한국기록관리학회지
25(3), 243-258, 2025.8
https://doi.org/10.14404/JKSARM.2025.25.3.243
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

오픈소스 기반의 AI 음성·텍스트 변환 기능 개발 및 대통령 음성을 통한 성능 분석

Development of an Open-Source-Based AI Speech-to-Text System and Performance Analysis Using Presidential Speech

배민수(Minsoo Bae)[1], 유영문(Young-Moon Yu)[2]

E-mail: getta811@korea.kr, yuym4599@korea.kr

1 제1저자 대통령기록관 공업연구사
2 교신저자 대통령기록관 공업연구관

논문접수 2025.07.15
최초심사 2025.07.26
게재확정 2025.08.22

ORCID

Minsoo Bae
https://orcid.org/0009-0000-7497-2943

Young-Moon Yu
https://orcid.org/0000-0001-7513-1644

초 록

본 연구는 오픈소스 기반의 AI 음성·텍스트 변환(STT, Speech To Text) 기능을 개발하고 대통령의 음성에 적용하여 그 성능을 분석하였다. 현재 다양한 고성능 STT 서비스가 상용화되어 있으나, 대부분 온라인 환경에서 유료로 제공되고 있다. 하지만 대통령 기록물의 특성상 온라인 서비스의 사용은 보안 등의 문제를 발생할 수 있고, 누적되는 기록물에 지속적인 비용을 들여 처리하는 것은 비효율적이다. 따라서, 대통령기록관은 오픈소스 기반의 AI 모델을 적용한 STT를 개발하고 오프라인으로 시험·운용 중이다. 본 연구에서는 이 기능을 통해 약 3시간 분량의 대통령 시청각기록물을 텍스트로 변환하고, 실제 텍스트와의 비교를 통해 오류율을 측정하였다. 그 결과 전체적으로 최신 상용 온라인 서비스 수준의 성능을 확인하였다. 또한, 발화 속도 및 녹음 품질을 추가로 추출하여 오류율과의 연관성을 분석하였다. 최종적으로 기록물의 활용을 위한 오픈소스 기반 AI 기술의 적용 가능성을 제시한다.

ABSTRACT

This study developed an open-source-based AI Speech-to-Text (STT) system and analyzed its performance by applying it to presidential speech. While various high-performance STT services are currently commercialized, most are provided online for a fee. However, because of the nature of presidential records, using online services can raise security concerns, and incurring continuous costs for processing accumulating records is inefficient. To address this, the Presidential Archives has developed an offline STT system based on open-source AI models, which is currently under testing and operation. In this study, approximately three hours of presidential audiovisual records were transcribed into text using this function, and the error rate was measured by comparing with the actual text. The results showed that the overall performance is comparable to the latest commercial online services. Additionally, speech rate and recording quality were extracted and analyzed for their correlation with the error rate. Finally, this research highlights the feasibility of applying open-source AI technologies for the utilization of records.

Keywords: 인공지능, 음성·텍스트 변환, 오픈소스 소프트웨어, 대통령기록관, 시청각기록물
Artificial intelligence, Speech-to-text, Open-source software, Presidential Archives, audiovisual records

1. 서론

대통령제를 채택하여 실시하고 있는 우리나라에서 대통령은 최고 권력자이며 국가의 중요 정책 결정에 있어 큰 영향을 행사할 수 있으므로, 재임 기간 중 생산한 대통령 기록물은 후대의 국정 평가에 영향을 미치는 국가의 중요한 자산이다(남태우 외, 2007). 따라서, 대통령 기록물은 역사적 가치를 지니며, 후대 연구자들에게 중요한 자료로 활용될 수 있으므로 체계적인 보존과 관리가 필수적이다. 그러나 자기 테이프류의 아날로그 매체에 수록된 영상·음성 시청각기록물은 시간이 흐르면서 물리적 손상과 음질 저하가 생길 수 있다. 따라서 대통령기록관은 시청각기록물을 안전하게 보존하기 위하여, 소장하고 있는 모든 아날로그 매체를 디지털화하고 있다(나미선, 한상효, 2016). 하지만 기록물의 안전한 보존에서 나아가, 필요한 정보를 검색하고 활용하기 위해서는 추가적인 정보 추출 과정이 필요하다. 디지털화되었을지라도, 영상·음성 파일은 정의된 구조 없이 시각·청각 신호를 사용자가 감상할 수 있는 형태로 저장한 비정형데이터이기 때문이다.

컴퓨터 비전 기술을 통해 추출된 주요 장면을 확인할 수 있는 영상과 달리, 음성의 수록 내용을 확인하기 위해서는 전체 분량을 일일이 청취 감각에만 의존해야 하는 번거로움이 있으며, 파일에서 특정 내용을 검색하는 것은 비정형데이터의 특성상 불가능하다. 이러한 문제점을 해결하기 위해 음성 기록물을 텍스트로 변환하는 AI 기반 STT (Speech To Text) 기술의 도입이 절실히 요구된다(안대진, 2017). 텍스트 형태로 변환된 시청각기록물은 내용 파악의 용이성을 높일 뿐만 아니라, 키워드 검색, 내용 분석 등 다양한 방식으로 활용될 수 있어 기록물의 가치를 극대화할 수 있다.

다만, 대통령기록관에서 보유하고 있는 기록물에는 비공개 대상이 포함되어 있다. 비공개 기록물은 공개될 경우 국가 안전 보장, 공익 등을 침해할 우려가 있어 공개하지 않기로 결정된 기록물이기 때문에, 기록물의 정보추출 과정은 모두 관내 오프라인 환경에서 이루어져야 한다. 또한, 제한된 예산 환경에서의 솔루션 개발 비용 부담을 덜고, 빠르게 발전하는 AI 기술의 지속적 업데이트를 통해 최신 모델을 지원받을 수 있도록 오픈소스 기반의 서비스 구축이 필요하다.

따라서, 영상·음성 기록물의 특성, AI 기반 STT 기술 도입의 필요성 및 적용 환경 등을 고려하여, 본 연구의 목적을 다음과 같이 설정하였다.

1. 대통령 시청각기록물의 효율적인 관리와 활용을 위해 오프라인 환경에서 구동할 수 있는 오픈소스 기반 STT 기능을 개발한다.

2. 실제 대통령의 영상 및 음성 기록물에 개발된 STT를 적용하여 오류율을 측정하고, 발화 속도 및 녹음 품질을 추출하여 오류율과의 연관성을 분석한다.

3. 오픈소스 기반 AI 기술의 기록물 관리 및 활용으로의 적용 가능성을 확인하고, 향후 관련 연구 및 시스템 구축에 대한 시사점을 제시한다.

2. 기술적 배경

2.1 음성·텍스트 변환 기술

STT는 일정한 간격으로 샘플링 된 사람의 음성 데이터를 자동으로 인식하여 텍스트 데이터로 변환하는 기술이다. STT는 오랜 역사와 함께 꾸준히 발전해 왔으며, 1970년대 초기에는 간단한 패턴 기반 기술(Malik et al.,

 https://doi.org/10.14404/JKSARM.2025.25.3.243

2021)이나 통계적 모델(Rabiner, 1989)이 주를 이루었으나, 2010년대 딥러닝 기술(Mikolov et al., 2010)의 발전과 함께 비약적인 성능 향상을 이루게 되었다. <표 1>에 STT에 적용된 주요 기술을 시대순으로 정리하여 그 특징 및 한계점을 요약하였다.

<표 1> 딥러닝까지의 STT 주요 기술

구분	주요 기술	특징	한계점
기술 초기	패턴 기반 음성인식 기술	· 언어학적 규칙과 음향학적 지식 기반 음성 분석 및 텍스트 변환 · 제한된 대상 및 목적에만 사용 가능	· 복잡하고 다양한 실제 발음 처리 및 새로운 언어나 환경에 대한 적응성 부족
통계적 모델 기반	HMM(Hidden Markov Model) GMM(Gaussian Mixture Model)	· 패턴 기반 방식에 비해 대량의 음성 신호 특징을 통계적으로 학습하여 성능이 크게 향상됨	· 통계적 모델링 기법의 한계로 복잡한 음향 환경 및 발음 변화에 취약
딥러닝 기반	RNN(Recurrent Neural Network) LSTM(Long Short-Term Memory) GRU(Gated Recurrent Unit)	· 순차적으로 입력되는 음성 데이터의 문맥 정보를 학습하여 이전 통계적 모델의 한계를 극복	· 장기 의존성 문제로 매우 긴 시퀀스 데이터 처리 시 정보 손실 발생

순환 신경망 구조의 딥러닝 기반 기술은 긴 문장이나 대화에서도 비교적 정확하게 음성을 인식할 수 있게 되었지만, 매우 긴 시퀀스의 데이터 처리 시 정보 손실이 발생하는 문제점(Long-Term Dependency Problem)을 여전히 가지고 있었다. 이는 문장이 길수록, 처음에 인식한 중요한 정보들이 다음 단계, 또 다음 단계로 넘어가면서 점점 희미해지거나 사라져 버리기 때문에 발생한다. 결과적으로, AI 모델은 문장의 뒷부분을 인식할 때 맨 앞부분에 있었던 중요한 문맥 정보를 제대로 기억하거나 활용하지 못하는 것이다.

이러한 딥러닝 기반 기술의 한계를 극복하고 STT의 성능을 한 단계 더 끌어올린 것이 바로 Transformer 네트워크(Vaswani et al., 2017)이다. Transformer 네트워크는 순환 신경망 구조를 사용하지 않고 Self-Attention 메커니즘을 통해 입력 데이터 내의 모든 위치 간의 관계를 파악하여 문맥 정보를 효과적으로 포착하였다. Transformer 네트워크 방식의 주요 특징 및 장점은 크게 두 가지이다.

첫째, 병렬 처리 방식을 통한 빠른 처리이다. RNN과 달리 순차적인 연산이 필요 없어 병렬 처리가 가능하여 빠른 학습 및 추론을 할 수 있다. 이는 대규모 데이터셋 학습에 유리하며, 모델 개발 시간 및 추론 시간을 단축할 수 있다. 특히 다량의 코어를 활용하여 범용 GPU(General Purpose Graphic Processing Unit)를 통해 효과적으로 병렬 처리할 수 있는 AI 모델이라 할 수 있다.

둘째, 장거리 의존성 모델링을 통한 성능 향상이다. Self-Attention 메커니즘은 입력 시퀀스 내의 먼 위치에 있는 단어 간의 관계도 직접적으로 파악할 수 있어, 긴 문맥 정보를 효과적으로 반영한다. 이는 RNN 기반 모델의 장기 의존성 문제를 해결하는 핵심적인 요소이다. 예를 들어, 문장 "나는 대통령기록관을 직장 동료들과 함께 갔다. 다른 기록관도 많이 가보았지만, 특히 그곳에는 매우 흥미로운 기록물들이 많았다"에서 RNN 등의 순환 신경망 모델은 “대통령기록관”이라는 단어에서 멀리 떨어질수록 그 중요도가 희미해져, 긴 문장에서 맥락 파악이 어려워지는 반면, Transformer 네트워크는 "대통령기록관"에서 멀리 떨어져 언급되고 있는 "기록물"이라는 단어와의 의미적 관계도 효과적으로 파악하여 STT의 정확도를 높일 수 있다. 다양한 STT 벤치마크 테스트에서 Transformer 네트워크 기반 모델들은 기존의 RNN 기반 모델들을 압도적인 성능 차이로 능가하고 있다. 이는 Transformer 네트워크의 효율적인 정보 처리 능력과 강력한 표현력 덕분이다.

이러한 장점들로 인해 Transformer 네트워크는 현재 STT뿐만 아니라 자연어 처리(NLP), 기계 번역 등 다양한 분야에서 핵심적인 기술로 자리매김하고 있으며, 대표적인 순차적 데이터인 음성 인식 기술의 발전을 선도하고 있다. Transformer 네트워크 기반 STT 기술은 다양한 분야에서 활용되고 있다. 대표적인 예로는 Whisper 모델이

https://doi.org/10.14404/JKSARM.2025.25.3.243

ChatGPT의 음성 입력으로 사용되고 있고, 그 외에도 음성 비서 (예: Siri, Alexa, Google Assistant), 자동 자막 생성, 음성 검색, 음성 기반 질의응답 시스템, 실시간 통역, 회의록 자동 작성 등에 활용되고 있다. 이러한 응용 분야들은 Transformer 네트워크의 뛰어난 성능과 효율성을 바탕으로 사용자들에게 편리하고 혁신적인 경험을 제공하고 있다.

2.2 오픈소스 소프트웨어 및 STT 서비스

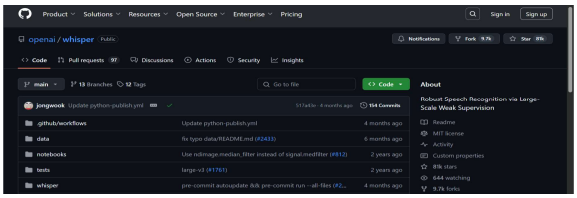
2.2.1 오픈소스 소프트웨어

최근 인공지능(AI) 분야에서 혁신적인 발전을 이끌고 있는 OpenAI社, Google社 등의 개발 주체에서 오픈소스(Open-source) 형태로 소프트웨어를 공개하는 것은 다양한 측면에서 업계에 긍정적인 영향을 미친다(권영환, 2021). 오픈소스 소프트웨어의 활성화 시 일반적인 이점을 다음과 같이 정리할 수 있다.

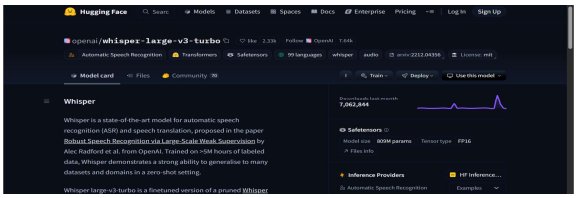
첫째, 연구 및 개발 촉진이다. 핵심 기술과 모델 구조를 공개함으로써, 전 세계의 연구자 및 개발자들이 해당 기술을 자유롭게 활용하고 개선할 수 있는 환경을 조성할 수 있다. 이는 아이디어 공유, 협업 증진, 새로운 응용 분야 발굴로 이어져 기술 혁신의 속도를 가속화하는 중요한 동력이 된다. 특히, 특정 기업이나 연구소의 독점적인 기술 개발보다는 개방적인 협력을 통해 더욱 창의적이고 다양한 접근 방식이 시도될 수 있다.

둘째, 기술 확산 및 접근성 향상이다. 오픈소스 모델은 누구나 무료로 접근하고 사용할 수 있으므로, 기술적인 제약이나 비용 부담 없이 AI 기술을 도입하고자 하는 개인, 기업, 연구 기관 등에 큰 이점을 제공한다. 이를 통해 AI 기술의 대중화를 촉진하고, 타 분야로의 기술 확산을 할 수 있다. 특히, 자금이나 인력이 부족한 스타트업이나 중소기업, 즉각적인 예산 확보가 어려운 공공 기관들에게 혁신적인 서비스를 개발하고 경쟁력을 확보하는 효율적인 방법이 될 수 있다.

셋째, 커뮤니티 구축 및 생태계 조성이다. 오픈소스 프로젝트는 개발자, 연구자, 사용자 간의 활발한 소통과 협력을 통해 발전해 나간다. 버그 수정, 기능 개선, 새로운 아이디어 제안 등 다양한 형태의 참여가 이루어지면서, 더욱 안정적이고 강력한 기술 생태계가 조성된다. 이러한 커뮤니티는 문제 해결을 위한 정보 공유, 기술 지원, 경험 교환 등 다양한 측면에서 사용자들에게 큰 도움을 제공한다. <그림 1>과 같은 GitHub, Hugging Face 등의 커뮤니티가 대표적인 예이다.



(a) GitHub



(b) Hugging Face

<그림 1> 오픈소스 커뮤니티에서 확인할 수 있는 Whisper 소스코드 및 모델

넷째, 윤리적 고려 및 투명성 증진이다. AI 모델의 작동 방식과 학습 데이터를 투명하게 공개함으로써, 모델의 예측 결과에 대한 이해도를 높이고 잠재적인 편향성이나 오류를 발견하고 개선할 수 있는 기회를 제공한다. 이는 AI 기술의 윤리적이고 책임감 있는 개발 및 활용을 위한 중요한 기반이 된다. 또한, 사용자들이 모델의 내부 동작 방식을 이해하고 보안성 및 안정성을 신뢰할 수 있도록 돕는다.

이러한 배경으로, 본 연구는 오픈소스 기반의 AI STT 기술을 선정하여 시스템을 개발하고, 그 성능을 분석하였

https://doi.org/10.14404/JKSARM.2025.25.3.243

다. 앞서 설명한 일반적인 오픈소스 소프트웨어 사용의 이점을 바탕으로 공공 기관에서 오픈소스 소프트웨어를 활용하는 주요 목적을 다음과 같이 요약할 수 있다.

1. 한정된 예산만 투입할 수 있는 공공 기관의 특성상, 집중적이고 막대한 비용이 요구되는 AI 핵심 기술 개발을 직접 수행하는 것은 불가능하다.

2. 또한, 공공 기관은 한정된 연구 인력이 다양한 업무를 동시에 수행하는 환경에 있으므로, 이들이 집중적으로 AI 기술을 연구하고 개발하는 것은 어렵다.

3. 기술 확산 및 AI 생태계 조성 등에 기여할 수 있는 오픈소스 소프트웨어를 사용하는 것은 공공의 이익을 추구해야 하는 공공 기관의 목적에 부합한다.

다만, 공공 기관은 오픈소스 소프트웨어 사용 시 <표 2>에 나타난 바와 같이 강한 카피레프트[1] 또는 퍼미시브 라이선스[2]의 법적 의무를 고려하고 준수해야 한다. 또한, 상용 소프트웨어와 달리 특정 개발자가 없으므로, 소프트웨어에 대한 깊은 이해를 바탕으로 유지보수를 책임질 수 있는 전문 인력 확보가 필수적이다. 이러한 법적 및 인력 확보 사항을 유의해야만 오픈소스 소프트웨어를 원활하게 사용할 수 있다.

<표 2> 주요 오픈소스 라이선스 비교표

라이선스명	유형	코드 공개 의무	상업적 이용	비고
GPL(GNU General Public License)	강한 카피레프트	필수	제한적	상업적 이용 시에도 소스 코드 공개 필수
MIT License	퍼미시브 라이선스	없음	가능	라이선스 고지문 배포 의무
Apache License 2.0	퍼미시브 라이선스	없음	가능	라이선스 및 소스코드 수정 사항 배포 의무
BSD License	퍼미시브 라이선스	없음	가능	라이선스 고지문 배포 의무

2.2.2 음성·텍스트 변환 서비스

현재 다양한 기업에서 고성능의 STT 모델들을 개발하여 서비스를 제공하고 있다(이해수 외, 2025). 이러한 모델은 라이선스, 사용 환경, 비용, 길이 제한, 오프라인 사용 가능 여부 등에서 차이를 보인다. <표 3>은 주요 STT 서비스들을 비교 분석한 결과이다. 이들은 모두 Transformer 기반의 최신 AI 음성인식 기술을 채용한 것으로 알려져 있으며, 매우 높은 수준의 음성·텍스트 변환 성능을 갖고 있다.

<표 3> 주요 STT 서비스 비교표

모델	라이선스	사용환경	비용	오프라인 사용	비고
Whisper	MIT	클라우드 API 및 로컬	0.36$/1시간(오프라인 무료)	가능	클라우드 또는 로컬
Google STT	상용	클라우드 API	0.24$/1시간	불가	클라우드 필수
Amazon Transcribe	상용	클라우드 API	1.8$/1시간	불가	클라우드 필수
Microsoft Azure STT	상용	클라우드 API	0.5$/1시간	불가	클라우드 필수
Naver Clova	상용	클라우드 API	960원/1시간	불가	클라우드 필수

1) Strong Copyleft: 소프트웨어의 사용·수정·배포에 대해 자유로운 퍼미시브 라이선스와 달리, 파생 저작물도 반드시 동일한 오픈 소스 라이선스로 공개해야 한다는 상속 의무가 있음

2) Permissive License: 자유로운 형태를 제공하는 라이선스를 의미하며, 소프트웨어의 사용·수정·배포에 대해 최소한의 제한만을 가지므로, 파생 저작물이 반드시 오픈소스가 되어야 한다는 의무가 없음

https://doi.org/10.14404/JKSARM.2025.25.3.243

본 연구에서는 제한된 환경에서 AI 기반의 STT 시스템을 개발하여 서비스하기 위하여, 다음과 같은 사항을 기준으로 AI STT 모델을 선정하였다.

1. 한정된 인력 및 예산을 바탕으로 AI 모델을 이용하여 자체적으로 시스템을 개발할 수 있어야 한다.
2. 효율적인 유지보수를 위하여 널리 알려진 모델이어야 하며, 퍼미시브 오픈소스 라이선스여야 한다.
3. 상용 유료 서비스와 동급 이상의 성능을 제공하기 위하여 Transformer 기반의 모델이어야 한다.
4. 비공개 기록물에 대해 음성·텍스트 변환해야 하므로, 오프라인으로 시스템을 구동할 수 있어야 한다.

결과적으로 OpenAI社의 Whisper(Radford et al., 2023) 모델을 대통령 시청각기록물에 대한 STT 기능 개발에 가장 적합한 선택이라고 판단하였으며, 그 세부적인 이유는 다음과 같다.

첫째, 높은 성능이다. Whisper 모델은 방대한 양의 다양한 음성 데이터(680,000시간 이상)를 학습하여 전반적으로 높은 음성 인식 정확도(한국어 기준 large-v3 모델 사용 시 문자 오류율 약 5.2%)를 보인다. 대표적 프롬프트 기반 생성형 AI 서비스인 ChatGPT의 음성 입력 기능에 사용되고 있는 만큼 긴 문맥을 이해하고 자연스러운 발화를 텍스트로 변환하는 능력이 매우 우수하다는 평가이다. 이는 대통령의 연설이나 담화와 같이 문맥이 중요하고 다양한 어투가 사용될 수 있는 음성 기록물에 매우 적합한 특징이다. 학습에 사용된 데이터의 양과 다양성은 모델의 일반화 성능을 높여 다양한 음향 환경과 화자의 발음에 대한 강인성을 확보하는 데 중요한 역할을 한다.

둘째, 오프라인 사용 가능 여부이다. 클라우드 기반 API 형태의 STT 서비스들은 네트워크 연결이 필수적으로, STT 처리 대상 데이터의 외부 노출에 따른 보안의 우려가 있을 수 있다. 반면, Whisper 모델은 로컬 환경에 설치하여 오프라인 구동이 가능하므로, 대통령 기록물의 민감성을 고려했을 때 각종 보안 문제에서 자유롭다.

셋째, 개방적인 라이선스이다. MIT 라이선스[3]는 수정·배포에 관대한 오픈소스 라이선스로, 상업적 이용을 포함한 거의 모든 형태의 사용을 허용하며, 파생 모델 개발 및 재배포에 대한 제약이 적다. 따라서, 내부 요구에 따른 소프트웨어 수정 및 확장 등의 유연성을 제공하며, 향후 다른 오픈소스와의 통합을 용이하게 할 수 있다.

넷째, 무료 사용이다. Whisper 모델은 무료로 공개되어 있어 구동 하드웨어만 있다면 구축 및 데이터 처리에 비용이 발생하지 않는다. 이는 예산 확보에 제약이 있을 수 있는 공공 기관에서 매우 매력적인 요소이다. 특히 누적되는 기록물에 대한 지속적인 사용에 대한 비용 부담이 없다는 점은 매우 중요하다. 이는 장기적인 관점에서 안정적인 시스템 운영 및 데이터 관리 측면에서 중요한 이점이다. 또한 지속적 업데이트를 통해 공개되는 최신 소스 코드 및 모델을 무료로 적용하여 추가적인 비용 없이 인식률 및 처리 속도의 향상도 기대할 수 있다.

2.3 대통령기록관 시청각기록관리시스템

대통령기록관은 대통령기록물의 전자적 관리를 위해 PAMS(Presidential Archives Management System)를 구축·운영하고 있으나, 방대한 양의 대통령 시청각 기록물의 특수성을 고려하여 안전하게 보존하고 효율적으로 활용하기 위하여 시청각기록관리시스템(MAM, Media Asset Management)을 별도로 운영하고 있다(김현숙, 2019). MAM은 PAMS의 하부 시스템으로 디지털(화) 파일의 포맷변환, 디지털 파일의 장기보존 등의 역할을 수행하며, PAMS 호출에 따른 보존 파일의 접근을 지원한다. 대량·대용량의 영상·음성·사진 기록물(파일)을 통합적으로 관리하고, 기록물(디지털 파일)의 장기적이고 안정적인 보존 및 활용에 필요한 기능을 제공하는 MAM은 시청각기록물의 전자적 관리를 위한 대통령기록관의 핵심 시스템이다. MAM의 주요 기능은 <표 4>와 같다.

3) 미국 매사추세츠 공과대학교에서 자교 학생들을 위해 개발되었으며, 소프트웨어의 자유로운 활용을 허용하되, 원 저작자의 저작권 고지 및 라이선스 내용을 유지할 것을 유일한 필수 조건으로 함

https://doi.org/10.14404/JKSARM.2025.25.3.243

<표 4> 대통령기록관 MAM 주요기능

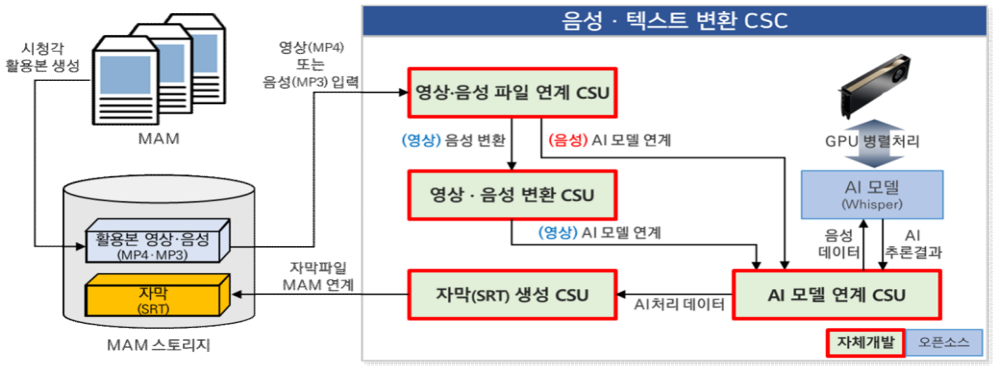
기능	설명
포맷 변환	다양한 포맷의 원본 기록물을 장기 보존에 적합한 개방형 표준의 단일 포맷으로 변환하고, 이는 새로운 기술 발생에 맞추어 일괄적인 포맷 변환을 용이하게 해준다. 동시에 서비스를 위한 저용량 포맷의 활용본을 생성한다. 결과적으로 포맷 변환을 통해 기록물의 안전한 보존과 사용자 편의성을 동시에 확보한다.
정보 추출	기록물의 메타데이터를 체계적으로 관리하기 위해 파일명, 생성일, 수정일, 해상도, 코덱 등 기술적인 정보를 추출하고 관리한다. 이러한 메타데이터는 기록물의 검색, 분류, 활용을 위한 시청각 파일 자체의 기본적인 정보 기반을 제공한다.
품질 검사	디지털(화) 파일의 품질을 보장하기 위해 다양한 자동 품질 검사를 수행한다. 예를 들어, 영상의 무음 구간, 블랙 화면 발생 여부, 재생 오류 등을 자동으로 탐지하여 알림으로서, 담당자의 육안 검사를 통해 최종 품질을 확인할 수 있도록 한다. 이는 기록물의 신뢰성을 확보하고 활용 가치를 높이는 데 중요한 과정이다.
장기 보존	MAM에서 관리되는 모든 시청각기록물은 장기적인 안전한 보존을 위해 LTO(Linear Tape-Open)와 같은 고밀도 자기 테이프에 최종적으로 원본 및 보존 파일을 저장한다. LTO 테이프는 긴 수명과 높은 저장 용량, 낮은 보관 비용 등의 장점을 가지고 있어 디지털 아카이브의 핵심 저장 매체로 활용되고 있다.

대통령기록관은 MAM을 통한 시청각기록물의 효율적인 관리 및 활용성 제고를 위해 OpenAI社의 Whisper 모델의 고속 구현체인 Faster-Whisper를 기반으로 AI 기반의 STT 기능을 자체 개발하여 MAM과 연계하였다. MAM에서 관리되는 대통령 음성·영상 기록물에 대해 자동으로 텍스트 변환을 하여, 자막파일이 생성된다.

3. 연구 방법

3.1 음성·텍스트 변환 기능 개발

본 연구에서 사용된 대통령기록관 내 STT 기능(CSC, Computer Software Component)은 총 4개의 CSU(Computer Software Units)로 구성·개발하였으며, MAM과 유기적으로 연동하여 신규 업로드되는 기록물에 대해 자동으로 텍스트 변환을 순차 처리할 수 있도록 설계되었다. 내부 구성 및 동작 흐름은 <그림 2>와 같다.



<그림 2> 음성·텍스트 변환 CSC 구성도

<표 5>에 나타난 STT 세부 기능을 통하여, 최종적으로 타임 태그가 추가된 자막 파일(SRT 포맷)이 생성된다. SRT 포맷은 시간 정보와 함께 텍스트를 포함하는 업계 표준의 자막 파일 형식으로, 영상·음성 편집 및 재생 프로그램 등 다양한 환경에서 활용될 수 있다. 생성된 자막 파일은 MAM의 별도 위치에 저장되어 해당 음성·영상 기록물과 연계되고, 사용자 호출 시 자막 파일을 다운로드 받아 <그림 3>과 같이 자막이 오버레이(Overlay) 된 상태로 영상을 재생할 수 있도록 하였다. 향후 활용을 위하여 추가로 타임 태그가 없는 순수 텍스트 형태(TXT)로도

https://doi.org/10.14404/JKSARM.2025.25.3.243

MAM에 저장하여 검색 및 분석으로의 활용 등 추가 연구에 사용할 수 있도록 하였다.

<표 5> 음성·텍스트 변환 CSC 구성 CSU

CSU	입력	출력	세부 기능
영상·음성 파일 연계 CSU	[from MAM] 포맷 변환 파일(MP4 또는 MP3)	[to 영상·음성 변환 또는 AI 모델 연계 CSU] 영상(MP4) 및 음성(MP3) 파일 위치 정보	· MAM 내 MP4 또는 MP3 파일 생성 감지 · 파일 유형(음성/영상) 확인 · 음성 코덱 미포함 MP4 파일의 경우 다음 파일 처리
영상·음성 변환 CSU	[from 영상·음성 파일 연계 CSU] 영상 파일 위치 정보	[to AI 모델 연계 CSU] 임시 음성 파일(MP3) 위치 정보	· FFmpeg 라이브러리 활용, MP4 파일에서 음성 트랙(MP3)을 추출하여 임시 음성 파일 생성
AI 모델 연계 CSU	[from 영상·음성 파일 연계 또는 영상·음성 변환 CSU] 음성 파일 위치 정보	[to 자막 생성 CSU] STT 텍스트 추출 결과	· Whisper 모델 호출을 통한 STT 수행 · Whisper 모델 구동 관련 설정 관리 · 텍스트 추출 완료 후 영상의 임시 음성 파일 삭제
자막 생성 CSU	[from AI 모델 연계 CSU] 텍스트 추출 결과	[to MAM] 타임 태그 포함 자막 파일(SRT)	· 타임 태그 추가를 통한 SRT 포맷 자막 파일 생성 · 타임 태그 없는 텍스트(TXT) 추가 생성(검색·분석용)

(a) 팟플레이어(카카오) (b) VLC Media Player(VideoLAN Organization)

<그림 3> 상용 프로그램을 통한 영상 재생 중 자막 오버레이

3.2 시스템 구성 환경

STT 동작 서버의 운영 체제는 Windows Server 2022를 사용하였으며, AI 모델 구동을 위한 GPU는 NVIDIA A40(쿠다코어 10,572개, VRAM 48GB)을 선정하였다. Transformer 네트워크 기반 Whisper의 GPU 가속을 활용하여 대량의 기록물을 빠르고 효율적으로 처리하기 위하여 다량의 GPU 코어 및 고용량의 비디오 메모리(VRAM)를 탑재한 워크스테이션급 이상의 제품이 요구된다. 특히 48GB의 비디오 메모리는 대용량 AI 모델(large)을 안정적으로 처리하는 데 필수적이다.

AI 모델을 동작하는 프레임워크는 Faster-Whisper를 사용하였는데, 이는 OpenNMT CTranslate2기반으로 Whisper를 개선한 구현체로서 AI 모델의 고속 추론이 가능하여 대량의 STT 처리를 빠르게 수행할 수 있도록 한다. Whisper는 하드웨어 요구사양에 따른 다양한 버전의 학습완료 모델을 제공하고 있으며, 파라메터의 크기(tiny39M ~ large1550M)에 비례하여 텍스트 추출의 성능이 상승하나, 처리시간 및 필요 하드웨어 사양도 함께 증가한다. 따라서, 본 연구에서는 대량의 시청각기록물의 텍스트 변환을 위하여 성능이 준수하고 속도도 상대적으로 빠른 turbo(large-v3-turbo)를 세부 모델로 선정하였다. Whisper는 large 및 turbo 모델을 기준으로 2025년 4월 기준 버전 3.0까지 배포되었다. 추가로, 본 연구에서는 STT 구동 및 시험을 위해 FFmpeg(영상 파일로부터

 https://doi.org/10.14404/JKSARM.2025.25.3.243

음성 추출), librosa(음성의 품질 분석), torch(Faster-Whisper를 GPU에서 구동) 등의 라이브러리 및 프레임워크를 사용하였다.

3.3 시험 대상

본 연구의 성능 분석 시험을 위해 대통령별 약 5분 분량의 음성·영상 기록물 각 3개씩 총 34개(이승만 대통령은 한글 육성 기록 1개)를 선정하였다. 다양한 대통령의 발화 특징 및 음향 환경에서의 모델 성능을 평가하기 위해 대통령들의 연설, 담화 등 주요 음성 기록을 포함하였다. 아날로그 매체의 경우 디지털화한 파일을 사용하였으며, 입력 포맷은 48kHz 스테레오 MP3 또는 동 규격의 음성 코덱을 포함한 MP4 파일이다. 대통령별 시험 대상 시청각 기록물 목록은 <표 6>과 같다. 오픈릴 테이프는 최초의 음성기록 자기 테이프 매체이며, U-matic은 최초의 상업용 카세트형 비디오 매체이다. 이들은 1970년대 이전 환경에서 녹음되어 음질이 떨어지는 편이다. 카세트 테이프는 플라스틱 카세트 안에 테이프가 있어 편리하고 휴대성이 뛰어났으며, 녹음장비의 발달로 일반적으로 좋은 음질을 갖고 있다. XDCAM은 광학 디스크나 플래시 메모리를 사용하는 전문가용 디지털 비디오 포맷으로, 아날로그 매체 대비 매우 뛰어난 음질을 갖고 있으며, 디지털 파일인 MP4는 고화질 콘텐츠를 작은 용량으로 압축하는 파일 형식으로, 인터넷 스트리밍과 다양한 기기에서 재생이 가능해 현재 가장 보편적으로 사용되는 포맷이다.

<표 6> 시험 대상 기록물 주요 정보(재임 순)

No.	대통령	생산년도	원본 매체 종류	수량	세부 자료 유형
1	이승만	1942	오픈릴 테이프	1	취임·퇴임사·특별 성명 등 공식 발표문 및 기타 육성녹음 자료 등
2	윤보선	1974	오픈릴 테이프	3	
3	박정희	1972~1974	카세트 테이프	3	
4	최규하	1979~1980	U-matic, 카세트 테이프	3	
5	전두환	1983~1987	오픈릴 테이프, 카세트 테이프	3	
6	노태우	1988~1990	카세트 테이프	3	
7	김영삼	1993~1997	오픈릴 테이프, 카세트 테이프	3	
8	김대중	1998~2001	카세트 테이프	3	
9	노무현	2003~2006	카세트 테이프	3	
10	이명박	2008	카세트 테이프, XDCAM, 디지털파일(MP4)	3	
11	박근혜	2014~2015	XDCAM	3	
12	문재인	2017~2022	디지털파일(MP4)	3	

3.4 추출 및 산출 데이터

본 연구에서는 개발한 STT 기능의 성능을 객관적으로 평가하고, 인식 오류에 영향을 미치는 요소를 분석하기 위해 다음과 같은 데이터를 추출 및 산출하였다.

3.4.1 STT 처리를 통한 텍스트 추출 결과물

Faster-Whisper 모델을 이용하여 각 시험 대상 음성 기록물에 대해 자동 음성 인식(STT)을 수행하고, 그 결과를 텍스트 형태로 저장하였다. 실제 구축한 시스템은 타임 태그가 처리된 자막 파일(SRT)을 생성하나, 정답(GT, Ground Truth) 텍스트와의 비교를 통한 오류율 분석을 위하여, 타임 태그가 제외된 텍스트(TXT)를 별도

https://doi.org/10.14404/JKSARM.2025.25.3.243

추출하여 시험에 사용하였다.

3.4.2 문자 오류율

결과의 정확도를 측정하기 위해, 사람이 직접 청취하여 수동으로 작성한 GT 텍스트와 STT 결과 텍스트를 비교하여 문자 오류율(CER, Character Error Rate)을 산출하였다. CER은 인식된 문자와 정답 문자 간의 편집 거리(Edit Distance), 즉 삽입, 삭제, 치환된 문자 수의 합을 정답 문자의 총수로 나누어 백분율로 나타낸 지표이다. CER 값이 낮을수록 STT 성능이 우수함을 의미한다. CER은 식 (1)과 같이 Levenshtein 거리를 사용하여 계산한다. 여기에서 S는 치환된 문자 수 (Number of Substitutions), D는 삭제된 문자 수 (Number of Deletions), I는 삽입된 문자 수 (Number of Insertions), N은 정답 텍스트의 총 문자 수 (Total Number of Characters in the Reference)를 나타낸다.

$$CER = \frac{S + D + I}{N} \times 100 \qquad (1)$$

3.4.3 발화 속도

음성 인식 오류율에 화자의 말하는 속도가 미치는 영향을 분석하기 위해, 각 시험 대상 기록물에서 화자의 평균 발화 속도(CPM, Characters Per Minute)를 측정하였다. 이는 식 (1)과 같이 수동으로 작성된 GT 텍스트의 총 문자수를 해당 음성 기록물의 길이(분)로 나누어 계산하였다. CPM는 화자의 말의 빠르기를 나타내는 지표이며, 화자의 말하는 속도가 빠를 경우 CPM은 증가하고, 반대의 경우 CPM이 감소한다.

$$CPM = \frac{\text{총 문자 수}(Character)}{\text{총 시간}(Minute)} \qquad (2)$$

3.4.3 음성의 선명도

음성 기록물의 음질이 인식 성능에 미치는 영향을 분석하기 위해, 음향 특징 중 하나인 Spectral Contrast를 추출하였다. Spectral Contrast는 오디오 신호의 각 프레임에서 주파수 스펙트럼의 피크(peak)와 밸리(valley) 간의 차이를 나타내는 지표로, 음성의 선명도 및 음질과 관련이 있다. 일반적으로 Spectral Contrast 값이 높을수록 음질이 좋고, 배경 잡음이 적음을 의미한다. 본 연구에서는 딥러닝 기반의 Librosa 라이브러리를 이용하여 각 음성 기록물로부터 Spectral Contrast 특징 벡터를 추출하고, 식 (3)과 같이 프레임별 Contrast 값의 평균값을 해당 기록물의 Spectral Contrast 값으로 사용하였다. M_i는 각 주파수 대역의 최댓값을 나타내며, m_i는 각 주파수 대역의 최소값, n은 주파수 대역 수를 의미한다. Spectral Contrast는 인간의 청각 시스템이 음색을 인지하는 방식과 유사하게, 시간의 흐름에 따라 변화하는 주파수 성분의 대비를 분석하여 음질을 정량화하는 데 유용하다.

$$Spectral\ Contrast = \frac{\sum_{i=1}^{n} (M_i - m_i)}{n} \qquad (3)$$

3.4.4 오류율과의 상관 계수

또한, 추출한 CER과 CPM 및 Spectral Contrast와의 상관관계를 분석하기 위하여, 상관 계수를 산출하였다. 상관 계수는 피어슨 상관 계수(Pearson Correlation Coefficient ,PCC)를 의미하며, 이는 두 변수 간의 선형 상관관계를 계량화하기 위한 수치이다. 변수 X, Y의 모집단 피어슨 상관 계수$\rho_{X,Y}$는 식 (4)와 같이 코시-슈바르츠 부등식에 의해 구할 수 있으며, $Cov(X, Y)$는 변수 X, Y의 공분산, $\sigma_X \sigma_Y$는 변수 X, Y의 모집단 표준편차를

https://doi.org/10.14404/JKSARM.2025.25.3.243

나타낸다. 피어슨 상관 계수는 +1과 −1 사이의 값을 가지며. +1은 완벽한 양의 선형 상관 관계인 반면, −1은 완벽한 음의 선형 상관 관계를 의미한다. 만약 발화속도 및 Spectral Contrast와 CER과의 피어슨 상관 계수가 0에 가깝다면, 선형적인 상관 관계가 없다고 판단할 수 있다.

$$\rho_{X,\,Y} = \frac{Cov(X,\,Y)}{\sigma_X \sigma_Y} \qquad (4)$$

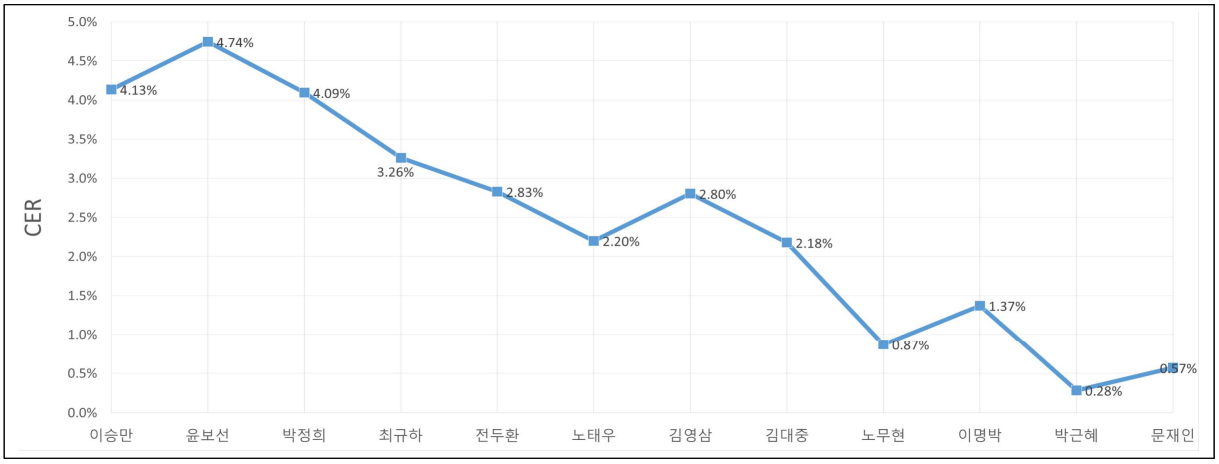
4. 시험 결과

4.1 문자 오류율 및 상관관계 분석

대통령별 음성 기록물에 대해 Whisper 모델을 이용하여 STT를 수행하고, 수동으로 작성된 GT 텍스트와 비교하여 CER을 산출한 결과, 개별 대통령의 CER 결과는 <표 7>과 같으며, 그 평균은 2.44%로 매우 낮은 오류율을 나타냈다. 이는 음절 100개당 2~3개의 오류가 있음을 의미하며, STT 기능이 실제 대통령 음성 기록물의 활용에 충분한 수준의 인식 성능임을 시사한다. 또한 대통령별 CER 그래프는 <그림 4>와 같으며, 역대 대통령 재임 순으로 오류율이 낮아지는 경향을 보였다. 우선 기록물의 원본 매체를 기준으로 그 결과를 파악하면, 오픈릴 테이프, 카세트 테이프 등 아날로그 매체가 대부분인 노무현 대통령 이전과, 디지털 파일이 많은 그 이후 대통령으로 구분하였을 때 디지털 매체에서 대체로 낮은 오류율을 보였음을 알 수 있다. 또한, 시험을 통하여 문자 오류율 및 상관관계 분석 결과, 오류율에는 녹음 상태, 즉 음성의 품질 및 지도학습 기반의 AI에서 인식률을 떨어뜨리는 요소(표준어 사용 여부 및 시대별 사용 언어 차이 등)가 영향을 끼친 것으로 확인하였다.

<표 7> 대통령별 CER 계산 결과(%)

이승만	윤보선	박정희	최규하	전두환	노태우	김영삼	김대중	노무현	이명박	박근혜	문재인	평균
4.13	4.74	4.09	3.26	2.83	2.20	2.80	2.18	0.87	1.37	0.28	0.57	2.44



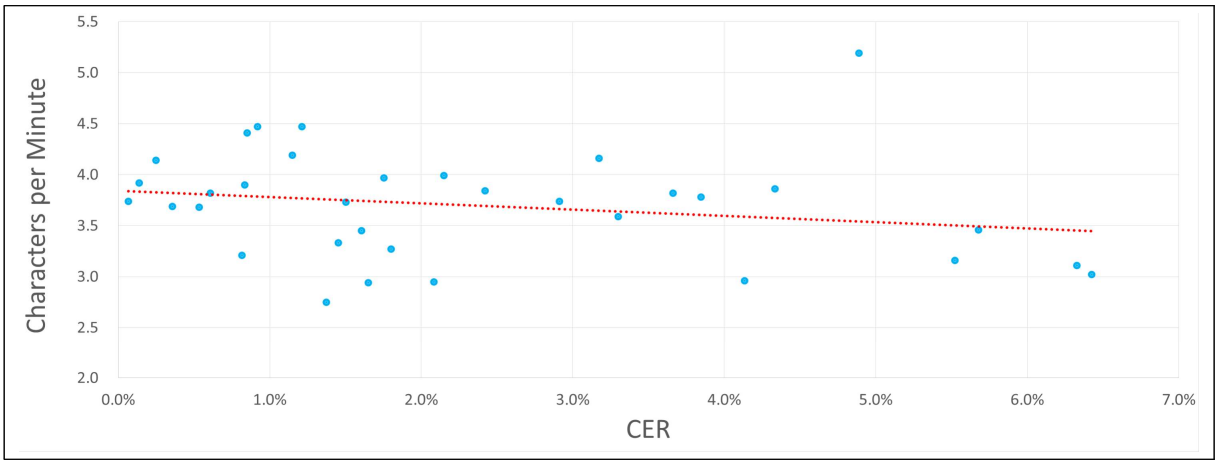
<그림 4> 대통령별 CER 그래프

4.1.1 CER과 발화 속도(CPM) 상관관계 분석

<그림 5>에 나타난 바와 같이 CPM과 CER과의 상관관계를 비교 분석한 결과, 대통령별 CER과 CPM 간에는 뚜렷한 상관관계를 발견하기 어려웠다. 추세선 역시 평행에 가까운 모습을 보여주고 있다. 발화 속도가 빠른 대통

https://doi.org/10.14404/JKSARM.2025.25.3.243

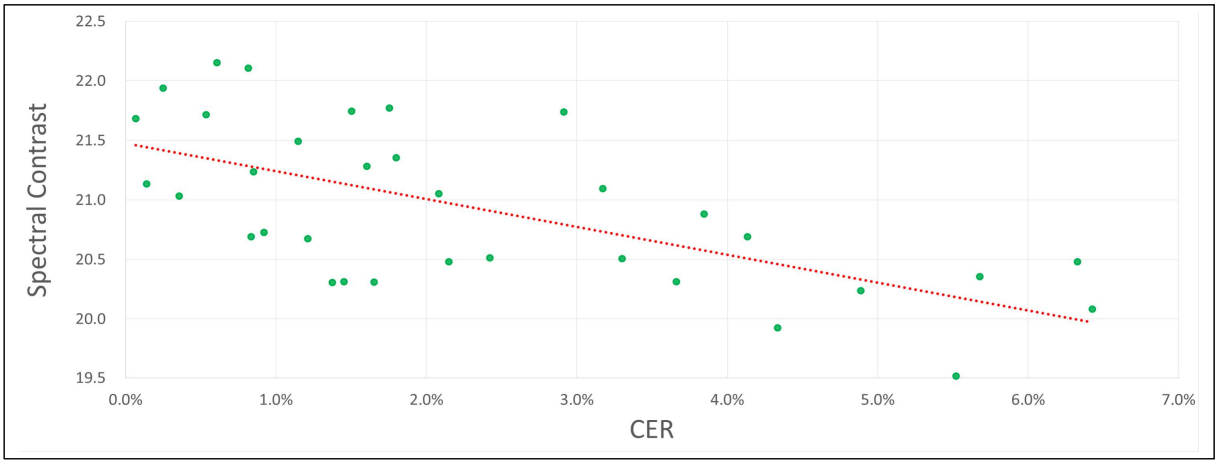
령의 기록에서 반드시 CER이 높게 나타나는 것도 아니며, 발화 속도가 느린 대통령의 기록에서 CER이 낮게 나타나는 경향도 보이지 않았다. 피어슨 상관 계수 역시 "-0.21"로 "0"에 근접한 수치를 보였다. 이는 발화 속도 자체보다는 개별 발음의 명확성, 억양의 특징 등 다른 요인들이 인식 성능에 더 큰 영향을 미칠 수 있음을 시사한다. 다만, 사람조차도 인지하기 어려울 정도로 극단적으로 빠르거나 느린 발화 속도의 경우 인식 오류를 증가시킬 가능성은 여전히 존재한다.



<그림 5> CER 대비 CPM의 분포도 및 추세선

4.1.2 CER과 Spectral Contrast 상관관계 분석

<그림 6>은 CPM과 Spectral Contrast와의 상관관계를 비교 분석한 결과를 보여주고 있다. 이를 통해 Spectral Contrast 값이 낮을수록 CER 값은 높아지는 추세를 확인할 수 있다. 추세선이 비교적 뚜렷한 하방을 보이고 있으며, 이는 음성 기록물의 녹음 품질이 STT 성능에 유의미한 영향을 미칠 수 있음을 시사한다. 피어슨 상관 계수 역시 "-0.64"로 유의미한 값을 나타내었다. Spectral Contrast가 높다는 것은 음성의 선명도가 높고 배경 잡음이 상대적으로 적다는 것을 의미하며, 이는 AI 모델이 음성 정보를 정확하게 파악하는 데 유리한 환경을 제공하는 것으로 해석할 수 있다. 다만, 녹음 상태뿐만 아니라 발음의 명확성, 억양의 특징 등 다른 요인들 역시 인식 성능에 더 큰 영향을 미칠 수 있기 때문에 Spectral Contrast와 CER이 완전히 비례관계에 있지는 않다.



<그림 6> CER 대비 Spectral Contrast의 분포도 및 추세선

https://doi.org/10.14404/JKSARM.2025.25.3.243

4.2 오류 발생 요인 분석

Whisper 모델은 지도학습 AI이며, 기본적으로 표준어 기반이고 현재 사용되는 언어 중심으로 학습되는 것으로 알려져 있다. 대통령 시청각 기록물에 시험적으로 수행한 STT 결과에서도 Whisper 모델 학습 환경이 오류 발생 주요 요인의 하나임을 파악하였다. 표준어 사용 여부 및 시대별 사용 언어 차이 등에 따른 오류 사항을 실제 시청각 기록물을 통한 구체적 STT 사례로 제시하고, 현재 구축한 STT 기능의 강점과 한계를 분석한다.

4.2.1 표준어 사용 환경에서의 높은 정확도

<표 8>의 예시에서 볼 수 있듯이, Whisper 모델은 비교적 현대에 주로 사용되는 표준어 사용 환경에서 매우 높은 정확도를 보인다. 특히, "2003년 8월", “건국 60주년”, "IT화"와 같이 숫자가 포함된 단어나 관용적으로 사용되는 영문 약어 등 다소 복잡한 어휘나 구문도 맥락을 정확하게 파악하여 오류 없이 텍스트로 변환하는 Transformer 모델의 강력한 성능을 확인할 수 있다. 이는 Transformer 아키텍처가 문장 전체의 의미를 파악하고, 그 맥락 안에서 단어를 선택하는 능력 덕분이다. 일부 오류는 사람이 청취하기에도 발음이 혼동될 수 있는 “4”와 “차”, “른”과 “는” 같은 글자에서 발견되었다.

<표 8> 표준어 사용 환경에서의 높은 정확도 예시

STT 결과	Ground Truth
국민 여러분 그런데 앞으로 5년 후는 그리고 10년 후는 어떻게 될까 혹시 중국에게 밀리는 것은 아닐까 이런 걱정을 하는 분들도 많이 계십니다 그러나 국민 여러분 너무 걱정하지 마십시오 우리도 그냥 손 놓고 있지는 않습니다 우리 하기 나름 아니겠습니까 정부도 대비하고 있습니다 이미 2003년 8월에 4세대 10대 성장동력을 선정해서 집중적인 투자를 하고 있습니다 부품 소재산업 그리고 전통산업의 IT화 금융 물류 서비스 산업도	국민 여러분 그런데 앞으로 5년 후는 그리고 10년 후는 어떻게 될까 혹시 중국에게 밀리는 것은 아닐까 이런 걱정을 하는 분들도 많이 계십니다 그러나 국민 여러분 너무 걱정하지 마십시오 우리도 그냥 손 놓고 있지는 않습니다 우리 하기 나름 아니겠습니까 정부도 대비하고 있습니다 이미 2003년 8월에 차세대 10대 성장동력을 선정해서 집중적인 투자를 하고 있습니다 부품 소재산업 그리고 전통산업의 IT화 금융 물류 서비스 산업도
올해로 대한민국 건국 60주년을 맞이합니다. 우리는 잃었던 땅을 되찾아 나라를 세웠고, 그 나라를 지키려고 목숨을 바쳤습니다. 모두가 하나같이 열심히 살았습니다. 그리하여 세계 역사상 최단기간에 산업화와 민주화라는 과업을 동시에 이뤄내었습니다. 오로지 우리의 의지와 우리의 힘으로 일구어내었습니다 지구상에서 가장 가난했던 나라가 세계 10위권의 경제대국이 되었습니다. 도움을 받던 나라에서 이제는 베푸른 나라로 바뀌었	올해로 대한민국 건국 60주년을 맞이합니다. 우리는 잃었던 땅을 되찾아 나라를 세웠고, 그 나라를 지키려고 목숨을 바쳤습니다. 모두가 하나같이 열심히 살았습니다. 그리하여 세계 역사상 최단기간에 산업화와 민주화라는 과업을 동시에 이뤄내었습니다. 오로지 우리의 의지와 우리의 힘으로 일구어내었습니다 지구상에서 가장 가난했던 나라가 세계 10위권의 경제대국이 되었습니다. 도움을 받던 나라에서 이제는 베푸는 나라로 바뀌었

4.2.2 방언 처리

<표 9>에 나타난 예시는 특정 지역의 방언이 포함된 발화에 대한 STT 결과이다. “금융실맹제”는 “금융실명제”를 나타내는 등 경상도 방언으로 인해 모음이 바뀌어 인식된 것을 확인할 수 있다. 또한, “체홉”, “조명희”, “연해주”, “썼습니다”와 같이 “ㅎ” 또는 “ㅆ”같이 일부 방언에서 자음이 약하게 들리는 이유로 인식 오류가 발생하였다. 앞에서 언급한 것처럼 Whisper 모델은 표준어 기반 학습으로, 방언에 대해서는 정확한 인식이 어렵고 표준어와 유사한 발음으로 오인하여 텍스트로 변환하는 경향을 보인다. 여기서 대형 AI 모델일지라도 훈련 데이터에 대한 편향을 가지는 것을 알 수 있으며, 다양한 훈련 데이터를 통해 편향성을 줄임으로써 그 성능을 높일 필요가 있음을 확인할 수 있다. 다만, 방언의 경우 실제 사람이 청취하는 것과 같은 텍스트로 추출되고 있고, 맞춤법이 틀렸다는 이유로 오류로 판단할 것인지에 대한 문제는 추가적인 검토가 필요하다.

https://doi.org/10.14404/JKSARM.2025.25.3.243

<표 9> 방언 처리 예시

STT 결과	Ground Truth
친애하는 국민 여러분, 드디어 우리는 금융실**맹**제를 실시합니다. 이 시간 이후 모든 금융거래는 실**맹**으로만 이루어집니다. 금융실**맹**제가 실시되지 않고는 이 땅의 부정부패를 원천적으로 봉쇄할 수가 없습니다. 정치와 경제의 검은 유착을 근원적으로 단절할 수가 없습니다 금융실**맹**거래의 정**책**이 없이는 이 땅의 진정한 분배정의를 구현할 수가 없습니다. 우리 사회의 도덕성을 **항**립할 수가 없습니다. 금융실**맹**제 없이는 건강한 민주주의도,	친애하는 국민 여러분, 드디어 우리는 금융실**명**제를 실시합니다. 이 시간 이후 모든 금융거래는 실**명**으로만 이루어집니다. 금융실**명**제가 실시되지 않고는 이 땅의 부정부패를 원천적으로 봉쇄할 수가 없습니다. 정치와 경제의 검은 유착을 근원적으로 단절할 수가 없습니다 금융실**명**거래의 정**착**이 없이는 이 땅의 진정한 분배정의를 구현할 수가 없습니다. 우리 사회의 도덕성을 **확**립할 수가 없습니다. 금융실**명**제 없이는 건강한 민주주의도,
러시아의 대문호 도스**텝**스키, 톨스토이와 함께 극동과 사할린을 문학에 담아낸 러시아 작가 안톤 체**업**을 한국인은 매우 사랑합니다. 이곳은 한국 문학에서도 중요한 공간이기도 합니다. 한국의 근대 소설가 이광수의 작품 유정은 시베리아와 바이칼 호수를 배경으로 하고 있습니다. 작가 조명**이**는 연**예**주에 살면서 이곳의 삶을 소설로 **섰**습니다	러시아의 대문호 도스**토옙**스키, 톨스토이와 함께 극동과 사할린을 문학에 담아낸 러시아 작가 안톤 체**홉**을 한국인은 매우 사랑합니다. 이곳은 한국 문학에서도 중요한 공간이기도 합니다. 한국의 근대 소설가 이광수의 작품 유정은 시베리아와 바이칼 호수를 배경으로 하고 있습니다. 작가 조명**희**는 연**해**주에 살면서 이곳의 삶을 소설로 **썼**습니다.

4.2.3 1980년대 이전의 용어 처리

<표 10>은 1980년대 이전에 주로 사용되어 현대에는 과거 대비 사회 환경 변화 등으로 주로 사용되지 않는 단어들의 인식 결과를 보여주고 있다. “김일성 집단”, “버마 폭거”, “위해”, “습격” 등 냉전시대에 주로 사용된 단어나 상대방을 표현하는 “저희”, 현재는 사용 빈도가 적은 “궤휼” 등은 Whisper 모델의 학습에 많이 포함되지 못한 것으로 보인다. 모델이 학습한 데이터의 빈도수가 낮을 경우 인식 오류가 발생할 가능성이 있으며, 특히 역사 기록물과 같이 오래된 어투나 표현이 사용된 경우에는 이러한 문제가 더욱 두드러질 수 있다. 이를 해결하기 위해서는 다양한 시대의 언어 데이터를 학습에 포함시켜 전이학습(Transfer Learning) 하는 방법도 고려할 수 있다.

<표 10> 1980년대 이전의 용어 처리 예시

STT 결과	Ground Truth
친애하는 국민 여러분 북한 **주민선** 집단이 저지른 이번 **범화** 폭**허**는 국가 원수인 본인에 대한 **이**해의 기도로서 세상 말씀드릴 필요도 없이 곧 우리 대한민국에 대한 선전포고나 다름없는 중대한 도발 행위인 것입니다. 이것은 우리의 생존과 **알**령을 파괴하려는 침략 전쟁의 순엄으로서 자유권의 발동을 통하여 **은**징, 보복을 받아 마땅한 전쟁 도발이 아닐 수 없는 것입니다. 우리는 그동안 카나다와 필리핀 순방 등 본인에 대한 **이**해의 기도와 1.21 청와대 습**적** 기도 등 저들의 무수한 **고**발에 대하여 인내와 자조로서 대처해 왔습니다. 그것은 우리의 힘이 약하기 때문이 아니라 우리의 **통**화의 의지가 그만큼 강했기 때문이며	친애하는 국민 여러분 북한 **김일성** 집단이 저지른 이번 **버마** 폭**거**는 국가 원수인 본인에 대한 **위**해의 기도로서 세상 말씀드릴 필요도 없이 곧 우리 대한민국에 대한 선전포고나 다름없는 중대한 도발 행위인 것입니다. 이것은 우리의 생존과 **안**령을 파괴하려는 침략 전쟁의 순엄으로서 자유권의 발동을 통하여 **응**징, 보복을 받아 마땅한 전쟁 도발이 아닐 수 없는 것입니다. 우리는 그동안 카나다와 필리핀 순방 등 본인에 대한 **위**해의 기도와 1.21 청와대 습**격** 기도 등 저들의 무수한 **도**발에 대하여 인내와 자조로서 대처해 왔습니다. 그것은 우리의 힘이 약하기 때문이 아니라 우리의 **평**화의 의지가 그만큼 강했기 때문이며
공화를 위하여 피를 흘리는 모든 친구들이여 저 외적의 거짓 선전에 속지 마시오 낙심 마시오. 저 외적이 백방으로 거짓 선전을 해서 우리를 속이려는 것입니다. 저희가 지금 전쟁을 이긴다고 큰소리를 합니다. 저**의** 군사는 세계에 이길 나라**에** 없다고 자랑합니다. 저희가 얼마 안에 세계를 다 전복한다고 떠듭니다. 이렇게 해서 우리들이 **고지** 듣고 낙심 낙망해서 싸움을 정지하고 항복하라는 것입니다. 그러므로 우리가 속지 말아야 됩니다. 저 1위는 궤**율**과 거짓으로 주장을 삼는 자들입니다. 이 전쟁에 저희가 얼마 성공했다는 것이 모두 궤**율**과 거짓으로 얻은 것입니다.	공화를 위하여 피를 흘리는 모든 친구들이여 저 외적의 거짓 선전에 속지 마시오 낙심 마시오. 저 외적이 백방으로 거짓 선전을 해서 우리를 속이려는 것입니다. 저희가 지금 전쟁을 이긴다고 큰소리를 합니다. 저**희** 군사는 세계에 이길 나라**이** 없다고 자랑합니다. 저희가 얼마 안에 세계를 다 전복한다고 떠듭니다. 이렇게 해서 우리들이 **곧이** 듣고 낙심 낙망해서 싸움을 정지하고 항복하라는 것입니다. 그러므로 우리가 속지 말아야 됩니다. 저 1위는 궤**휼**과 거짓으로 주장을 삼는 자들입니다. 이 전쟁에 저희가 얼마 성공했다는 것이 모두 궤**휼**과 거짓으로 얻은 것입니다.

https://doi.org/10.14404/JKSARM.2025.25.3.243

5. 결론

본 연구는 오픈소스를 기반으로 하여 AI 음성·텍스트 변환(STT) 기능을 개발하고, 대통령의 시청각기록물에 적용하여 그 성능을 심층적으로 분석하였다. 시험 결과, 전반적으로 평균 3% 이하의 매우 낮은 문자 오류율(CER)을 확인하여, 오픈소스 기반 AI STT 기술이 긴 세대에 걸쳐 생산되고 있는 대통령 시청각기록물의 텍스트 변환 작업에 충분히 활용 가능성이 높음을 확인 할 수 있었다. 다만 아무리 오류율이 낮을지라도 기록물의 원문 보존 측면에서 소량의 오류가 끼칠 수 있는 영향성뿐만 아니라, AI로 인해 나타날 수 있는 윤리적인 문제를 명확히 검토하여 기술을 활용할 필요가 있음을 인지해야 한다. 본 연구를 통해 변환된 텍스트는 내부 기록관리 목적으로 현재 활용하고 있으며, 향후 추가적인 검토를 통해 확대·적용할 것이다.

실제 음성과 인식 문자 간의 차이를 분석한 결과, 발화 속도는 인식 성능에 뚜렷한 영향을 미치지 않는 것으로 나타났으나, 음성의 뚜렷함을 나타낼 수 있는 Spectral Contrast는 CER과 유의미한 상관관계를 보이는 것으로 분석되었다. 즉, 녹음 품질이 좋을수록 오류율이 떨어지는 경향을 확인하였다. 또한, 표준어 사용 환경에서는 높은 인식률을 보였으나, 방언이나 현대에 잘 사용되지 않는 어휘에 대해서는 일부 오류가 발생하는 것을 확인할 수 있었다. 이는 학습 데이터의 편향성 및 언어 변화에 대한 적응력 향상의 필요성을 시사한다.

본 연구를 통해 빠르게 발전하는 AI 기술을 대통령 기록물 관리에 적극적으로 적용하기 위해 오픈소스 기반 AI 기술을 활용하는 것이 매우 타당하다는 결론을 내릴 수 있다. OpenAI社의 Whisper와 같은 오픈소스 모델은 높은 성능을 제공하면서도 무료로 사용 가능하며, 필요에 따라 시스템을 자체적으로 구축하고 확장할 수 있는 유연성을 제공한다. 또한, Hugging Face와 같은 AI 모델 플랫폼을 통해 지속적으로 공개되는 새롭게 추가 학습되고 개선된 AI 모델들을 즉시 무료로 활용할 수 있다는 장점도 가지고 있다.

마지막으로, 정책·학술·교육적 관점으로 구분하여 본 연구의 시사점을 도출하였다.

1. 정책적 관점: 오픈소스 기반 AI 기술을 공공기록물 관리에 적극 도입함으로써 기술 독립성을 확보하고, 예산 절감을 실현 가능. 특히, 공공 부문에서의 기술 독립성 확보는 장기적인 디지털 주권 확보와 직결됨

2. 학술적 관점: AI 기반 STT 성능에 영향을 미치는 요인에 대한 추가적인 언어학적·정보과학적 분석 및 Whisper와 같은 모델을 한국어에 특화된 학습 데이터로 개선하는 과제 등이 연계될 수 있음

3. 교육적 관점: AI 음성인식 기술을 활용한 교육용 콘텐츠 자막화 및 검색 기능 등이 가능해져, 학습자 맞춤형 교육자료 제공 등이 쉬워질 수 있으며, 기록물을 활용한 디지털 교육 활용 가능성 증가

향후 이번 STT 기능 개발을 확장하여, 시청각기록물로부터 추출된 텍스트를 기록관리 및 서비스의 편의성 증가에 활용할 수 있을 것이다. Whisper와 같은 Transformer 네트워크 기반의 BERT(Bidirectional Encoder Representations from Transformers) 기술을 이용해 추출된 텍스트를 요약하여 제공하거나, 변환된 텍스트를 이용한 효율적인 전문 검색 기능을 추가로 검토해 볼 수 있다. 더 나아가, 영상 처리 분야에서도 오픈소스 AI 기술을 적극적으로 활용하여 대통령 시청각기록물의 품질 향상 및 활용 가치 증진을 위한 연구를 수행할 필요가 있다. 예를 들어, 오래된 흑백 영상을 컬러 영상으로 복원하거나, 저해상도 영상을 고해상도로 개선하는 등의 기술은 기록물의 보존 및 활용 가치를 높이고 사용자들에게 더욱 생생한 경험을 제공할 수 있을 것이다.

참고문헌

권영환 (2021). 오픈소스 활성화를 위한 오픈소스 연구개발 생태계 연구(RE-102). 소프트웨어정책연구소.

 https://doi.org/10.14404/JKSARM.2025.25.3.243

김현숙 (2019). 디지털기반 대통령기록관리체계 모델 재설계. 기록관리 이슈페이퍼, 7, 1-17.

나미선, 한상효 (2016). 대통령기록관 신청사 보존·복원 인프라 구축. 기록인, 34, 22-31.

남태우, 오지영, 유보현 (2007). 대통령기록물관리법에 관한 연구. 한국기록관리학회지, 7(2), 165-188. https://doi.org/10.14404/JKSARM.2007.7.2.165

안대진 (2017). 지능형 기록정보서비스 방안. 기록인, 41, 38-45.

이해수, 유재흥, 안미소, 안성원 (2025). 2024년 국내외 인공지능 산업 동향 연구(RE-189). 소프트웨어정책연구소.

FFmpeg (n.d.). FFmpeg Website. Available: https://ffmpeg.org

GitHub (n.d.). GitHub Website. Available: https://github.com

Hugging Face (n.d.). Hugging Face Website. Available: https://huggingface.co

Malik, M., Kamran, M., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. Multimedia Tools and Applications, 80(6), 9411-9457. https://doi.org/10.1007/s11042-020-10073-7

McFee, B., Colin, R., Dawen, L., Daniel P.W. E., Matt, M., Eric, B., & Oriol, N. (n.d.). Librosa. Librosa Documentation. Available: https://librosa.org/doc/latest/index.html

Mikolov, T., Karafiát, M., Burget, L., Cernocky, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In Interspeech, 1045-1048. https://doi.org/10.21437/Interspeech.2010-343

OpenNMT (n.d.). CTranslate2. OpenNMT. Available: https://opennmt.net/CTranslate2

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286. https://doi.org/10.1109/5.18626

Radford, A., Kim, Jong wook, Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. International Conference on Machine Learning, ICML 2023 Proceedings of Machine Learning Research, 202, 28492-28518. https://doi.org/10.48550/arXiv.2212.04356

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 6000-6010. https://doi.org/10.48550/arXiv.1706.03762

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

An, Dae-jin (2017). Plans for Intelligent Archival Information Services. The Korean Archivist, 41, 38-45.

Kim, Hyunsook (2019). Redesigning the Digital-Based Presidential Records Management System Model. Archival Management Issue Paper, 7, 1-17.

Kwon, Younghwan (2021). A Study on the Open Source R&D Ecosystem for Open Source Activation(RE-102). Software Policy & Research Institute.

Lee, Hae-soo, Yoo, Jae-heung, Ahn, Miso, & Ahn, Seong-won (2025). Artificial Intelligence Industry Trends in Korea and Overseas in 2024(RE-189). Software Policy & Research Institute.

Na, Mi sun & Han, Sang hyo (2016). Establishment of Preservation and Restoration Infrastructure for the New Presidential Archives Building. The Korean Archivist, 34, 22-31.

Nam, Tae-woo, Oh, Ji-young, & Yoo, Bo-hyun (2007). A Study of President Records Management Law. Journal of Korean Society of Archives and Records Management, 7(2), 165-188. https://doi.org/10.14404/JKSARM.2007.7.2.165