

일반논문 (Regular Paper)

방송공학회논문지 제31권 제5호, 2026년 9월 (JBE Vol.31, No.5, September 2026)

<https://doi.org/10.5909/JBE.2026.31.5.1028>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

국회방송 환경에서 AI 음성 인식 모델의 다중 구간 성능 비교 및 폐쇄 자막 자동 생성 연구

박 성 민^{a)†}

A Multi-Segment Performance Comparison of AI Speech Recognition Models for Automatic Closed Caption Generation in National Assembly Broadcasting Environment

Sung Min Park^{a)†}

요 약

현재 국회방송은 폐쇄 자막 서비스를 위해 연간 7억원 이상의 외주 속기 예산을 지출하고 있다. 본 논문에서는 국회방송의 폐쇄 자막 방송과 국회 회의록 자동 생성을 위한 AI 음성 인식 모델의 성능을 비교 연구하였다. 본 연구는 Faster-Whisper, CLOVA Speech, Google STT 세 가지 모델을 국회방송 본회의 유튜브 실제 영상을 활용하여 동일한 5분×5구간(총 1,500초) 조건에서 WER, CER 두 가지 음성 인식 평가 지표를 통해 성능을 비교하였다. 단일 구간 실험 결과 CLOVA Speech의 WER은 4.71%, CER은 1.08%로 세 가지 모델 중 가장 낮은 오류율을 보였으며 다중 구간 검증 실험에서도 CLOVA Speech가 평균 WER 27.44%, CER 6.18%로 가장 우수한 성능을 보였다. 특히 CLOVA Speech 도입 시 연간 비용은 9,799,042원인 반면 기존 외주 속기 비용은 720,000,000원으로 AI 음성 인식 모델 적용 시 약 98.6%의 비용 절감 효과가 기대된다.

Abstract

Currently, the National Assembly Broadcasting System spends more than 700 million Korean Won annually on outsourced stenography for closed caption services. This paper presents a comparative study of AI speech recognition models for the automatic generation of closed captions for National Assembly broadcasts and official parliamentary records. In this study, three models – Faster-Whisper, CLOVA Speech, and Google STT – were evaluated using actual footage from National Assembly plenary sessions on YouTube under identical conditions of five 5-minute segments (total 1,500 seconds), comparing performance through two speech recognition metrics, Word Error Rate (WER) and Character Error Rate (CER). In the single-segment experiment, CLOVA Speech achieved the lowest error rate with a WER of 4.71% and CER of 1.08%. In the multi-segment validation experiment, CLOVA Speech also demonstrated superior performance with an average WER of 27.44% and CER of 6.18%. Notably, the annual cost of implementing CLOVA Speech is 9,799,042 Korean Won, compared to the existing outsourced stenography cost of 720,000,000 Korean Won, suggesting an expected cost reduction of approximately 98.6% upon adoption of the AI speech recognition model.

Keyword : Automatic Speech Recognition, Closed Caption, Character Error Rate, Cost Reduction, National Assembly Television, Word Error Rate, Whisper

Copyright © 2026 Korean Institute of Broadcast and Media Engineers. All rights reserved.

“This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered.”

1. 서론

국회방송은 1995년 구내방송을 시작으로 2004년 5월에 개국하여 2010년 1월부터 종일 방송을 하고 있다. 또한 국회방송은 SNS 및 케이블TV, 위성방송, IPTV, 인터넷, 유튜브 등으로 실시간 및 녹화방송을 하고 있다^[1].

자막 방송은 현대 미디어 소비의 기본 요소가 되고 있으며 방송 TV, 스트리밍 서비스, 온라인 비디오까지 자막 서비스를 하고 있다. 넷플릭스 사용자의 80% 이상이 한 달에 최소 한번 자막 방송을 시청하고 있다. 미국 내 전체 시청 시간의 절반을 자막과 함께 콘텐츠를 시청하고 있다^[2].

최근 유튜브 방송과 넷플릭스 등 OTT 플랫폼의 글로벌 콘텐츠 미디어 환경이 급성장함에 따라 전통적인 TV 시청 습관에서 점차 텍스트를 통해 시청하고 있어 자막 방송은 소통 차원에서 매우 중요한 역할을 하고 있다.

국내 방송 사업자들은 대부분 청각 장애인에게는 수화 서비스와 방송 자막 서비스를 제공하고 있다. 이러한 관점에서 국회방송은 다채널·다매체 시대에 대응하기 위해 AI 기반 음성 인식 자막 서비스 도입을 적극적으로 검토해야 한다. 최근 지상파 방송사들이 본방송 외에 유튜브, OTT, SNS 등 다양한 플랫폼을 통해 콘텐츠를 제공하고 있는 것처럼 국회방송 역시 다채널 환경에서 모든 채널에 실시간 자막을 제공하기 위해서는 AI 음성 인식 기반의 자동화 시스템이 필수적이다.

현재 국회방송의 폐쇄 자막은 외주 속기사에 의해 제공되고 있으며 국회사무처의 공식 회의록은 별도로 작성되고 있다. 동일한 발언에 대해 방송 자막과 국회 공식 회의록이 이중으로 제작됨으로써 인력과 비용 측면에서 비효율성이 발생하고 있다. 현재 국회방송은 대부분 생방송 스케줄에 한정하여 폐쇄 자막 서비스를 제공하고 있다. 다만, 상임위

원회는 생중계만 폐쇄 자막을 제공하고 있다. 향후 청각 장애인의 방송 접근권 보장과 무음 시청 증가 등 방송 시청 행태 변화를 고려할 때 AI 음성 인식 기반 폐쇄 자막 시스템을 모든 위원회 중계로 확대 적용하는 것이 필요하다. 또한 국회방송의 폐쇄 자막 서비스를 보면 24시간 무제한 토론 및 야간 속기를 계속하다 보면 국회 내용에 약간씩 오류가 발생하는 문제점이 발생하고 있다.

최근 자막 방송은 소음이 많은 대중교통 등 다양한 환경에서 시청자를 수용하여 시청 경험을 향상시키고 포용도 및 사용자의 만족도를 향상시키고 있다. 또한 폐쇄 자막은 처음에는 청각 장애인 용도로 개발되었는데 이제는 언어와 문화 장벽을 넘는 콘텐츠에서 점점 더 상호 연결되는 콘텐츠가 요구되면서 폐쇄 자막은 중요한 역할을 하고 있다^[3].

따라서 기존 연구에서는 음성 인식 평가를 위해 대부분 (WER: Word Error Rate) 단어 오류율 측정 지표를 사용하였는데 본 논문에서는 방송 자막의 특성을 고려하여 평가 지표를 WER(단어 오류율)과 CER(문자 오류율) 두 가지 핵심 지표를 사용하여 음성 인식 성능 평가를 수행하였다.

해외 주요 의회에서는 AI 기반 실시간 자막 시스템을 도입하고 있으나 아직 국내 국회사무처와 국회방송의 속기는 여전히 외주 속기사 방식과 수동 수기 방식에 의해 국회 기록물을 보존 및 방송하고 있어 AI 시대에 맞는 음성 인식 기술 시스템 도입이 필요하다고 판단된다. 또한 특히 국회에서 이중 자막 비용을 절감하기 위해서 AI 음성 인식 자막 시스템 도입을 검토할 필요가 있다.

이에 본 연구에서는 국회방송의 실제 영상을 가지고 음성 인식 모델 3가지를 비교하였다. 또한 이미 해외 의회에서는 AI 음성 인식 자막이 이미 적용 중이다. AI 음성 인식을 하게 되면 한 번만 처리하면 되기 때문에 실시간 처리가 가능하고 이중 비용과 이중 인력이 불필요하다. 따라서 국회방송의 폐쇄 자막 서비스 비용 절감과 청각 장애인의 방송 접근성 향상을 위해 해외 의회 방송 사례와 같이 AI 음성 인식 시스템을 단계적으로 도입할 필요가 있다. 구체적으로 국회방송 PGM 신호 기반의 실시간 음성 인식 자막 서비스를 본회의 중계방송을 시작으로 상임위원회 중계방송과 뉴스 등으로 순차적으로 확대 적용함으로써 속기사의 의존도를 낮추고 자막 서비스의 효율성을 높일 수 있다. 다만, AI 음성 인식 시스템 도입 시 다음과 같은 기술적 한계

a) 국회사무처(National Assembly Secretariat)

‡ Corresponding Author : 박성민(Sung Min Park)

E-mail: psm7601@hanmail.net

Tel: +82-2-6788-3284

ORCID: <https://orcid.org/0000-0001-5374-9240>

※ 본 논문에서 제안하는 AI 음성 인식 기술은 그 독자성을 바탕으로 현재 특허 출원 절차가 진행 중이며(특허 출원 10-2026-0123055) 본 논문에서는 이에 대한 실험 데이터와 비용 분석을 중심으로 기술한다.

· Manuscript received July 4, 2026; Revised July 10, 2026; Accepted July 10, 2026.

를 고려해야 한다. 네트워크 중단 시 실시간 자막 서비스가 중단될 수 있다. 둘째, 숫자와 단어가 혼용된 발언의 경우 오인식이 발생할 수 있다. 아울러 복수의 마이크에서 동시 입력되거나 여러 의원이 동시에 발언하는 경우 화자 구분 및 음성 인식 정확도가 저하될 수 있다. 또한, 국회 보안상 외부 클라우드 API 활용에 제약이 따를 수 있다. 따라서 이러한 기술적 한계를 보완하면서 본회의 중계를 시작으로 단계적으로 확대 적용하는 것이 바람직하다.

음성 인식을 이용하여 자막을 자동으로 변환하고 종결어미와 시간을 활용해 문장을 분리하여 자막을 생산함으로써 자막 생성에 드는 시간과 노동력을 감소시킬 수 있도록 하는 시스템을 개발하였다^[4].

현재 국회방송의 방송 자막 서비스와 국회사무처의 공식 속기록은 동일한 발언에 대해 별도의 인력과 비용으로 이중 제작되고 있어 비효율성이 발생하고 있다. AI 음성 인식 시스템 도입 시 단일 처리로 방송 자막과 공식 회의록을 동시에 생성할 수 있어 이중 속기 문제를 해소하고 예산 절감 효과를 극대화할 수 있다. 또한 방송 자막과 공식 회의록의 내용이 동일한 음성 인식 결과로부터 생성됨으로써 두 기록 간의 불일치를 방지하고 국회 정보의 신뢰도와 투명성을 향상시키는데 기여할 수 있을 것으로 판단된다.

본 논문은 다음과 같이 구성한다. II장에서는 음성 인식 모델 현황을 검토한다. III장에서는 해외 의회의 음성 인식 기술 적용 사례를 기술하였다. 그리고 IV장에서는 방송 자막 음성 인식 시스템 설계 구성을 기술하였다. 특히 V장에서는 음성 인식 모델의 분석 결과를 제시하였다. VI장에서는 본 논문의 결론을 제시하고 시사점을 제시하였다.

II. 음성 인식 모델 현황

1. 음성 인식 기술 모델

IEEE Access(2025)에 따르면 음성 인식 모델 중에서 Transformer 모델은 기존 Seq2Seq 모델의 정보 손실의 한계를 극복하고 정확하고 효율적인 음성 인식 시스템을 제공하는 기술이며, 병렬 처리와 동시 처리를 가능하게 하고 실시간 번역과 다국어 인식 및 감정 감지 등 다양한 음성

관련 분야에서 성공적으로 활용되고 있다. Transformer 모델은 효율성을 향상시킨 딥러닝 모델로 기상 예보, 의료 분야, 스마트 시티 등 다양한 분야에서 긍정적인 영향을 미치고 있다. 음성 인식 분야에서는 Transformer 기반 모델이 가장 많은 연구에서 사용되고 있으며 OpenAI의 Whisper가 두 번째로 많이 활용되는 모델로 나타났다. 이는 Whisper가 다국어 지원과 높은 인식 정확도를 바탕으로 전 세계적으로 널리 활용되고 있음을 보여준다^[5]. 따라서 본 논문에서는 글로벌 관점에서 Whisper가 핵심 음성 인식 모델로 평가하여 연구에 적용하려고 한다.

El Bahri et al.(2025)은 Codecarbon, Pyjoule, Powerapi 측정 도구를 활용하여 Python 및 PyTorch 기반으로 Kaggle 데이터셋의 20,000개 오디오 클립을 대상으로 클라우드 기반 음성 인식 시스템의 성능을 평가하였다. 분석 결과 Google STT가 모든 평가 지표에서 우수한 성능을 보였으며 Whisper 대비 WER이 13.4% 낮았고 처리 시간은 51% 빠른 것으로 나타났다^[6].

청각 장애인의 미디어 접근성을 향상시키기 위해 음성 인식 서버 및 스마트 폰과 연계하여 음성을 통해 자동으로 자막을 생성하고 표시하는 음성 자막 자동 변환 시스템을 개발하였다^[7].

1.1 Naver CLOVA Speech 모델

Naver CLOVA와 Google STT 모델의 전사 정확도를 비교하였다. 초등학교 1, 2학년의 일반 아동 20명과 언어학습 장애 아동 20명을 대상으로 80개의 글을 수집하여 분석하였다. 분석 결과 전사 오류율에 있어 Naver CLOVA는 15.17%로 나타났고 Google STT는 30.11%보다 통계적으로 유의하게 낮은 것으로 제시되었다. 그리고 Naver CLOVA가 Google STT보다 정확도가 높았으나 아동 집단 간 차이는 나타나지 않았다^[8].

1.2 Deepgram Nava 모델과 Azure Speech 모델

이 논문에서는 TORGO 구음장에 발화 코퍼스(Corpus)를 활용하여 총 8개의 상용 음성-텍스트 변환 서비스를 평가하였다. 평가 대상은 4개의 전통적인 ASR 시스템(AssemblyAI, Whisper large-v3, Deepgram Nova-3, Nova-3 Medical)과 4개의 MLLM 기반 시스템(GPT-4o, GPT-4o

Mini, Gemini 2.5 Pro, Gemini 2.5 Flash)으로 구성되었다. 분석 결과, 장애의 중증도에 따라 성능이 저하되었다. 경증 구음장애 환경에서는 일반적인 발화 벤치마크에 근접하는 3~5%의 단어 오류율(WER)을 달성한 반면, 중증 구음장애 환경에서는 모든 시스템에서 WER이 49%를 초과하였다^[9].

마이크로소프트(Microsoft)는 애저(Azure) 프로젝트의 일환으로 대화 및 구어 받아쓰기 시나리오에 특화된 음성-텍스트 변환(Speech-to-text) 서비스를 제공하고 있다. 원어민(L1) 대비 비원어민(L2) 화자의 낭독체 음성에서 상용 ASR 서비스의 단어 오류율(WER)이 약 2배 증가하는 것으로 나타났다^[10].

1.3 ElevenLabs Scribe 모델

M. Pietroni(2026) 연구에서는 의료 문서 자동 생성에 사용되는 의사-환자 간 대화에 초점을 맞추어, 폴란드어를 지원하는 여러 최고 수준(SOTA)의 ASR 모델들을 비교하였다. ElevenLabs Scribe 모델은 의학 용어가 밀집되어 있고 소음 및 화자 중첩이 심한 임상 음성 환경에서 Whisper가 우수한 반면, 다국어 모델들은 의학 약어, 라틴어 용어 및 수치 데이터를 처리하는 데 어려움을 겪었다. 테스트된 모델들은 단어 에러율(WER)과 음절 에러율(CER)을 기준으로 평가하였다. 실험 결과, 오픈 소스 모델 중에서는 Whisper 모델이 가장 우수한 성능을 보였다. 반면, ElevenLabs의 Scribe 모델은 일반 벤치마크와 의료 데이터 모두에서 폴란드어에 대해 가장 뛰어난 성능을 발휘함을 보여주었다^[11].

2. 오픈 소스 음성 인식 기술의 모델

오픈 소스 음성 인식 모듈은 Nvidia nemo Canary, Reazon Speech, Open AI Whisper, Wave2vec 2.0, IBM Granite Speech, Distil-Whisper, Moonshine, Faster Whisper 등 최근에는 LLM 기반의 음성 인식 엔진으로 발전하고 있다.

2.1 Reazon Speech 모델

Reazon Speech는 일본 TV 프로그램에서 수집된 15,000 시간 이상이 지속적으로 성장하고 있는 코퍼스이다. Reazon Speech로 학습된 자동 음성 인식(ASR) 모델은

JSUT basic5000 테스트 세트에서 8.23%, Common Voice v8.0 테스트 세트에서 9.93%의 음절 에러율(CER)을 기록하고 있으며 최고 수준(State-of-the-art, SOTA)의 결과를 달성하였다. Whisper large-v2 모델과 동등한 수준이다^[12].

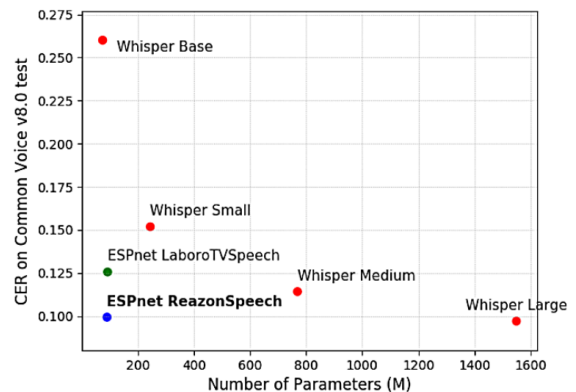


그림 1. 공동 음성에서의 CER 음성 인식 정확도^[12]

Fig. 1. CER Speech Recognition Accuracy in Collaborative Voice Environments^[12]

2.2 Whisper Large V3 Turbo 모델

Alec Radford(2023) 논문에서는 인터넷상의 대규모 오디오 전사 데이터를 단순 예측하기 위해 학습된 음성 처리 시스템의 성능을 조사하였다. 68만 시간에 달하는 다국어 및 다중 작업 지도로 학습된 이 모델은 표준 벤치마크 데이터셋에서 우수한 성능을 보였다. 별도의 데이터셋 특화 미세 조정 없이도 기존의 완전 지도 학습 기반 연구들과 경쟁할 만한 성과를 달성하였다. 나아가 인간의 인지 능력과 비교했을 때, 본 모델은 인간 수준의 정확도와 강건성(Robustness)에 근접하는 결과를 보였다^[13].

Whisper Large V3 Turbo 모델은 디코더를 32층에서 4개 층으로 줄여 파라미터 수를 15.5억에서 8.09억으로 축소하여 동일한 정확도로 2-5배 빠른 전사 속도를 제공한다^[14].

Bain et al.(2023)은 Whisper의 타임스탬프 부정확성과 단어 정보 미제공 문제를 해결하기 위해 WhisperX를 개발하였다. WhisperX는 음성 활동 감지(VAD)와 강제 음소 정렬을 통해 단어 수준의 정확한 시간 기록을 제공하였다^[15].

2.3 Wav2Vec 2.0 모델

Wav2Vec 2.0 모델은 원시 파형의 잠재 표현을 마스킹하

고 양자화된 음성 표현에 대한 대조 학습 과제를 해결하는 자기 지도 학습 기반의 프레임워크이다. 음성 처리 분야에서 레이블이 없는 대규모 데이터로 사전 학습하는 방식으로 높은 잠재력을 입증하였으며 단 10분의 레이블 데이터 만으로도 Librispeech 데이터셋 기준 WER 4.8%와 8.2%를 달성하는 우수한 음성 인식 성능을 보였다^[16].

따라서 본 논문에서는 다양한 글로벌 음성 인식 모델 중 한국어 환경에 최적화된 모델을 선정하기 위해 오픈 소스 기반의 Faster-Whisper, 한국어 특화 상용 서비스인 네이버 CLOVA Speech, 글로벌 표준 음성 인식 서비스인 Google STT를 대상으로 비교 실험을 수행하였다.

III. 각국 의회의 음성 인식 기술 적용

1. 각국 의회의 음성 인식 기술 현황

국제의회연맹(IPU)은 의회 세션의 음성을 텍스트로 자동 변환하는 AI 기반 Hansard 속기록 및 회의록 시스템을 개발하고 있다. 음성 녹음이나 라이브 피드를 AI 시스템에 전송하면 실시간으로 처리하여 의회 토론을 정확하게 문서화할 수 있다. IPU는 AI 보조 음성 인식 기술을 도입하여 의회 토론 및 위원회 회의의 실시간 대본을 자동으로 생성함으로써 의회 절차의 접근성과 투명성을 향상시키고 있다^[17].

캐나다 의회는 AI 기반 음성 인식 시스템을 도입하여 의원 발언을 텍스트로 자동 변환하고 있다. 다만 AI 시스템이 미국식 영어 철자를 선호하거나 특정 고유명사 인식에 어려움을 겪는 한계점이 보고되고 있어 지속적인 개선이 필요한 것으로 나타났다^[18].

Nascimento et al.(2024)는 포르투갈 의회 회의를 대상으로 의회 토론 자동 전사 시스템인 STAAR(Speech To Auto Assembly Republica)을 개발하였다. STAAR은 OpenAI Whisper 모델을 기반으로 설계된 최신 ASR 기술로 다국어 및 다중 화자 환경에서 안정적인 성능을 보였다. 기존 ASR 시스템 대비 오류율이 1.7%에서 11.3% 감소하였으며 WER 6.1%를 기록하였다. 또한 낭독 연설이 즉흥 연설보다 더 높은 인식 정확도를 보이는 것으로 나타났다. STAAR은

포르투갈 의회 본회의에 실제 적용되어 회의록 자동 생성을 통해 공식 의회 기록 작성의 효율성을 크게 향상시켰다^[19].

Kawahara(2012)에 따르면 의회 회의록 후편집 작업 분석 결과 채우기 단어 삭제가 51%, 담화 표지 삭제가 23.5% 차지하는 것으로 나타났다. 일본 의회는 자동 음성 인식 시스템(Automatic Speech Recognition: ASR)을 개발하여 2010년부터 실제 의회에 배치하였으며 약 90%의 문자 인식 정확도를 달성하였다. 이 시스템은 방대한 의회 회의록 자료를 효율적으로 활용하기 위해 개발되었으며 통계적 기계번역(SMT) 방식을 적용하여 즉흥적 발언에 대한 인식 정확도를 향상시켰다^[20].

Proksch et al.(2024)는 정치 연구에서 음성 인식 활용 가능성을 평가하기 위해 두 가지 Google STT API와 YouTube 자동 자막을 비교 분석하였다. 자동 전사 결과와 수동 전사 결과를 비교하여 단어 오류율(WER)을 측정하고 정치 연설 환경에서의 음성 인식 성능을 평가하였다^[21].

IV. 방송 자막 음성 인식 시스템의 설계 구성

1. 전체 시스템 구조도

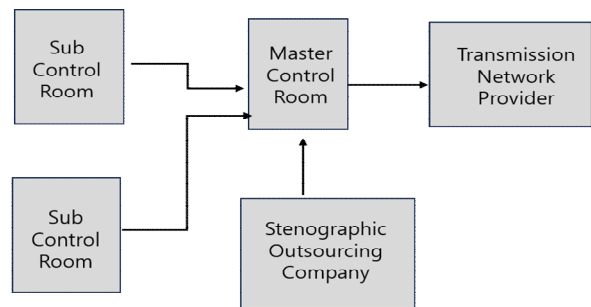


그림 2. 현재 국회방송 자막 방송 흐름도

Fig. 2. Current Closed Caption Broadcasting Workflow of National Assembly Broadcasting

국회방송에 AI 음성 인식 시스템을 적용할 경우 주조정실에서 AI 음성 인식 모듈과 VAD 필터를 거쳐 텍스트 후처리 과정을 통해 방송 자막 생성과 속기록 데이터베이스를 자동 저장이 가능하다. 다만 국회 공식 속기록은 본회의

뿐만 아니라 모든 상임위원회 회의에 적용되기 때문에 현재 제안된 국회방송 주조정실 기반 시스템만으로는 전체 속기록의 자동화를 구현하는 데 한계가 있다. 따라서 전체 속기록의 시스템 자동화를 위해서는 각 상임위원회 회의장의 음향 시스템에 AI 음성 인식 모듈을 별도로 적용하는 것이 더 효율적일 것으로 판단된다.

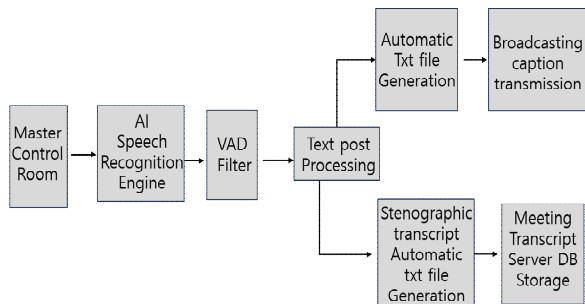


그림 3. 국회방송 AI 음성 인식 모듈의 방송 자막 송출 다이어그램
Fig. 3. Broadcast Caption Output Diagram of AI Speech Recognition Module for National Assembly Broadcasting

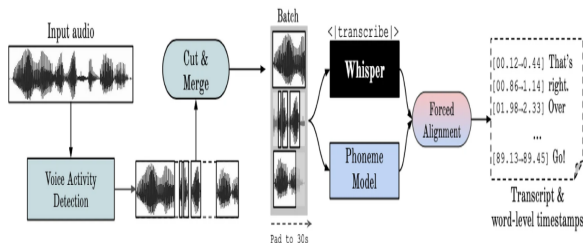


그림 4. WhisperX를 이용한 음성 텍스트 프로세스 다이어그램^[22]
Fig. 4. Speech-to-Text Process Diagram Using WhisperX^[22]

2. 음성 인식 모듈 설계

본 연구에서 적용한 Faster-Whisper 음성 인식 모듈은 VAD 필터를 적용하여 국회방송 장시간 오디오 처리 시 발생하는 환각 현상을 원천 차단하고 타임스탬프 기반의 속기록을 자동 생성한다. VAD 필터는 실제 방송 환경에서 무음 및 소음 구간을 자동으로 제거하여 인식 정확도를 높이며 실시간 파일 저장 기능을 통해 방송 중단 시에도 그 이전까지의 결과물이 보존되는 강점이 있다.

Faster-Whisper는 Transformer 모델 전용 고속 추론 엔진인 CTranslate2를 기반으로 Whisper를 재구현한 것으로 기

존 Whisper 대비 처리 속도가 4배 향상되었다^[23].

표 1은 Whisper와 Google STT의 평균 성능을 비교한 것이다.

표 1. Whisper와 Google STT 평균 성능 지표^[24]

Table 1. Average Performance Metrics of Whisper and Google STT^[24]

Metrics	Google speech to text	Whisper
World error rate(WER)	14.9%	17.2%
Total processing time	122min	238min
Processing time per audio hour	5.4min/h	10.7min/h

표 2는 Whisper와 Faster-Whisper의 처리 속도를 비교한 것이다.

표 2. Whisper와 Faster-Whisper 음성 인식 엔진 비교표

Table 2. Comparison of Whisper and Faster-Whisper Speech Recognition Engines

Item	Whisper	Faster-Whisper
Processing Time	1-2 hours	10-15 minutes
Hallucination	Occurred	Completely Eliminated
Timestamp	Not Provided	Auto-Generated
Memory	Insufficient Risk	Real-time Clearing
Stenography Format	Text Only	Time + Text

표 3은 포르투갈 의회 STAAR 시스템과 본 연구를 비교한 것이다.

표 3. 포르투갈 의회 STAAR과 본 연구 비교

Table 3. Comparison Between Portuguese Parliament STAAR and This Study

Item	STAAR	This Study
Error Rate	1.7%-11.3%	CLOVA 4.71%
Application	Portuguese Parliament	Korean National Assembly and National Assembly Broadcasting
Engin	Single	3 Models Comparison
Language	Portuguese	Korean

3. 음성 인식 품질 측정 지표

문자 오류율(CER: Character Error Rate)은 음성 인식 결

과의 텍스트 정확도를 문자 단위로 측정하는 평가 지표이다. CER은 단어 분리가 명확하지 않은 중국어, 일본어 등의 언어에서 주로 활용되며 한국어와 같이 형태소 단위로 의미가 변화하는 언어의 정확도 평가에도 적합한 지표이다^[25].

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

여기서 S 는 잘못 인식된 문자 수(대체), D 는 삭제된 문자 수, I 는 삽입된 문자 수를 의미하며 N 은 참조 텍스트의 전체 문자 수를 나타낸다.

음성 인식 모델의 성능을 평가하기 위해 가장 널리 사용되는 지표인 단어 오류율(WER: Word Error Rate)은 잘못 전사된 단어의 비율을 측정하는 지표로 다음과 같이 계산된다^[26].

$$WER = \frac{S + D + I}{N} \times 100 \quad (2)$$

여기서 S 는 대체된 단어 수, D 는 삭제된 단어 수, I 는 삽입된 단어 수, N 은 참조 텍스트의 전체 단어 수를 의미한다. WER은 값이 낮을수록 음성 인식 정확도가 높음을 나타낸다.

4. 비교 실험 설계

4.1 실험 데이터

본 논문은 제434회 국회 임시회 제3차 본회의(2026년 4월 10일) 국회방송 유튜브 채널에 공개된 유튜브 영상을 실험 데이터로 사용하였다. 동일한 비교 실험을 위해 본회의 15분, 23분, 31분, 39분, 47분대 구간을 각 5분(300초)씩 구분하여 파일 형식은 MP3, 샘플링 레이트 48,000Hz 스테레오로 추출하였다. Faster-Whisper, CLOVA Speech, Google STT 세 가지 모델에 동일한 구간을 바탕으로 WER, CER 두 가지 평가 지표를 기준으로 비교 분석하였다. 다만, Google STT는 API 제한으로 인해 50초 단위로 분할 후 결과를 병합하는 방식을 적용하였다. 또한 나라장터의 공개 자료를 활용하여 AI 음성 인식 모델 도입 시 국회 예산 비용 절감 효과를 추정하였다.

V. 방송 자막 음성 인식 결과

1. 실험 및 결과 구성

본 절은 VAD 필터의 장시간 처리 성능을 별도 검증한 결과이며 유튜브 전체 구간 중 5개 구간을 비교 실험을 수행하였다.

기존 Whisper 모델에서 환각 현상과 언어 혼입 오류가 유발되었으나 VAD 필터를 적용한 결과 Faster-Whisper 활용 시 오류 없이 깔끔한 결과로 출력되었다.

이는 속기록 자동 생성 시스템 구축 적용 가능성을 높이는 중요한 결과가 될 것으로 판단된다.

2. 음성 인식 성능 비교

Faster-Whisper의 평균 신뢰도는 로그 우도값(Log-likelihood) -0.110으로, 이를 지수 변환 및 백분율로 환산 시 약 89.6%에 해당한다. 이는 CLOVA Speech(95.7%) 및 Google STT(87.0%)의 내부 신뢰도 점수 산출 방식(Confidence Score Metrics)과 수학적 정의가 비슷하다. 다만, 해당 수치들을 정량적으로 단순 비교하는 것은 적절하지 않으며 각 엔진의 상대적 지표로 해석하는 것이 타당할 것으로 판단된다. 표 4는 음성 인식 엔진 모델을 비교하여 제시하였다.

표 4. 음성 인식 엔진 비교표

Table 4. Comparison of Speech Recognition Engines

Category	Faster-Whisper	CLOVA Speech	Google Speech STT
Confidence	avg_logprob -0.110	95.7%	87%
Real-time Processing	Not Available	Available	Available
Technical Terminology	Partial Errors	Recognized	Partial Errors
Speaker Identification	Not Available	Available	Not Available
Proper Nouns	Errors	Available	Available
Hallucination	VAD Filter Applied	None	None

본 논문에서 CLOVA Speech 모델이 국회방송 환경에 최적 모델로 선정한 근거는 5가지와 같다. 첫째, 음성 인식 성능 측면에서 5개 구간 다중 실험 결과 평균 WER

27.44%, CER 6.18%로 세 모델 중 가장 낮은 오류율로 나타났다. 둘째, 기술적 적합성 측면에서 실시간 스트리밍과 화자 인식 기능을 지원하고 있어 국회 본회의 생중계 환경에 즉시 적용이 가능하다고 판단하였다. 셋째, 전문 용어 처리 측면에서 국회 법률 용어와 의원 이름 및 예산 수치 등 국회 특수의 용어를 정확하게 음성 인식하였다. 넷째, 기존 외주 속기 예산 대비 비용 측면에서 약 98.6%의 예산 절감 효과가 기대된다. 다섯째, 구두점 자동 생성과 타임스탬프 기능을 통해 속기록 자동 생성이 가능하였으며 국내 방송사 적용 사례가 있어 안정성을 검증하였다.

표 5. 음성 인식의 최적 모델 선정 기준

Table 5. Criteria for Optimal Speech Recognition Model Selection

Selection Criteria	Faster-Whisper	CLOVA Speech	Google STT
WER (5-segment avg.)	32.06%	27.44%	49.70%
CER (5-segment avg.)	10.46%	6.18%	18.57%
Technical Terminology	Partial Errors	Recognized	Partial Errors
Real-time Streaming	Not Available	Available	Available
Speaker Identification	Not Available	Available	Not Available
Auto Punctuation	Available	Available	Not Available
Cost Reduction	Free	98.6%	Paid
Broadcasting Application	Not Available	Available	Not Available

3. 음성 인식 모델의 오류 분석

표 6은 각 음성 인식 모델의 오류 유형별 인식 결과를 비교하여 제시하였다.

표 6. 각 음성 인식 모델의 오류 유형별 비교표

Table 6. Comparison of Error Types Across Speech Recognition Models

Model	Error Category	Reference Text	Error Type	Error Type
ModelError CategoryReference TextRecognition ResultError TypeFaster-Whisper	Technical Terminology	유류세 인하는	유류세인하는	Substitution Error
Faster-Whisper	Proper Noun	우원식 의장님	우원시 교장님	Substitution Error
Google STT	Technical Terminology	유류세 인하는	유류세이나는	Omission Error
CLOVA Speech	None			

** Faster-Whisper: avg_logprob -0.110

환산값: $\exp(-0.110) \times 100 = 89.6\%$ 단, 측정 방식이 달라 직접 비교 불가함.

3.1 음성 인식 모델별 성능 차이 원인 분석

네이버의 CLOVA Speech는 대규모 한국어 음성 데이터를 바탕으로 학습된 한국어 특화 모델로 국회 본회의에서 빈번하게 등장하는 법률 용어, 의원 이름, 예산 수치 등 대부분의 전문 용어 처리에 강점이 있다. 특히 형태소 분석 기반의 언어 모델이 적용되어 한국어 참가어 특성인 조사 결합과 어미변화를 정확하게 처리할 수 있었다. 이로 인해 WER 4.71%, CER 1.08%의 최저 오류율을 보였다.

Faster-Whisper는 OpenAI Whisper를 CTranslate2 엔진으로 재구축된 오픈 소스 기반의 68만 시간의 다국어 데이터를 처리하고 있다. 범용 다국어 모델의 특성상 한국어 전문 용어 처리에서 CLOVA Speech 대비 상대적으로 낮은 결과를 달성하였으나 VAD 필터 적용으로 환각 현상이 완전히 제거되었고 타임스탬프가 자동 생성되는 강점을 보였다. WER 7.06%, CER 1.43%를 달성하였으며 예산 없이 로컬 환경에서 실용적 가치가 보였다.

글로벌 표준 범용 음성 인식 서비스는 Google STT로 다국어 지원에 최적화되어 있다는 평가는 받으나 한국어 전문 용어 처리에서 상대적으로 낮은 성능을 달성하였다. 특히 구두점이 자동 생성되지 않아 속기록으로 활용하기 위해서는 추가 후편집 작업이 필요로 판단되며, WER 13.33%, CER 5.44%를 기록하였다. 이는 국회방송과 같이 전문 용어가 많은 특수 환경에서 범용 모델의 한계를 보여주는 분석 결과이다.

CLOVA Speech가 국회방송 환경에서 가장 우수한 성능을 보인 이유는 다음과 같은 관점에서 추정할 수 있다. 첫째, 음성 인식 학습 데이터 측면에서 CLOVA Speech는 한국어 뉴스, 방송 등 대규모 한국어 음성 데이터를 바탕으로 학습되어 국회방송 환경에 최적화되어 있다고 판단된다. 둘째, 전문 용어 측면에서 법률 용어, 의원 이름 등 국회 특수 용어 인식에 강점을 보였다. 셋째, 음성 인식의 화자 특성 측면에서 다수의 화자가 교체되는 국회 본회의 환경에서도 안정적인 인식률을 보였다. 넷째, 모델 구조 측면에서 한국어 형태소 기반 언어 모델이 사용되어 교착어 특성인 조사 결합과 어미변화를 정확한 처리를 보였다. 반면 Faster-Whisper와 Google STT는 범용 다국어 모델로 한국어 전문 용어 처리하는데 한계를 보여주었다.

4. 음성 인식의 처리 시간 비교

CLOVA Speech를 국회방송 본회의 유튜브 1분 클립에 적용한 음성 인식 출력 결과는 그림 5에 제시하였다.

그만큼 4조 2천억 원이 시중에 나오면 통화량이 늘어나고 결국은 물가를 자극합니다. 물가를 올리게 된다는 말씀입니다. 하지만 유류세 인하는 아닙니다. 800원 정도 되는 유류세의 탄력세율을 만약 50%까지 낮춘다면 기름값을 300원 넘게 낮출 수 있습니다. 현재는 30%까지 낮출 수 있는데, 정부는 그나마 휘발유는 15%밖에 낮추고 있지 않습니다. 줄어드는 세수도 최고 가격 제도로 정산해 주는 것은 4조 원 정도입니다. 다시 말씀드리면 굳이 돈을 뿌리며 시장을 왜곡하면서 기름값을 낮추는 것보다 세금을 덜 걷어서 물가를 자극하지 않고 기름값을 낮추는 것이 상식적이라는 이야기입니다. 정부는 유류세의 추가 인화와 탄력세율 50% 조정을 적극적으로 검토해야 합니다.

그림 5. CLOVA Speech 적용 출력 결과

Fig. 5. Output Results of CLOVA Speech Application

본 실험은 국회방송 유튜브에서 본회의 중계 영상의 50분대 구간을 30초 클립을 Google Speech STT에 적용하여 음성 인식 성능을 평가하였다. 실험 결과 86.8% 결과를 도출하였다. 음성 인식된 텍스트 결과는 그림 6과 같다.



그림 6. Google Cloud Speech 적용 출력 결과

Fig. 6. Output Results of Google Cloud Speech Application

Google Speech STT의 분석 결과를 ‘유류세이나는’으로 음성 인식이 잘못 인식하는 오류가 발생한다. 이는 한국어 전문 용어 처리에 있어서 한국어 특화 모델 대비 낮은 성능을 보이는 것으로 나타났다. 특히 구두점이 자동 생성되지 않아 속기록으로 활용하는데 추가 후편집 작업이 필요할 것으로 보인다.

Faster-Whisper는 국회방송 유튜브 본회의 영상 1시간 영상을 이용하여 VAD 필터를 적용하였다. 적용 결과 환각 현상 없이 깔끔한 결과를 보였다, 그리고 타임스탬프가 자동 생성되었으며 그림 7과 같다.

[51:40] 존경하는 국민 여러분, 우월적 의정 능력과
[51:56] 선배 동료 의원 여러분, 조국혁신당 황윤하 의원입니다.
[52:01] 20년 전 노무현 대통령께서는 수도권 집중 억제와
[52:08] 수도권 발전을 견제하셨다는 원대한 결단으로
[52:12] 수도권 광역권 전방위라는 장벽에 부딪히
[52:14] 비록 그 꿈은 잠시 멈춰섰지만
[52:17] 국가균형발전이라는 시대적 소명은
[52:20] 그 한순간도 퇴색된 적이 없습니다.
[52:23] 단, 한순간도 내 뜻을 이어받은 이재명 정부는
[52:26] 이재명 수도 배종환 원수를
[52:29] 균형성장 분야 제1호 국정 과제이자
[52:31] 123대 국정 과제 중 하나로 확립했습니다.
[52:39] 이는 이름뿐인 행정도시가 아닌
[52:41] 실질적인 행정수도 시대를 열어서
[52:43] 수도권 일극 체제를 깨고
[52:46] 대한민국 전체를 살리겠다는 강력한 의지의 표명입니다.
[52:50] 구청장이 국토의 11% 불과한
[52:54] 수도권에 밀집해서 초저후생과 지방소멸이라는
[52:57] 국가적 재앙을 초래하고 있습니다.
[53:00] 세종특별자치시 행정수도로 삼아서
[53:04] 국토의 다극 체제를 만드는 것만이
[53:06] 지속가능한 대한민국을 만드는 유일한
[53:11] 출구전략입니다.

그림 7. Faster-Whisper VAD 필터 적용 출력 결과

Fig. 7. Output Results of Faster-Whisper with VAD Filter Application

Faster-Whisper VAD 필터를 적용하여 국회방송 제434회 본회의 영상 57분 14초를 처리한 결과 처리 시간 76.90분이 소요되었으며 평균 신뢰도 -0.110을 달성하였다. VAD 필터 적용으로 기존 Whisper에서 발생하던 환각 현상이 완전히 제거되었으며 타임스탬프가 자동 생성되어 속기록 형식으로 출력되었다. 다만 ‘유류세인화는’, ‘우원시 교장님’ 등 전문 용어 및 고유명사 오인식 오류가 일부 발생하였다.

5. 음성 인식 서비스의 비용 절감 효과 분석

CLOVA Speech API 비용 산정 근거는 본회의 100일 기준이며 하루 평균 4시간 기준으로 장문 인식 요금은 5원/15초이며 4시간을 하면 960구간에 5원을 환산하면 day당 4,800원(VAT별도)이다^[28]. 반면 나라장터 공개 자료에 따르면 국회방송 폐쇄 자막 방송의 속기 시간은 2025년 401,709분이고 2026년 445,411분이다. 폐쇄 자막 방송 각 비용은 각각 7억 9천 300만원, 7억 2,000만원의 예산이 지출될 예정이다^[29]. 예산을 대입해서 계산하면 98.6% 절감 효과가 기대된다.

표 7. 음성 인식 모델 비용과 폐쇄 자막 방송 예산 비교표

Table 7. Cost Comparison Between Speech Recognition Models and Closed Caption Broadcasting Budget

Category	2025 Annual Cost	2026 Annual Cost	Cost per Minute
CLOVA Cost	8,837,598KRW	9,799,042KRW	20KRW
Closed Caption Broadcasting Cost	793,000,000KRW	720,000,000KRW	1,974KRW / 1,617KRW
Budget Savings	784,162,402KRW	710,200,958KRW	

본 논문에서 인용한 CLOVA Speech API 관련 정보는 네이버 클라우드 공식 문서를 기반으로 하였으며, 서비스 정책 변경에 따라 수치가 달라질 수 있음을 밝힌다.

국회 공식 속기록(2026년 4월 10일 제434회 본회의)을 참조 텍스트로 사용하여 WER(단어 오류율)과 CER(문자 오류율)을 측정한 결과는 표 8과 같다. WER은 낮을수록 우수한 성능을 나타내며, CLOVA Speech가 4.71%로 가장 우수하였고, Faster-Whisper 7.06%, Google STT 13.33% 순으로 나타났다. CER은 CLOVA Speech 1.08%, Faster-Whisper 1.43%, Google STT 5.44%로 CLOVA Speech가 가장 낮은 오류율로 나타났다. 추정 결과 CLOVA Speech는 한국어 특화 학습 데이터를 바탕으로 국회 회의 음성 인식 및 속기록 자동 생성에 가장 적합한 모델임을 증명하였다. 다만 본 연구는 각 모델별 평가 구간의 길이가 서로 달라 직접적인 통계 비교에 한계가 있음을 밝힌다. 표 8은 음성 인식 모델의 평가 지표별 분석 결과를 제시하였다.

표 8. 음성 인식 모델의 평가 지표별 분석 결과

Table 8. Analysis Results by Evaluation Metrics of Speech Recognition Models

Item	Faster-Whisper	CLOVA Speech	Google Speech STT
WER	7.06%	4.71%	13.33%
CER	1.43%	1.08%	5.44%

표 9는 다중 구간 검증 실험(5분×5구간) 결과 Faster-Whisper WER은 32.06%이고, CER은 10.46%로 나타났다.

표 9. 구간별 음성 인식 에러율 분석 결과

Table 9. Analysis results of speech recognition error rates by segment

Segment	WER	CER	Playback Time
Segment 1 (15min)	31.77%	5.04%	300 sec
Segment 2 (23min)	29.01%	12.64%	300 sec
Segment 3 (31min)	47.27%	16.22%	300 sec
Segment 4 (39min)	27.33%	4.88%	300 sec
Segment 5 (47min)	24.94%	13.50%	300 sec
Overall Mean	32.06%	10.46%	1500 sec

표 10은 유튜브 영상을 5분씩 5구간을 분석한 결과로 CLOVA Speech는 1위로 27.44%, CER 6.18%이다. 그리고

표 10. 구간별 음성 인식 3개 모델 에러율 분석 결과

Table 10. Analysis Results of Speech Recognition Error Rates by Segment Across Three Models

Model Name	Average WER	Average CER	Segment
Faster-Whisper	32.06%	10.46%	5 Minutes×5 Intervals
CLOVA Speech	27.44%	6.18%	5 Minutes×5 Intervals
Google STT	49.70%	18.57%	5 Minutes×5 Intervals

** 동일한 5분×5구간(총 1,500초)으로 3개 모델을 동일 조건에서 비교하였음.
Google STT는 50초 단위 분할 후 결과를 병합하는 방식으로 처리하였음.

Faster-Whisper 2위로 WER 32.06%, CER 10.46%이다. 마지막으로 Google STT 3위로 WER 49.70%, CER 18.57%의 분석 결과로 확인되었다. 따라서 다중 구간 검증 실험(5분×5구간) 결과 CLOVA Speech가 평균 WER 27.44%, CER 6.18%로 가장 우수한 성능을 보였다. 이는 Faster-Whisper는 WER 32.06%, CER 10.46%, Google STT는 WER 49.70%, CER 18.57%로 확인되었다. 따라서 단일 구간 실험 결과와 동일하게 CLOVA Speech가 국회방송 환경에 가장 적합한 모델임을 재확인하였다.

Ⅵ. 결론 및 시사점

본 연구는 국회방송 유튜브 본회의 영상을 활용하여 CLOVA Speech, Google STT, Faster-Whisper 세 가지 AI 음성 인식 모델의 성능을 비교 분석하였다. WER, CER 두 가지 음성 인식 평가 지표와 처리 시간을 반영하여 국회방송 환경에 최적화된 모델을 선정하였다.

분석 결과 CLOVA Speech가 WER 4.71%, CER 1.08%로 세 가지 모델 중 가장 낮은 오류율을 달성하였으며 Faster-Whisper는 WER 7.06%, CER 1.43%로 확인되었다. Google STT는 WER 13.33%, CER 5.44%로 상대적으로 낮은 성능을 보였다. 비용 분석 결과 기존 외주 속기 비용 720,000,000원 대비 CLOVA Speech 도입 시 연간 9,799,042원으로 약 98.6%의 비용 절감 효과가 기대된다. 네이버 CLOVA Speech는 한국어 인식률이 가장 우수하고 실시간 스트리밍을 지원하며 방송사 적용 사례가 있어 국회방송 환경에 가장 적합한 모델로 선정되었다. Google

STT는 글로벌 표준으로 다국어 지원과 실시간 처리가 가능하나 한국어 전문 용어 인식에서 상대적으로 낮은 성능을 보이는 것으로 나타났다. 따라서 국회방송 PGM 오디오에 CLOVA Speech 스트리밍 API를 적용하면 본회의 및 위원회 자막 자동 송출과 속기록 동시 저장이 가능할 것으로 판단된다.

본 논문의 학술적 기여는 다음과 같다. 국내 공공방송 특수 환경에서 상용 모델(CLOVA Speech, Google STT)과 오픈 소스 모델(Faster-Whisper)을 동일한 조건으로 비교 분석하였다. 기존 연구들이 일반 음성 환경에서 단일 모델을 평가한 것이다. 반면 본 연구는 국회방송 방송 환경에서 3개 모델을 체계적으로 비교하였다. 또한, 조달청 나라장터 자료를 사용하여 AI 음성 인식 도입 시 연간 약 98.6%의 외주 속기 비용 절감 가능성을 정량적으로 추정하였다. 이는 단순한 성능 비교를 넘어 실무 적용 가능성과 경제적 효과를 함께 제시한 연구로서 의의가 있다. 특히, 방송 자막 서비스와 국회 공식 회의록의 이중 속기 문제를 AI 음성 인식을 통해 해결하는 통합 시스템을 제안하였다. 이는 청각 장애인의 방송 접근성 향상과 국회 정보의 투명성 제고에 기여할 수 있다.

본 연구의 한계점은 다음과 같다. 첫째, 국회방송 PGM 신호와 실시간 연결 실험을 수행하지 못하였다. 둘째, 공식 회의록 전체 구간을 활용한 WER 정량 측정에 한계가 있었다. 셋째, 실험 데이터가 제한되어 다양한 방송 환경에서 검증이 부족하였다. 넷째, Faster-Whisper의 경우 장시간 처리 시 환각 현상이 부분적으로 발생하였다.

참 고 문 헌 (References)

- [1] Sung min Park, "A Ripple Effect Analysis of Radio Introduction to National Assembly TV," The Korean Institute of Broadcast and Media Engineers, Vol. 26, No. 4, Sep. 2021.
doi: <https://doi.org/10.5909/JBE.2021.26.5.575>
- [2] Grant Scullion, Michael Crabb, and Gemma Webster, "Why Do People Turn on Subtitles? A Systematic Review of Audiences and Motivations for Subtitles Activation and Enjoyment," In Proceedings of the 2026 ACM International Conference on Interactive Media Experiences (IMX '26). Association for Computing Machinery, New York, NY, USA, 279 - 297, 2026.
doi: <https://doi.org/10.1145/3788851.3805012>
- [3] <https://www.eliteasia.co/the-importance-of-subtitles-and-closed-captions-for-global-marketing/> (Accessed 5. June. 2026.)
- [4] W.-S. Son and E.-K. Kim, "Subtitle Automatic Generation System using Speech to Text," The Journal of the Korea institute of electronic communication sciences, vol. 16, no. 1, pp. 81 - 88, Feb. 2021.
doi: <https://doi.org/10.13067/JKIECS.2021.16.1.81>
- [5] Y. O. Sharrah et al., "Advancements in Speech Recognition: A Systematic Review of Deep Learning Transformer Models, Trends, Innovations, and Future Directions," IEEE Access, vol. 13, pp. 46925-46940, 2025.
doi: <https://doi.org/10.1109/ACCESS.2025.3550855>
- [6] El Bahri, J., et al., "Comparative Analysis of Energy Consumption and Carbon Footprint in Automatic Speech Recognition Systems: A Case Study Comparing Whisper and Google Speech-to-Text," Computer Sciences & Mathematics Forum, vol. 1, no. 1, p. 6, 2025.
doi: <https://doi.org/10.3390/cmsf2025010006>
- [7] Choi Mi-Ae, et al., "Development and Enhancement of Automatic Caption Generation System based on Speech-to-Text for the Hearing Impaired," Proceedings of the 2020 Korea Institute of Broadcast and Media Engineers Summer Conference, pp. 465-468, 2020.
- [8] Yang Hui Jae, et al., "Comparison of Automatic Speech Recognition System for School-aged Children's Narratives: Naver Clova Speech and Google Speech-to-Text," Communication Sciences and Disorders, 28(1), 30-38, 2023.
doi: <https://doi.org/10.12963/csd.23952>
- [9] El Bahri, J.; Kouissi, M.; Achkari Begdouri, M., "Comparative Analysis of Energy Consumption and Carbon Footprint in Automatic Speech Recognition Systems: A Case Study Comparing Whisper and Google Speech-to-Text," Comput. Sci. Math. Forum 2025, Vol. 10, no. 6.
doi: <https://doi.org/10.3390/cmsf2025010006>
- [10] Ronald Cumbal1, Birger Moell1, Jos'e Lopes2, Olov Engwall1, "You don't understand me!" Comparing ASR results for L1 and L2 speakers of Swedish," 2405, 13379, arXiv, 2024. <https://arxiv.org/abs/2405.13379>
- [11] M. Pietroni et al., "Quality of Automatic Speech Recognition - Polish Language case study - from Wav2Vec to Scribe ElevenLabs," arXiv preprint arXiv:2603.02246, Feb. 2026. Available: <https://arxiv.org/abs/2603.02246>
- [12] Yin, Y., Mori, D., & Fujimoto, S, "ReazonSpeech: A Free and Massive Corpus for Japanese ASR," Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 2023.
doi: <https://doi.org/10.21437/Interspeech.2023-1406>
- [13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," In International Conference on Machine Learning, PMLR, pp. 28490-28508, 2023.
doi: <https://doi.org/10.5555/3618408.3619590>
- [14] <https://whispernotes.app/ko/blog/introducing-whisper-large-v3-turbo/> (Accessed 6. June. 2026.)
- [15] Bain, M., Huh, J., Han, T., & Zisserman, A., "WhisperX: Time-

- Accurate Speech Transcription of Long-Form Audio,” Proceedings of the Interspeech 2023, 4489 – 4493, 2023.
doi: <https://doi.org/10.21437/Interspeech.2023-78>
- [16] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449 – 1260, 2020.
doi: <https://doi.org/10.5555/3495724.3496769>
- [17] <https://www.ipu.org/ai-use-cases/ai-assisted-speech-transcription-parliamentary-proceedings>(Accessed 4. June. 2026.)
- [18] <https://www.ipu.org/ai-use-cases/automatic-audio-transcription-plenary-sessions-and-committee-meetings>(Accessed 4. June. 2026.)
- [19] Nascimento, P., Ferreira, J. C., & Batista, F., “Automatic transcription system for parliamentary debates in the context of assembly of the republic of Portugal,” International Journal of Speech Technology, vol. 27, no.3, pp. 613 – 635, 2024.
doi: <https://doi.org/10.1007/s10772-024-10126-4>
- [20] Kawahara, T., “Transcription System Using Automatic Speech Recognition for the Japanese Parliament (Diet),” Proceedings of the AAAI Conference on Artificial Intelligence, 26(2), 2224 – 2228, 2012.
doi: <https://doi.org/10.1609/aaai.v26i2.18962>
- [21] Proksch, S.-O. et al., “Testing the Validity of ASR,” Political Analysis,” 2019, doi.org/10.1017/pan.2018.62
doi: <https://doi.org/10.1017/pan.2018.62>
- [22] Bain, M., Huh, J., Han, T., & Zisserman, A., “WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Proceedings of the Interspeech 2023,” pp. 4489 – 4493, 2023.
doi: <https://doi.org/10.48550/arXiv.2303.00747>
- [23] D.Macháček et al., “Turning Whisper into Real-Time Transcription,” arXiv:2307.14743, 2023.
doi: <https://doi.org/10.18653/v1/2023.ijcnlp-demo.3>
- [24] El Bahri J et al., “Comparative Analysis of Energy Consumption,” MDPI, 2025.
doi: <https://doi.org/10.3390/cmsf2025010006>
- [25] <https://www.llamaindex.ai/glossary/what-is-character-error-rate> (Accessed 8. June. 2026.)
- [26] H. de Vos and S. Verberne, “Political corpus creation,” arXiv: 2304.08137, 2023.
doi: <https://doi.org/10.48550/arXiv.2304.08137>
- [27] Varona, A. et al., “A Bilingual Basque-Spanish Dataset,” Applied Sciences, 2024.
doi: <https://doi.org/10.3390/app14051951>
- [28] <https://www.ncloud.com/product/aiService/clovaSpeech#pricing> (Accessed 8. June. 2026.)
- [29] <https://www.g2b.go.kr> (Accessed 8. June. 2026.)
- [30] https://record.assembly.go.kr/assembly/mnts/total/22.do;jsessionid=pa-ncAcmDnDQyYeJpGKtL7ZJIEJbusd9h9VQDlec9lHKFYsqzblJ-1844123710?class_id_sch=1(Accessed 8. June. 2026.)

저 자 소 개



박 성 민

- 2000년 : 강원대학교 전기공학과 졸업
- 2006년 : 서울과학기술대학교 매체공학과 대학원 졸업
- 2015년 : 한국공학대학교 경제학 박사 졸업
- 2000년 ~ 2003년 : 한국교육방송공사 (EBS) 역임
- 2003년 ~ 현재 : 국회사무처 방송국 주무관 재직
- ORCID : <https://orcid.org/0000-0001-5374-9240>
- 주관심분야 : 미디어 인공 지능 및 딥러닝 기반 영상 처리, AI 융합 분야, 방송 정책