# LLM을 활용한 STT 문맥 기반 할루시네이션 감소 연구

이남경, 김수현, 박현석, 김지현, 김희원, 이건희*
에이치디씨랩스

{nk_lee, soohyun.kim, hyunskki, jh991219, ive2go, Gunhee_Lee}@hdc-labs.com

# Reducing Hallucination in Whisper-Based Speech Recognition Using Context-Aware LLM Correction Framework

Namkyeong LEE, Soohyun Kim, Hyunseok Park, Jihyeon Kim, Heewon Kim, Gunhee Lee*
HDC LABS.

요 약

본 논문은 Whisper large-v3-turbo 기반 음성 인식(STT) 시스템에서 발생하는 할루시네이션(hallucination)을 효과적으로 완화하기 위해, 대형 언어 모델(LLM)을 활용한 문맥 기반 탐지 프레임워크를 제안한다. 제안하는 방법은 음성 인식 결과의 앞뒤 발화를 활용해 문맥 기반으로 할루시네이션을 탐지하고, 자주 탐지되는 단어나 문장을 기록하여 이후 실시간 교정이 가능하도록 설계된다. 실제 산업 현장의 회의록 데이터에 적용한 결과, 본 프레임워크는 전문가가 오류로 판단한 발화 중 평균 38.87%를 LLM이 탐지하였으며, LLM이 탐지한 발화 중 67.91%는 실제 오류와 일치하였다. 이는 후속 회의에서 반복적인 할루시네이션 발생 가능성을 줄이고, STT 기반 기록의 신뢰도를 향상시키는데 기여할 수 있음을 시사한다.

## Ⅰ. 서 론

최근 음성-텍스트 변환(STT: Speech-To-Text) 기술은 다양한 산업 현장에서 활용되며, 사람의 업무 효율성을 크게 향상시키고 있다. 하지만 STT 모델은 학습 데이터에 의존해 추론을 수행하기 때문에, 실제 발화와 무관하거나 존재하지 않는 정보를 생성하는 할루시네이션(hallucination) 현상이 자주 발생한다. 이러한 오류는 회의, 상담, 법률 등 높은 정확성이 요구되는 환경에서 심각한 문제를 초래할 수 있다.

기존 연구에서는 STT 모델에 도메인 특화 데이터를 추가로 학습시켜 파인튜닝을 진행하거나[1][2], 확률 기반 후처리 기법을 활용하여 오류를 줄이고자 하였다. 또한 최근에는 질문을 단위별로 나눈 뒤, LLM(Large Language Model)을 통해 검증함으로써 할루시네이션을 교정하려는 시도도 이루어지고 있다.[3][4] 그러나 실제 환경에서의 복잡한 발화 구조에서는 단순한 단어 빈도나 확률 값만으로 오류 여부를 정확히 판단하기 어렵고, 동일한 문장이라도 대화 맥락에 따라 의미가 달라지는 경우가 많다. 특히, 일부 발언은 이전/이후 발화와의 연계를 통해서만 올바르게 해석될 수 있다. 이로 인해 문맥을 고려하지 않은 교정은 불필요한 수정이나 의미 왜곡을 초래할 수 있다.

이에 본 연구에서는 LLM의 문맥 파악 능력을 활용하여, Whisper 기반 STT 시스템에서 발생하는 할루시네이션을 문맥 기반으로 탐지하고 교정할 수 있는 프레임워크를 제안한다. 제안하는 방법은 앞뒤 문맥을 바탕으로 할루시네이션 여부를 판단하며, 반복적으로 탐지된 오류 표현을 기록하여 이후 실시간으로 교정이 가능하도록 설계되었다. 실제 산업 현장에서 수집된 실시간 회의록 데이터를 기반으로 실험한 결과, 제안한 프레임워크는 유사 환경의 후속 회의에서 할루시네이션 발생률을 유의미하게 감소시켰으며, 이를 통해 실제 환경에서도 신뢰도 높은 기록을 가능하게 함을 입증하였다.

또한, 본 연구에서는 특정 주제나 도메인에만 최적화된 프롬프트가 아닌, 다양한 회의 주제에 적용 가능한 일반화된 프롬프트 기법을 도입함으로써, 다양한 상황에 대한 범용성을 확보하였다.

본 연구의 주요 기여는 다음과 같다.

1. Whisper 기반 STT 결과의 할루시네이션을 문맥 기반으로 탐지하고, 실시간으로 후속 처리를 할 수 있는 프레임워크를 제안한다.
2. 실제 산업 현장의 회의 데이터를 활용하여 제안 프레임워크가 실환경에서도 작동함을 입증한다.
3. 다양한 회의 주제 및 맥락에 대응 가능한 일반화된 프롬프트 기법을 적용하여 프레임워크의 범용성과 실효성을 강화한다.

## Ⅱ. 본론

본 연구에서는 Whisper 기반 STT 모델에서 발생하는 할루시네이션을 감소시키기 위한 프레임워크를 제안한다.

1. 데이터셋 구축 및 전처리

본 연구는 실제 기업 회의에서 수집된 실시간 회의 음성 데이터를 기반으로 실험을 수행하였다. 데이터셋은 회의 원본 음성 파일, Whisper 모델을 통해 자동 생성된 STT 결과, 그리고 전문가가 교정한 정답 텍스트 세 가지로 구성된다. 원본 음성은 전처리 과정을 거쳐 유의미한 발화 구간만을 블록 단위로 추출하였다. 무음이 일정 시간 지속되거나 음성 구간의 길이가 특정 기준을 초과한 경우에만 해당 블록을 Whisper 모델에 입력하여, 실제 발화를 중심으로 한 STT 결과를 생성하였다. 이렇게 처리된 텍스트는 전문가에 의해 교정된 정답 텍스트와 비교되며, 음성과 불일치하거나 문맥상 부적절한 단어나 문장을 할루시네이션으로 라벨링하였다.

2. 프레임워크 구성 요소

제안하는 프레임워크는 Whisper 의 STT 결과를 입력으로 받아, LLM 을 활용한 문맥 기반 분석과 반복 오류 관리를 통해 최종적으로 신뢰도 높은 텍스트를 산출한다. 주요 구성 요소는 다음과 같다.

(1) 문맥 기반 할루시네이션 탐지 모듈

Whisper 모델로부터 생성된 문장을 중심으로, 앞뒤 발화와의 문맥적 연관성을 LLM 을 통해 평가한다. 이를 통해 해당 문장이 실제 발화 흐름과 의미적으로 일치하는지를 판단하고, 문맥과 무관한 표현을 할루시네이션으로 탐지한다. 특히 Whisper 모델은 방송·자막·영상 데이터 등 다양한 비회화적 학습데이터를 포함하고 있어, 실제 회의 맥락과 무관한 표현(예: “다음 영상에서 만나요”)이 출력되는 경우가 있다. 또한 “네”, “음” 등 의미 없는 단어가 반복적으로 생성되는 현상도 자주 발생한다. 본 모듈은 이러한 표현을 문맥 불일치로 식별하여 제거함으로써, 발화 기록의 정확도와 신뢰도를 향상시킨다.

(2) 오류 패턴 기록 및 실시간 교정 모듈

실험 과정에서 반복적으로 등장하는 할루시네이션 유형(예: 특정 단어의 지속적 오인식 등)은 별도의 오류 패턴 데이터베이스에 저장된다. 이 데이터베이스는 이후 STT 결과에 대한 실시간 탐지 및 교정에 활용되며, 유사한 발화에서 반복 오류를 자동으로 수정하는 데 기여한다.

3. 성능 평가

제안한 프레임워크를 실제 기업 회의 데이터에 적용하여 성능을 검증하였다. 다양한 주제의 회의에서 Whisper 의 STT 결과, 전문가 교정본, 그리고 프레임워크 적용 결과로 비교하여 할루시네이션 발생률의 감소폭을 측정하였다.

HDC 랩스에서 수집된 실무 회의 데이터를 활용한 실험 결과, 1 시간 회의 기준 Whisper 모델의 할루시네이션 발생률은 평균 15.85%로 측정되었으며, 제안한 프레임워크를 적용한 경우 10.38%로 약 34.5% 감소하였다.

또한, 오류 패턴 데이터베이스에 등록된 할루시네이션 표현이 후속 회의에서 재등장 한 경우, 프레임워크는 이를 자동 탐지 및 교정하여 할루시네이션 발생률을 추가적으로 2% 이상 감소시킬 수 있었다.

또한, 제안 프레임워크의 핵심 구성 요소인 LLM 기반 문맥 분석의 실질적인 탐지 성능을 평가하기 위해 실제 회의 데이터를 대상으로 사람이 판단한 오류와 LLM 이 판단한 오류간의 일치율을 비교하였다. 아래 표는 회의별 LLM 탐지 결과를 정리한 것으로, 탐지율(Recall), 정확도(Precision), 그리고 사람이 판단한 오류 중 LLM 이 탐지하지 못한 비율(미탐지율)을 함께 제시한다.

**표1. LLM 기반 할루시네이션 탐지 성능 요약**

| | 탐지율(%) | 정확도(%) | 미탐지율(%) |
|---|---|---|---|
| Gemma 2-9B-INT4 | 34.26 | 64.38 | 13.64 |
| Qwen2.5-7B-Instruct | 43.48 | 71.43 | 15.85 |

이는 제안된 시스템이 단순 후처리에 그치지 않고 문맥 기반 반복 오류 억제 시스템으로서 지속적인 성능 개선과 실시간 적용이 가능함을 보여준다.

## Ⅲ. 결론

본 연구는 Whisper 기반 STT 시스템의 할루시네이션 문제를 완화하기 위해, LLM 을 활용한 문맥 탐지 및 교정 프레임워크를 제안하였다. 제안된 방법은 앞뒤 문맥을 고려해 오류를 탐지하고, 반복 표현을 기록하여 후속 회의에서 실시간 교정이 가능하도록 설계되었다.

실제 기업 회의 데이터를 활용한 실험에서, 할루시네이션 발생률은 기존 대비 약 34.5% 감소하였으며, 오류 패턴 기록을 통한 추가 감축 효과도 확인되었다. 또한 다양한 회의 주제에 적용 가능한 일반화된 프롬프트 기법을 통해 범용성과 실효성을 확보하였다.

향후에는 오류 패턴 데이터베이스를 지속적으로 보강하고, 경량화 모델을 도입하여 실시간성과 적용 범위를 더욱 확대할 계획이다. 아울러, LLM 의 문맥 이해력과 오류 식별 정확도를 더욱 향상시키기 위해 프롬프트를 강화할 계획이다.

## 참 고 문 헌


[1] 민동욱, 남승수 and 최대선. (2024). KcBERT 를 활용한 한국어 음성인식 텍스트 정확도 향상 연구. 정보과학회논문지, 51(12), 1115-1124.

[2] 진혜원 , 이아현 , 채예진 , 박수현 , 강유진 and 이수원 (2021). LSTM 기반의 Seq2Seq 모델을 이용한 한국어 음성인식 오류 교정 방법. 한국컴퓨터정보학회논문지, 26(10), 1 - 7.

[3] 이수정, 이하영, 허성수 and 최원익. (2025). 대형 언어 모델 응답의 신뢰성 향상을 위한 환각 탐지 및 설명 모델. 정보과학회논문지, 52(5), 404-414.

[4] Fang, Yangui, et al. "Fewer Hallucinations, More Verification: A Three-Stage LLM-Based Framework for ASR Error Correction."arXiv preprint arXiv:2505.24347(2025).