and Natural Language Processing (NLP) algorithms would be used in this component to make intelligent decisions based on the input data.

4.1.2. WEB/WAS

This module provides the user interface for controlling and monitoring the integrated services. The front end is most likely built with JavaScript frameworks such as React for an interactive and responsive user experience.

4.1.3. Voice Gateway

This component monitors and collects the voice stream during a call and delivers it to the necessary details. This could involve using APIs for real-time voice data capturing and streaming.

4.1.4. Speech Analyzer

This component comprises a Speech-to-Text (STT) service that converts voice data to text and a Text Analyzer (TA) that extracts emotional data from the converted text. STT could be implemented using services like Google Speech-to-Text API or IBM Watson, while TA would utilize NLP and sentiment analysis models to extract emotional data.

4.1.5. Voice Analyzer

This component extracts the speaker's emotional state using voice data [4]. This could be achieved using various voice/emotion recognition APIs or services.

4.1.6. Media Server

This server processes the speaker's voice data according to the A. I Assistant's judgment. This might include tasks like voice modulation, data compression or expansion, etc.

4.1.7. Speech Bot

This component intervenes in the conversation between the speakers based on the A. I Assistant's judgment. It could use chatbot APIs or custom-built chatbot models to generate human-like text responses.

4.1.8. Emotion Analyzer

The emotion analysis module in our system draws on principles from multi-domain sentiment analysis, allowing for robust emotion recognition across various service sectors.

4.2. System Scenario

1. When a call begins between speakers, the Voice Gateway monitors and captures the necessary voice data. This module could employ real-time data capturing and voice analysis techniques.
2. The Voice Gateway then sends this voice data to the Speech/Voice Analyzer. The Speech Analyzer component converts the raw voice data into text data using Speech-to-Text (STT) technologies. The Voice Analyzer, on the other hand, assesses the emotion or tone of the speaker directly from the voice data.
3. Once the Speech/Voice Analyzer has analyzed the data, it is sent to the A.I. Assistant. The A.I. Assistant's role is crucial, making judgments based on the analyzed data. The decisions could range from assessing the speaker's emotional state to deciding what response or action is needed.
4. Depending on the judgment of the A.I. Assistant, necessary services or actions are triggered. This could involve processing the speaker's voice data via the Media Server, letting the Speech Bot intervene in the conversation, or providing insights into the consultant's client interface.

4.3. System Technologies

The call between the customer and the consultant is conducted in real-time using WebRTC technology. WebRTC is an open-source project that enables real-time communication between web browsers without plugins. This allows direct data exchange between web browsers like voice, text, and video.

4.3.1. Real-Time Communication

The call between the customer and the consultant is conducted in real-time using WebRTC technology [5]. WebRTC is an open-source project that enables real-time communication between web browsers without plugins. This allows direct data exchange between web browsers like voice, text, and video. The real-time communication in our system is further facilitated by implementing CoTURN TURN Server [6] for efficient voice transmission, ensuring smooth and uninterrupted communication during customer consultation. The design of our real-time communication system benefits from the comparative analysis of real-time communication protocols, especially in customer support environments.

4.3.2. Signaling Server

The signaling server is a protocol server that facilitates customer and consultant interactions. For real-time messaging between the customer and the consultant, our system integrates Spring Boot WebSocket [7], ensuring seamless interaction.

4.3.3. LINCALL Server

The LINCALL server performs various functions, including session management, interactions with the database, providing multiple data, and integration with APIs. It provides an integrated communication environment and processes and analyzes necessary real-time information. The LINCALL server is inspired by novel approaches to call center support using voice analysis. This includes the handling of voice recognition, emotion analysis, and other key functionalities that enhance the support structure within the call center.

4.3.4. ML Server

The ML server is the core server of LINCALL, performing voice recognition, emotion analysis, STT (Speech-to-Text) functions, and voice filtering [8]. It uses commercialized or open APIs and employs machine learning algorithms to extract and analyze useful information from voice data. Our system utilizes advanced voice processing technologies, modeled after successful AI-based approaches to enhancing call center efficiency [9]. This contribution leverages existing advancements to provide a robust solution for real-time emotion detection and analysis.

4.3.5. Intelligent Q&A

The implemented system is designed to convert voice to text and analyze conversations to understand the customers' emotions. Utilizing the Universal Sentence Encoder [10] for effective text representation enables nuanced emotion detection, allowing consultants to respond to customer emotions more targeted and empathetic.

The automatic question recommendation subsystem specifically aids counselors in the counseling domain. It begins by crawling general counseling manuals and information from counselingcenters, which are then organized into a database. These collected texts are encoded using the Universal Sentence Encoder, a procedure mirrored for texts arising during counseling. Questions with high similarity are selected and recommended by calculating the cosine similarity between the counseling text and the database information. This process equips counselors with the tools necessary to conduct more effective counseling, creating a more adaptive and responsive consultation experience.

Table 1. Cosine similarity sample result

Compare Sentence	**Cosine Similarity**
No, why don't you put a spoon in it?	0.2615303099155426
How can I register?	0.9689831733703613
The food I ordered came too late. Can I cancel the order?	0.6478094458580017
My order hasn't arrived in an hour, but you should have informed me beforehand.	0.23400570452213287
Please refund	0.9578129960214632
I don't know the rules, you must compensate	0.34066885709762573
How do I check when my shipment is coming?	0.7346420288085938
The period of use of the card has expired. How can I get the card reissued?	0.5416097640991211
I want to top up my mobile phone bill. How can I do that?	0.47308844327926636

Through the experimental results, a recommendation threshold was set only to suggest questions when the cosine similarity exceeds 0.4. This criterion enables more accurate question recommendations and provides more meaningful assistance during the counseling process.

4.3.6. STT(Speech-To-Text)

In this study, commercially used APIs were utilized, and a method of adding neologisms tailored to the unique environment of the call center was adopted. This contributed to enhancing the operational efficiency and accuracy of communication within the call center.

5. Experimental Results

The experiment in this study was conducted among college students with experience in telephone counseling jobs. The research was structured around scenarios in call centers for food delivery, courier services, credit card companies, telecommunications companies, and insurance companies, with 100 participants in the scenarios.

5.1. Experimental Design

The participants were given specific counseling scenarios and conducted consultations with virtual customers using the system designed in this study. The system's various features, such as Real-Time Transcript, Voice Emotion Analysis, Text Emotion Analysis, etc., were evaluated to see how they operate in counseling situations.

5.1.1. Statistical Hypotheses

For a more rigorous analysis, the study formulated the following hypotheses:
H_0: The proposed system does not significantly improve counseling efficiency.

H_1: The proposed system significantly improves counseling efficiency.

5.2. Experimental Results

5.2.1. Counseling Efficiency

Table 2. Participants' responses on the effectiveness of the system's features

Response Category	**Percentage of Respondents**
Helpful	72%
Unnecessary	13%
Somewhat Helpful	15%

Most participants (72%) responded that the system's features were helpful in real-time conversation analysis. 13% found it unnecessary, and 15% responded that it was not significantly helpful but better than having nothing.

Statistical Test for Hypothesis 1
To test the hypotheses, a chi-square goodness-of-fit test was conducted. The p-value was found to be less than 0.05, suggesting that H_0 can be rejected in favor of H_1.

5.2.2. Operational Accuracy

78% of participants rated the system's accuracy as good.

5.2.3. Application of Neologisms

Although the college students did not specifically evaluate the application of neologisms, the application was confirmed to be successful. The system was able to accurately recognize phrases like "call center," "the delivery is late," "tell me the credit card limit," and "when will the package arrive?"

5.2.4. User Satisfaction

Table 3. User Satisfaction Ratings

Satisfaction Category	**Percentage of Respondents**
Satisfied	67%
Unsatisfied	33%

Among those dissatisfied, 64% felt that the interface was not excellent, 24% did not think it would be particularly helpful, and the rest did not respond.

6. Conclusion and Future Research Directions

In this paper, a new system is proposed that aims to more precisely understand the emotional state of a customer by analyzing both their voice and text simultaneously. This system analyzes the customer's negative emotions complementarily and notifies the consultant when angry, allowing for a more sensitive response to the customer's emotions. Additionally, the ability to analyze keywords from the conversation offers deep insights into the customer's needs and emotions.

The system greatly reduces the burden on consultants and increases customer satisfaction. By integrating machine learning-based emotion analysis and real-time voice processing technology, this system operates as a tool that significantly enhances customer service efficiency. In fact, its effectiveness was proven when applied to services such as food delivery.

Future plans focus on utilizing more data to increase the accuracy of the system's emotion analysis and improve the performance of voice processing. Additionally, the development of multimodal emotion recognition technology using not only voice information, but also facial expressions, movement, and biometric signals is targeted, aiming to implement a personalized intelligent system.

REFERENCES

[1] Consiglio, Chiara & Borgogni, Daniela & Alessandri, Laura & Barbaranelli, Claudio, (2016) "Inbound Call Centers and Emotional Dissonance in the Job Demands – Resources Model", Frontiers in Psychology, Vol. 7, No. 1133, ppAugust 2016.

[2] Hochschild, Arlie R., (1983) "The Managed Heart: Commercialization of Human Feeling", University of California Press, Vol., No., pp.

[3] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N., (2017) "Attention is all you need", Advances in Neural Information Processing Systems, Vol. NIPS '17, pp5998-6008.

[4] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina, (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arXiv:1810.04805, Vol., No., pp.

[5] WebRTC, (2021) "WebRTC: APIs and RTCWEB Protocols of the HTML5 Real-Time Web", https://webrtc.org/.

[6] CoTurn, (2021) "CoTURN TURN Server", https://github.com/coturn/coturn.

[7] Spring Boot, (2020) "Spring Boot WebSocket", https://spring.io/guides/gs/messaging-stomp-websocket/.

[8] GitHub, (2021) "speech-emotion-recognition • GitHub Topics • GitHub", https://github.com/topics/speech-emotion-recognition.

[9] Goodfellow, Ian & Bengio, Yoshua & Courville, Aaron, (2016) "Deep Learning", MIT Press, Vol., No., pp.

[10] Google, (2020) "Universal Sentence Encoder", https://tfhub.dev/google/universal-sentence-encoder/4.

AUTHORS

Changmook Oh, HCI, AI,Homomorphic Encryption, Database