STT API와 ChatGPT API를 활용한 강의 요약 애플리케이션 구현

김진웅[1], 금보성[1], 김태국[2*]
[1]국립부경대학교 컴퓨터·인공지능공학부 학생, [2]국립부경대학교 컴퓨터·인공지능공학부 교수

Implementing a Lecture Summary Application Using STT API and ChatGPT API

Jin-Woong Kim[1], Bo-Seong Geum[1], Tae-Kook Kim[2*]
[1]Student, School of Computer and Artificial Intelligence Enginerring, Pukyong National University
[2]Professor, School of Computer and Artificial Intelligence Enginerring, Pukyong National University

요 약 ChatGPT의 등장 이후 다양한 분야에서 이를 활용하려는 시도가 활발히 이루어지고 있다. 본 논문에서는 인터넷 연결이 가능한 모바일 기기에서 사용할 수 있는 강의 요약 애플리케이션을 제안한다. 제안된 애플리케이션은 다음의 세 가지 주요 기능을 제공한다. 첫째, 강연자의 음성을 녹음하여 파일로 저장하는 기능, 둘째, 저장된 음성 파일을 문자로 변환하는 음성 인식(Speech-to-Text) 기능, 셋째, 변환된 텍스트를 기반으로 ChatGPT API를 활용하여 요약본을 생성하는 기능이다. 코로나19 종식 선언 이후 비대면 온라인 강의가 대면 수업으로 전환되면서 영상 다시보기가 어려워지는 상황에서, 본 애플리케이션은 강의 내용을 자동으로 요약하여 제공함으로써 학습자의 복습 시간을 줄이고 학습 효율성을 높이는 데 기여할 수 있다. 또한 회의록 요약 등 다양한 실무 분야에도 확장 가능성이 클 것으로 기대된다.

주제어 : 강의 요약, STT(Speech to Text), ChatGPT, API, 사물인터넷

Abstract Since the emergence of ChatGPT, there has been growing interest in its application across various domains. This paper proposes a lecture summarization application designed for mobile devices with Internet connectivity. The application offers three core functionalities: (1) recording and saving the lecturer's voice as an audio file, (2) converting the recorded audio into text using a speech-to-text engine, and (3) generating a summarized version of the lecture using the ChatGPT API based on the transcribed text. With the transition from online to face-to-face classes following the end of the COVID-19 pandemic, the ability to rewatch recorded lectures has diminished. The proposed application addresses this issue by providing automated summaries, thereby reducing review time and enhancing learning efficiency. Furthermore, the system holds significant potential for broader applications, including automated meeting minutes summarization and other professional use cases.

Key Words : Lecture summary, STT(Speech to Text), ChatGPT, API, Internet of Things (IoT)

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00242528).
*교신저자 : 김태국(king@pknu.ac.kr)
접수일 2025년 05월 14일 수정일 2025년 06월 02일 심사완료일 2025년 06월 11일

1. 서론

COVID-19의 종식으로 인해 감염 예방을 목적으로 운영되던 비대면 온라인 강의가 대부분 대면 강의로 전환되었다. 그러나 대면 강의의 경우, 온라인 강의와 달리 강의 영상이 별도로 녹화되지 않아 수업 내용을 다시 확인하기 어렵다는 불편함이 존재한다. 이에 따라 학생들은 강의 중 녹음과 필기를 병행하며 학습 내용을 보완하고 있지만, 녹음 파일만으로 복습을 진행하는 데에는 한계가 있다. 그리고 특정 내용을 빠르게 찾기 위해서는 필기 내용과 음성 파일을 일일이 대조해야 하는 번거로움이 따른다.

본 논문에서는 이러한 문제를 해결하기 위해, 모바일 기기와 오픈 API(Open API)[1] 기술을 활용하여 강의 내용을 자동으로 녹음하고, 필사하며, 요약할 수 있는 애플리케이션을 구현하였다. 사용자는 모바일 기기를 통해 강연자의 음성을 녹음하고, 해당 음성은 STT(Speech to Text) API를 통해 텍스트로 변환되며, 이후 ChatGPT API를 통해 요약문이 생성된다. 이를 통해 사용자는 강의 음성, 텍스트 필사본, 요약본을 종합적으로 활용하여 보다 효율적인 학습이 가능하다.

2. 관련 연구

기존의 STT 기술은 인공지능 비서, 영상 자막 생성, 음성 명령 처리 등 다양한 분야에서 널리 활용되어 왔다. 한편, 텍스트 요약 기술은 회의록 요약이나 뉴스 기사 요약 등 정보의 압축 및 전달을 목적으로 한 응용 사례에 주로 사용되고 있다. 그러나 강의 내용을 녹음하고, 이를 텍스트로 변환한 후 자동으로 요약까지 제공하는 통합 애플리케이션에 대한 연구는 아직 활발히 이루어지지 않은 실정이다. 지금까지는 주로 요약 알고리즘이나 STT 성능 향상에 초점을 맞춘 개별 연구들이 주를 이루었다.

배영준 등은 PageRank 알고리즘을 기반으로 한 TextRank 요약 알고리즘을 제안하였다. 이 연구에서는 여러 발화자의 음성을 텍스트로 변환한 후, 단어의 출현 빈도를 기반으로 회의록을 요약하는 방식을 적용하였다[2].

임지원 등은 소음을 효과적으로 필터링하고 STT 성능을 향상시키는 방법을 제안하였다. 이들은 음성 신호의 주파수 성분 간 상호 연관성을 분석하여 보다 정확한 음성 인식 결과를 도출하고자 하였다[3].

김진웅 등은 API 통신을 활용한 강의 요약 애플리케이션을 연구하였으며, 시스템 구현 가능성을 검토하였다. 이 연구는 강의 콘텐츠 요약 기능의 적용 가능성을 보여준 초기 연구로 의의가 있다[4].

또한, 이재걸[5], 한광록[6], 이소연[7], 이건희[8] 등은 의미 있는 핵심어를 추출하여 텍스트 콘텐츠를 요약하는 알고리즘 개발에 주력하였다. 이와 함께, 이혜정[9], 박해긍[10], 윤기혁[11] 등의 연구에서는 텍스트 마이닝 기법을 통해 자연어 데이터를 처리하고, 이로부터 유의미한 정보를 추출하는 방법론을 제시하였다.

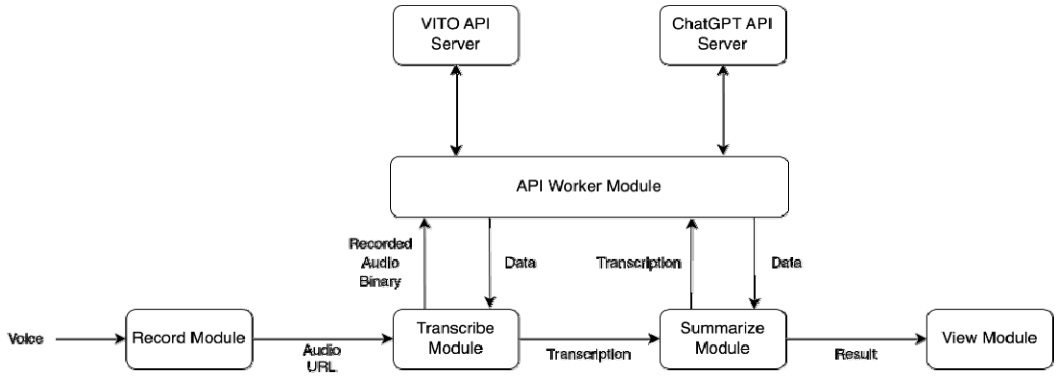
이와 같은 선행연구들은 음성 인식, 텍스트 요약, 자연어 처리 기술의 발전에 기여했으나, 본 논문에서 제안하는 바와 같이 강의의 전체 흐름(녹음, 필사, 요약) 하나의 통합 애플리케이션으로 구현한 사례는 찾기 어려우며, 이에 대한 실증적 연구는 부족한 실정이다.

3. 강의 요약 애플리케이션 설계

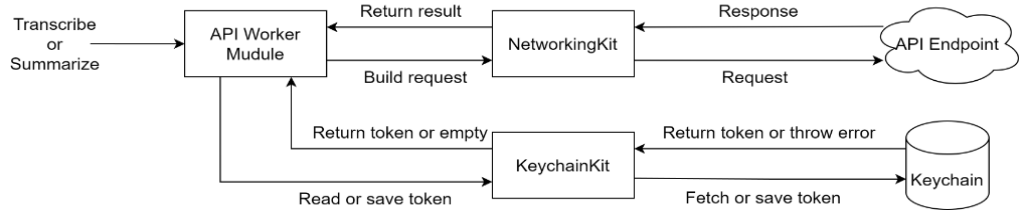
본 연구에서 구현한 강의 요약 애플리케이션은 각 기능을 모듈화하여 구조적으로 설계되었다. [Fig. 1]은 각 모듈의 역할과 작동 흐름을 나타낸 순서도이다. 각 모듈은 고유한 역할을 수행하며, 전체 애플리케이션은 프레임워크 형태로 빌드(build)되어 통합적으로 구성되었다.

〈Table 1〉 STT API

Service	Korean Support	SDK Support	Pricing
Apple Speech[12]	Support	- Only on Apple's platforms	- Server: supports sessions at one-minute intervals only - On-device: free of charge
Google Cloud STT[13]	Support	- REST API provisioned - Available in Python and seven other languages	- Free tier: 1H per month - Excess: $0.0024 per minute
VITO Speech[14]	Support	- REST API provisioned	- Up to 100H free in total
AWS Transcribes[15]	Support	- Available in Python and six other languages	- Free for 1H per month



[Fig. 1] Application flowchart



[Fig. 2] API Worker Module Flow

제안된 애플리케이션은 음성을 텍스트로 변환하기 위한 STT 기술과, 변환된 텍스트를 요약하기 위한 ChatGPT API를 활용하며, 이들 기능은 API 통신을 통해 처리된다. 〈Table 1〉은 공개된 주요 STT API 목록을 나타내며, 본 연구에서는 한국어 인식이 가능하고 총 100시간의 무료 사용 시간을 제공하는 VITO STT API를 채택하였다. 해당 API 서버는 요청에 대해 JSON(JavaScript Object Notation) 형식으로 응답하며, 토큰 관련 정보에는 토큰 값과 유효 시간 등이 포함되고, 필사본 정보에는 발화 시간과 발화 내용이 포함된다.

3.1 API Worker Module

API Worker Module은 STT API 및 ChatGPT API와의 통신을 담당하는 핵심 모듈이다. 이 모듈은 실제 네트워크 통신을 수행하는 NetworkingKit과, API 인증 토큰을 안전하게 저장하고 호출할 수 있는 KeychainKit을 포함하여 구성된다. [Fig. 2]는 API Worker Module의 동작 과정을 나타낸다.

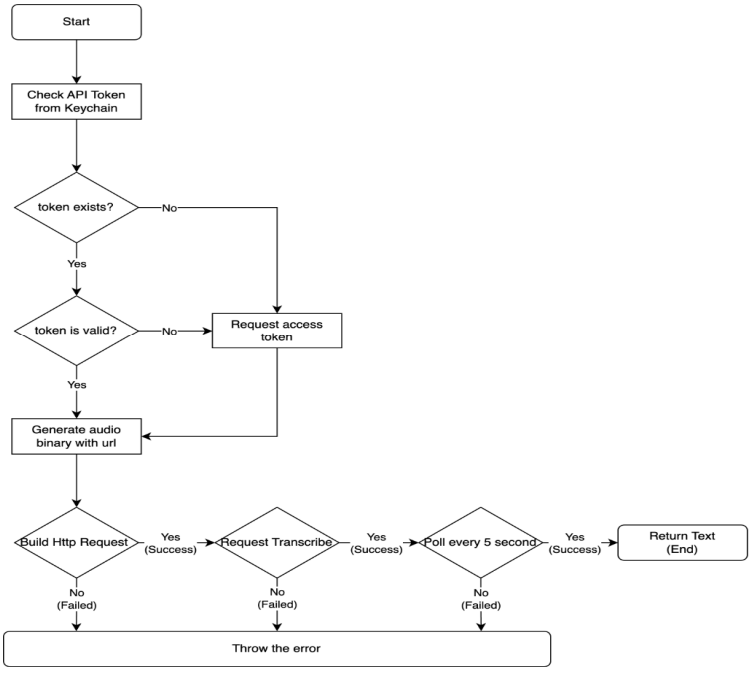
VITO STT API와 ChatGPT API는 서로 다른 HTTP 요청 형식을 요구하므로, 본 연구에서는 각 요청 조건을 추상화하여 유연하게 처리할 수 있도록 설계하였다. 예를 들어, 네트워크 통신 시 HTTP Header에 포함되는 Content-Type의 미디어 타입(Media Type)을 추상화한 후, 해당하는 타입을 입력값으로 전달받도록 구성하였다.

3.1.1 Transcribe Process

[Fig. 3]은 입력으로 제공된 오디오 바이너리 데이터를 기반으로 음성 인식(transcribe)을 수행하는 VITO API 서비스의 처리 흐름을 나타낸다.

API Worker Module 내의 VITO API 서비스는 입력값으로 녹음된 오디오 파일의 URL을 수신한다. 네트워크 통신을 시작하기에 앞서, 먼저 KeychainKit을 통해 저장된 인증 토큰의 존재 여부 및 유효성을 확인한다. 서버에서 획득한 토큰은 만료 시간이 유닉스 시간(Unix time) 형식으로 제공되며, 이를 현재 시간과 비교하여 유효성을 판단한다. 만약 토큰이 없거나 유효하지 않을 경우, 서버로부터 새로운 토큰을 요청하여 획득한 후 이후 과정을 진행한다.

STT API 서버는 HTTP 요청 시 Content-Type을 multipart/form-data로 지정하며, 요청 본문(body)에는 오디오 바이너리 데이터를 포함하도록 명시되어 있



[Fig. 3] Flowchart of Transcribing

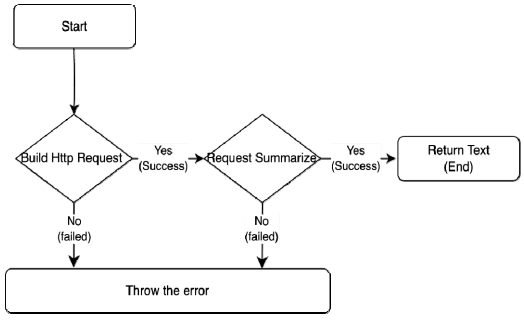
다. 이에 따라 Record 모듈을 통해 녹음된 음성 파일을 바이너리 형태로 변환하고, 앞서 획득한 토큰과 함께 전송하여 Transcribe 요청을 수행한다.

이 과정에서 다음과 같은 사유로 오류가 발생할 수 있으며, 오류 발생 시 프로세스는 즉시 종료된다. 주요 오류 원인으로는 (1) 서버 URL 변경으로 인한 기존 주소의 무효화, (2) 발급받은 토큰의 오류, (3) 오디오 파일 자체의 손상 또는 형식 오류 등이 있다.

Transcribe 요청이 정상적으로 수신되면, 서버는 고유 ID를 반환한다. 클라이언트는 해당 ID를 기반으로 5초 간격으로 GET 방식의 요청을 반복 전송하며 처리 상태를 확인한다. 음성 인식이 완료되었을 경우, 서버는 응답 본문(Response Body)에 필사본 데이터를 포함하여 반환하며, 이 데이터를 가공하여 최종 결과로 출력한다. 단, 요청 후 60초 이상이 경과하거나 응답 상태 메시지에서 오류가 감지될 경우에도 프로세스는 중단된다.

3.1.2 Summarize Process

ChatGPT API는 VITO STT API와는 달리, client id와 client secret 기반의 인증 방식이 아닌, 단일 API Key를 활용한 접근 방식을 채택하고 있다. 또한, HTTP 요청 시 Content-Type을 application/json으로 지정해야 하며, 요청 본문은 요구되는 JSON 형식에 맞추어 구성되어야 한다. [Fig. 4]는 텍스트 요약 처리 과정을 나타낸다.



[Fig. 4] Flowchart of Summarization

API Worker Module의 GPT API 서비스는 STT 과정을 통해 생성된 텍스트 파일의 URL을 입력값으로 수신한다. API Key 인증 방식에서는 별도의 토큰 유효성 확인이나 재발급 절차가 필요하지 않기 때문에, 즉시 HTTP 요청을 구성하여 요약 요청을 수행할 수 있다.

또한, ChatGPT API는 요청과 동시에 응답으로 요약 결과를 반환하므로, VITO STT API와 달리 결과를 확인하기 위한 주기적인 요청(예: 5초 간격의 polling)이 필

〈Table 2〉 Network Endpoint

Endpoint				
Endpoint URL		STT API Endpoint URL		ChatGPT API
Endpoint Path		Require Access Path	Request Transcribe Path	Request Summarize Path
HTTP Header	Content-Type	x-www-form -urlencoded	multipart/form-data	application/JSON
	Accept	application/JSON		
HTTP Method		POST	POST/GET	POST
HTTP Body		user secrets	audio binary	JSON type data

요하지 않다. 단, 이 과정에서도 API Key가 유효하지 않거나 서버의 도메인 변경으로 인해 지정된 엔드포인트(Endpoint)가 무효화될 경우 오류가 발생할 수 있으며, 이 경우 프로세스는 즉시 종료된다.

3.1.3 Network and Keychain

각 API 서버는 고유한 URL, 경로(Path), 그리고 HTTP Header 구성을 요구한다. 이러한 요소들을 일관되게 처리하고 재사용성을 높이기 위해, 네트워크 통신 인터페이스를 〈Table 2〉에 제시된 Endpoint 유형으로 추상화하였다. '네트워크 엔드포인트(Network Endpoint)'란 컴퓨터 네트워크에 연결되는 모든 장치를 의미한다. 입력값을 Endpoint 타입으로 통일함으로써 코드의 재사용성과 유지보수성을 향상시킬 수 있다.

또한, 각 서비스에서 요구하는 인증 정보를 Keychain에 저장하기 위해 인증 토큰 구조 역시 추상화하였다. 서비스 이름과 해당 토큰 정보를 하나의 JSON 데이터로 인코딩할 수 있는 구조체로 정의하여, 다양한 서비스의 시크릿 값을 저장할 수 있도록 하였다. STT API의 경우 유효 기간(expiration time)이 존재하는 액세스 토큰을 저장하며, ChatGPT API의 경우에는 만료 기간이 없는 단일 API Key가 저장된다.

3.2 Other Modules

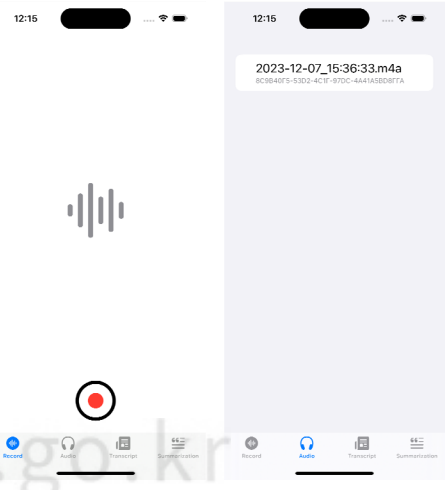
Record 모듈은 모바일 디바이스의 마이크 기능을 활용하여 음성 녹음을 수행하며, Transcribe 및 Summarize 기능은 API 워커 모듈을 통해 처리된다. 애플리케이션은 사용자의 '녹음 시작' 입력을 트리거로 받아 Record 모듈을 통해 녹음을 시작하고, '녹음 중지' 명령이 입력되면 녹음을 종료한 후 음성 파일을 생성하여 기기 내에 저장한다.

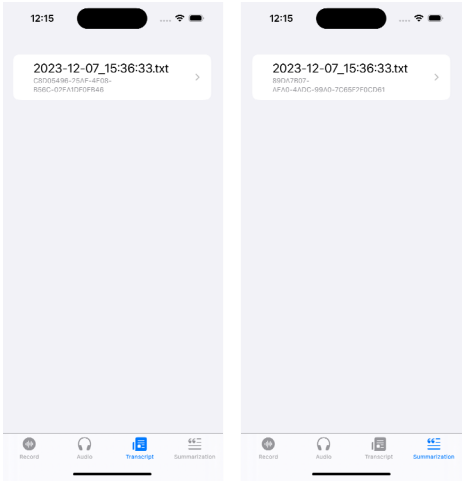
생성된 음성 파일은 Transcribe 모듈에 전달되며, 이는 API 워커 모듈을 통해 STT 처리 요청을 수행하고 필사본 텍스트를 생성한다. 생성된 텍스트 역시 로컬 저장소에 저장된다.

이후 Summarize 모듈은 해당 텍스트 파일을 기반으로 API Worker Module을 통해 요약 요청을 수행하며, 생성된 요약 결과 역시 기기 내에 저장된다. 이러한 일련의 프로세스를 통해 사용자는 음성 파일, 텍스트 필사본, 요약본을 통합적으로 활용할 수 있다.

4. 애플리케이션 동작 실험

개발된 애플리케이션을 실제로 실행하여 각 기능의 동작 여부를 확인하였다. [Fig. 5]는 애플리케이션의 주요 동작 화면을 보여주며, 순서대로 음성을 녹음하는 화면, 생성된 오디오 파일을 확인하는 화면, STT를 통해 생성된 필사본 확인 화면, 그리고 최종 요약본을 확인하는 화면으로 구성된다.

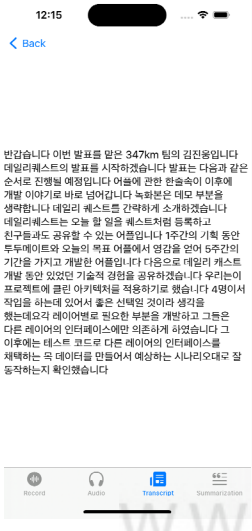


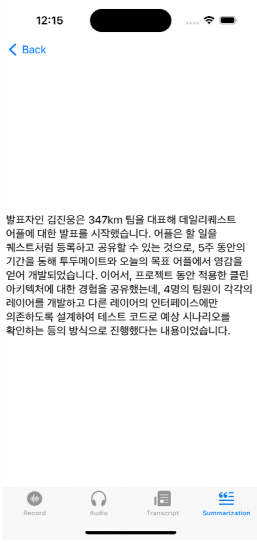


[Fig. 5] Application main operation screen

애플리케이션은 사용자가 녹음을 시작한 시간을 기준으로 파일명을 자동 생성하며, 각 파일은 범용 고유 식별자(UUID, Universally Unique Identifier)를 부제목으로 포함하여 고유하게 식별된다. 범용 고유 식별자는 컴퓨터 시스템에서 객체를 고유하게 식별하는 데 사용되는 128비트 레이블이다.

실험에는 실제 프로젝트 발표에 사용된 대본을 기반으로 음성을 녹음하였다. 해당 내용은 애플리케이션 개발에 소요된 시간, 개발 아이디어의 기원, 그리고 향후 개발 방향성에 관한 내용으로 구성되어 있다. [Fig. 6]은 녹음된 음성에 대한 STT 결과와 요약본을 나타낸다. VITO STT API를 사용하여 음성 파일을 텍스트로 변환(STT)하였고, ChatGPT API를 활용하여 변환된 텍스트를 요약하였다.





[Fig. 6] Transcription and summary details

〈Table 3〉은 애플리케이션 동작에서의 문제점을 나타낸다. 실험 결과, 음성 인식 과정에서 일부 발음이 부정확하게 인식되는 사례가 확인되었으며, 이로 인해 필사본 내 일부 단어가 실제 발화 내용과 다르게 표기되는 문제가 발생하였다. 예를 들어, “어플에 관한 한줄 소개 이후에 개발 이야기로...”를 “어플에 관한 한줄속이 이후에 개발 이야기로...”로 텍스트 변환에 오류가 있었다. ChatGPT API는 이와 같은 오류 상황에서도 유사한 의미의 단어 중 문맥상 자연스러운 단어를 선택하여 요약문을 생성하는 경향을 보였다. 그러나 발음 오류가 빈번하게 발생하는 경우에는 필사본의 정확도가 떨어지며, 이에 따라 요약 결과의 품질에도 영향을 줄 수 있으므로, 이러한 문제를 보완하기 위한 후속 연구가 필요하다.

〈Table 3〉 Operational issue in the application

Content	Description
Voice Recognition	Some pronunciations are recognized incorrectly
Recording quality	Voice recording quality issues due to ambient noise and recording from a distance

5. 결론

본 연구에서는 모바일 기기와 STT, ChatGPT API를 활용하여 강의 내용을 녹음하고, 텍스트로 필사하며, 요

약문을 자동으로 생성하는 강의 요약 애플리케이션을 설계하고 구현하였다. 제안된 애플리케이션은 사용자가 녹음을 시작하고 종료하는 입력만으로, 이후의 필사 및 요약 과정이 자동으로 수행되도록 설계되었으며, 각 단계는 네트워크 통신을 통해 처리된 결과를 기반으로 동작한다.

제안된 시스템은 특히 대학 강의 현장에서 유용하게 활용될 수 있다. 사용자는 단순한 조작만으로 강의 내용을 자동으로 기록하고 요약할 수 있어, 강의 내용을 복습하거나 정리할 때 높은 편의성과 효율성을 제공한다.

다만, 실험 과정에서 확인된 바와 같이, 강연자의 발음에 따라 음성 인식 정확도가 낮아지는 경우가 있으며, 이는 필사본의 품질 저하로 이어지고, 궁극적으로 요약 결과의 신뢰성에도 영향을 줄 수 있다. 이러한 한계를 극복하기 위해, 향후에는 발음 오류 보정 알고리즘 도입, 사전 정의된 사용자 발음 프로필 기반의 음성 학습 등 추가적인 기술적 보완이 요구된다.

REFERENCES

[1] J.M.Choi, "Quiz Generation System Using Google Forms and ChatGPT," *Journal of Internet of Things and Convergence*, Vol.10, No.6, pp.105-110, 2024.

[2] Y.J.Bae, H.T.Jang, T.W.Hong, H.Y.Lee, "Automatic Meeting Summary System using Enhanced TextRank Algorithm," *The Journal of Korea Institute of Information, Electronics, and Communication Technology*, Vol.11, No.5, pp.467-474, 2018.

[3] J.W.Lim, Y.H.Hwang, K.H.Kim, "Noise filtering method based on voice frequency correlation to increase," *The Korean Institute of Broadcast and Media Engineers*, pp.176-129, 2021.

[4] J.W.Kim, B.S.Geum, T.K.Kim, "A Lecture Summarization Application Using STT and ChatGPT," *Annual Conference of KIPS*, pp.297-298, 2023.

[5] J.K.Lee, S.B.Park, S.J.Lee, "Meeting Minutes Summarization using Two-step Sentence Extraction," *The Journal of Korea Institute of Intelligent Systems*, Vol.20, No.6, pp.741-747, 2010.

[6] K.R.Han, S.K.Oh, K.W.Rim, H.Y.Lee, "Document Summarization using Topic Phrase Extraction and Query-based Summarization," *The Journal of KIISE*, Vol.31, No.4, pp.488-497, 2004.

[7] S.Y.Lee, J.E.Choi, T.W.Hong, S.Y.Yoo, "A Study on the Content Summary Based on Attention Algorithm," *The Journal of Digital Contents Society*, Vol.22, No.9, pp.1487-1491, 2021.

[8] K.H.Lee, S.H.Na, J.H.Lim, T.W.Hong, T.H.Kim, D.S.Chang, "PrefixLM for Korean Text Summarization," *The Journal of KIISE*, Vol.49, No.6, pp.475-487, 2022.

[9] H.J.Lee, K.-H. Youn, "The Analysis of Research Trends in Social Service Quality Using Text Mining and Topic Modeling," *Journal of Internet of Things and Convergence*, Vol.8, No.3, pp.29-40, 2022.

[10] H.K.Park, K.H.Youn, "An Analysis on Media Trends in Public Agency for Social Service Applying Text Mining," *Journal of Internet of Things and Convergence*, Vol.8, No.2, pp.41-48, 2022.

[11] K.H.Youn, "Trend Analysis of Fraudulent Claims by Long Term Care Institutions for the Elderly using Text Mining and BIGKinds," *Journal of Internet of Things and Convergence*, Vol.8, No.2, pp.13-24, 2022.

[12] Apple, https://developer.apple.com/documentation/

[13] Google, https://cloud.google.com/speech-to-text/

[14] VITO, https://developers.vito.ai/

[15] Amazon, https://aws.amazon.com/ko/transcribe/

김 진 웅(Jin-Woong Kim) [준회원]

■ 2016년 3월 ~ 2024년 2월 : 국립부경대학교 컴퓨터·인공지능 공학부

〈관심분야〉
모바일, 아키텍처 패턴

금 보 성(Bo-Seong Geum) [준회원]

■ 2019년 3월 ~ 2024년 8월 : 국립부경대학교 컴퓨터·인공지능 공학부

〈관심분야〉
모바일, 인공지능(AI)

김 태 국(Tae-Kook Kim) [종신회원]

- 2004년 8월 : 고려대학교 전기전자전파공학부(공학사)
- 2006년 8월 : 고려대학교 메카트로닉스학과(공학석사)
- 2014년 8월 : 고려대학교 모바일솔루션학과(공학박사)
- 2016년 3월 ~ 2022년 2월 : 동명대학교 AI학부 교수
- 2022년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능 공학부 교수

〈관심분야〉
사물인터넷(IoT), 콘텐츠 전송 네트워크(CDN), 이동성, 인공지능(AI), 빅데이터, 모바일 서비스