

실제 강연 환경에서의 실시간 AI 음성번역 오류 사례 분석: 평가 측면의 시사점

배문정(영남대학교)

1. 서론

최근 인공지능(AI) 기술의 비약적인 발전으로 실시간 AI 음성 통번역 서비스가 국제회의, 기업 행사, 대학 강의 등 다양한 현장에 도입되고 있다(권용만, 2025; 노인호, 2025; 한혜선, 2026). 기존의 AI 음성 통번역 시스템은 자동 음성 인식(ASR), 전사(STT), 기계번역(MT), 음성 합성(TTS)이 순차적으로 결합된 구조로 이루어졌는데 최근에는 구글의 제미니(Gemini)나 오픈AI의 챗GPT(ChatGPT) 등에서 음성 입력과 출력을 하나의 신경망이 종단간(end-to-end) 처리하는 음성 대 음성(speech-to-speech) 통역 기능을 잇달아 내놓고 있다.

이러한 흐름 속에서 국내외 통번역학계에서는 AI 기반 음성 통번역 기술의 현장 적용 가능성과 인간통역사의 역할 변화를 검토하기 위한 품질 분석 연구가 축적되고 있다(김혜림, 2025; 배문정, 2025; Fantinuoli & Prandi, 2021; WHO, 2025).

그런데 이들 선행연구에서 활용된 기존의 품질 평가 틀은 실시간 통번역 현장의 멀티모달(multimodal)적 특성을 충분히 반영하지 못하고 있다. 강연이나 회의에 대한 실시간 통역은 단순한 텍스트 변환 과정이 아니다. 발

화가 이루어지는 장소, 발표 자료, 화자의 제스처, 청중과의 상호작용 그리고 화자의 발음, 억양, 휴지, 말실수와 같은 다양한 요소들이 의미 해석에 중요한 영향을 미친다. 기존의 평가 방식은 산출물의 전반적 품질을 정량적으로 비교하는 데는 유용하지만, 오류가 어떠한 입력 조건과 담화적 맥락에서 발생하는지, 그리고 그 오류가 실제 의사소통에서 어떠한 의미 왜곡을 초래하는지를 충분히 드러내는 데에는 한계가 있다.

이에 본 연구는 실제 한국어 강연 현장에서 제공된 AI 기반 실시간 한-영 음성번역과 전문 인간통역사의 순차통역을 비교하여, AI 음성번역에서 나타나는 오류의 유형과 발생 조건을 탐색하고자 한다. 이를 위해 기존 평가 틀을 그대로 적용하여 오류를 미리 정해진 범주로 분류하는 대신, 실제 코퍼스에서 관찰되는 오류 사례를 수집하고 코딩화하여 그 발생 원인과 양상에 따라 비슷한 유형의 오류끼리 범주화하는 방식을 채택한다. 아울러 인간통역사가 어떤 방식으로 오류 가능성을 회피하거나 화자의 의도를 반영하여 통역하는지도 함께 살펴본다. 단, 인간통역의 경우 전체 통역 산출물에 대한 전수 평가 대신 AI 음성번역에서 식별된 오류 지점을 중심으로 동일한 담화 맥락에서 인간통역사가 어떠한 전략을 사용하는지 검토한다. 이는 본 연구의 목적이 점수 비교를 통해 인간통역과 AI 음성번역의 우열을 가리는 것이 아니기 때문이다. 본 연구에서 인간통역 데이터는 실시간 멀티모달 의사소통 환경에서 현 단계의 AI 음성번역이 아직 수행하기 어려운 인간통역 고유의 의미 해석 및 의사소통 조정 기능을 보여주기 위한 비교 기준점으로 기능한다. 궁극적으로 본 연구는 AI 음성 통번역의 특성을 반영한 평가 기준 구축을 위한 실증적 근거를 제공하고자 한다.

한편, 본 연구에서 다루는 핵심 개념과 분석 대상의 범위를 명확히 하기 위해 관련 용어를 다음과 같이 정립하고자 한다. 본 연구의 분석 대상은 연사의 발화가 실시간으로 전사된 후 텍스트로 번역되어 제시되는 형태이다. 이에 본 연구에서는 이를 ‘AI 음성번역’으로 지칭한다. 다만 선행연구에는 실시간 번역 텍스트에 음성 합성이 결합되어 최종 산출물이 음성으로 제공되는 경우도 포함되어 있으므로, 이러한 유형을 포괄하는 용어로는 ‘AI 음성 통번역’을 사용한다. 단, 선행연구 검토 과정에서는 각 연구에서 사용된 고유한 용어(예: 자동 동시 음성번역, AI 통역, 자동자막, 실시간 음성번역

등)을 그대로 반영하여 기술하였음을 밝혀둔다.

2. 문헌 고찰

본 문헌 고찰에서는 실시간 AI 음성 통번역의 품질을 검토한 국내외 선행연구를 대상으로, 각 연구의 분석 대상과 통번역 품질 평가 기준 및 방법을 살펴보고자 한다. 이를 통해 기존 연구들이 실시간 AI 음성 통번역의 품질과 한계를 어떤 관점에서 규명해 왔는지 검토하고, 본 연구의 분석 방법이 기존 평가 방법과 구별되는 지점을 도출하고자 한다.

2.1 실시간 음성 통번역 품질 연구 사례

먼저 판티누올리와 프란디(Fantinuoli & Prandi, 2021)는 Azure 음성 번역(Azure Speech Translation)을 활용한 자동 동시 음성번역을 인간 동시통역과 비교한 연구이다. 이들은 실제 영어 강연 자료를 대상으로 영어-이탈리아어 방향의 자동 동시 음성번역 산출물과 전문 통역사의 동시통역 전사본을 비교하였다. 특히 이 연구는 참조 번역본을 기반으로 하는 BLEU(Papineni et al., 2002), TER(Snoover et al., 2006), METEOR(Banerjee & Lavie, 2005) 등 기존 자동평가 지표가 의사소통 상황을 반영하지 못하는 한계를 지적하며, 티셀리우스(Tiselius, 2009)의 프레임워크를 기반으로 한 사용자 중심 평가 방식을 적용했다. ‘이해 용이성(intelligibility)’과 ‘정보 충실도(informativeness)’의 두 차원에서 산출물을 평가하는 것으로써 통번역 전공 평가자 6명이 각 산출물을 6점 척도로 평가하였다. 분석 결과, 인간통역은 ‘이해 용이성’에서 자동번역을 크게 앞선 반면, ‘정보 충실도’에서는 자동번역이 인간통역과 유사하거나 일부 우세했다. 이는 기계가 발화 내용을 충실하게 전사 및 번역하는 데 강점이 있는 반면, 청자가 쉽게 이해할 수 있도록 정보를 재구성하고 자연스러운 목표 언어 표현을 산출하는 데에는 인간통역이 우세함을 시사한다. 나아가 저자들은 이러한 정량 평가에 다 담기지 않은 인간통역만의 특징을 추가 기술했다. 즉, 인간통역에서 나타나는 일반화, 첨가, 의도적 생략 등의 전략적 개입이 청자의 이해를 돕는 기능을 할 수 있는데, 기계 산

출물에는 이러한 화용적 조정이 거의 나타나지 않는다고 설명하였다.

WHO(2025)는 국제기구의 실제 다국어 회의 맥락에서 AI 통역의 사용 가능성과 위험을 평가한 사례이다. 실시간 음성통역 서비스, Wordly에 2024년 세계보건총회(World Health Assembly) 연설 중 일부를 입력하여 UN 공식 언어 6개 간 실시간 음성통역 품질을 평가했다. 전사본을 만들기는 했지만 평가 대상은 음성 산출물이었다. WHO 언어서비스 직원 및 제네바대 통번역 교육 관련 전문가들이 수행한 평가는 UN 경쟁시험 평가 기준을 기반으로 내용(Content), 표현력(Expression), 전달력(Delivery)의 세 가지 항목에 대해 3단계 척도로 - Good(1), Poor(0.5), Unacceptable(0) - 이루어졌다. 단, AI 통역임을 감안해 ‘내용’에 중점을 두어 평가하고 나머지 두 개 항목은 내용에 영향이 있을 경우에 고려하는 것으로 하였다. 특별히 ‘평판 위험’을 ‘내용’에 대한 세부 항목으로 두었는데 이는 정치적, 외교적 문제를 야기할 수 있는 통역을 의미한다. 평가 대상 음성통역 산출물 90건 중 단 한 건만이 통과 점수(75% 이상)를 받았고 90건 모두에서 적어도 하나의 평판 위험 요소가 발견되었다. WHO 기준에서는 평판 위험이 단 한번이라도 발견되면, 평가에서 통과하지 못하는 것으로 간주했다. 음성 전사(STT), 기계번역(MT), 음성 합성(TTS)의 3단계를 거치다보니 지연 시간이 길어진다는 문제점도 발견되었다.

김혜림(2025)은 중국어-한국어 AI 통역에서 발생하는 오류를 분석하고, 이를 바탕으로 AI 통역의 위험요인을 도출하고자 하였다. 이 연구에서도 실제 있었던 연설문 음성을 중국의 AI 통역 서비스인 iFlytek에 입력하여 한국어 통역을 생성하였다. 이 시스템은 원발화를 음성 인식하여 전사하고, 이를 기계번역한 뒤 자막과 음성으로 출력하는 방식으로 작동한다. 연구자는 원발화(ST1), 자동전사 결과(ST2), 한국어 번역 산출물(TT)의 대조쌍을 구축함으로써, 오류가 원발화의 전사 단계에서 발생했는지, 또는 전사 이후의 번역 단계에서 발생했는지를 구분하여 분석하였다. 평가 기준은 WHO(2025)의 AI 통역 평가 틀을 참고하되 ‘전달력’은 제외하고 크게 ‘내용’과 ‘표현력’을 기준으로 하였다. 평가 항목별로 몇 차례의 오류가 발생했는지를 수치로 보여 주고 AI 통역의 특징 및 위험 요인을 기술하였다. 주요 위험 요인은 비정형적 구어를 적절히 파악하지 못하는 문제, 고유명사 처리 실패, 원발화에 없

는 내용을 생성하는 환각, 문학 텍스트의 문화적, 문학적 의미 전달 실패, 동음이의어 오류 등으로 정리되었다.

두와 루(Du & Lu, 2024)와 배문정(2025)은 모두 피더슨(Pedersen, 2017)의 자막 번역 평가 모델인 FAR 모델을 기반으로 자동자막 및 실시간 음성번역의 품질을 평가하였다. FAR 모델은 원발화의 의미가 목표 언어에 적절히 재현되는지에 관한 기능적 동가성(Functional Equivalence), 목표 언어 표현이 해당 언어의 규범 및 관습에 부합하는지에 관한 수용성(Acceptability), 그리고 자막이 시청자 또는 사용자가 읽고 이해하기에 적절한 형태로 제시되는지에 관한 가독성(Readability)을 함께 고려하며, 오류의 심각도에 따라 가중 감점하는 방식이다. 두 연구 모두 음성 인식 오류와 부적절한 발화 분절이 후속 번역 오류 및 자막 이해의 저하로 이어진다는 점을 확인했다. 그러나 두 연구는 비교 대상과 FAR 모델의 활용 방식에서 차이를 보인다. 두와 루(2024)는 중국어와 영어 영상 각 1편씩을 대상으로 중국어-영어, 영어-중국어 방향의 인간번역 자막과 iflyrec이 생성한 기계 자막을 동일한 FAR 기준으로 점수화하여 인간번역과 기계번역의 품질 격차를 정량적으로 제시하였다. 이에 비해 배문정(2025)은 인간번역 또는 인간통역을 비교 대상으로 설정하지 않고, 영어 및 한국어 연설 영상 각 2건씩을 EventCat이라는 실시간 회의 통역 플랫폼에 입력하여 산출된 영어-한국어 및 한국어-영어 실시간 음성 번역 결과를 대상으로 하였다. 언어 방향별, 연사 특성별로 다르게 나타난 오류의 수와 양상을 진단하였다. 특히 배문정(2025)은 FAR 모델을 실시간 회의 환경에 맞게 확장하여 ‘기능적 동가성’ 차원의 오류를 발화 분절 오류, 음성 인식 전사 오류, 번역엔진 오역으로 세분화하고, 한-영 번역에서는 한국어의 주어 탈락 특성으로 인한 주어 누락 오류를 별도 범주로 설정하였다. 따라서 두와 루(2024)가 FAR 모델을 활용하여 인간번역과 기계자막 간의 상대적 품질 차이를 수치로 보여준 연구라면, 배문정(2025)은 동일 모델을 토대로 실시간 음성번역의 처리 단계와 언어 방향에 따른 오류 발생 기제를 보다 구체적으로 규명한 연구로 볼 수 있다.

2.2 기존 평가 틀의 한계 및 대안적 접근법

지금까지 검토한 선행연구는 서로 다른 명칭의 평가 틀을 사용하였음에

도, 대체로 통역 품질을 일정한 평가 항목에 따라 점수화하고 그 결과를 제시한 뒤, 정량화된 평가만으로 충분히 설명되지 않는 AI 음성 통번역의 오류 양상과 인간통역과의 차이를 별도로 기술한다는 공통점을 보인다. 판티누올리와 프란디(2021)가 활용한 평가 틀에서 정보 충실도(informativeness)는 원발화의 핵심 정보가 목표 언어 산출물에 얼마나 충분히 보존 및 전달되었는지를 평가하는 항목이다. 이는 WHO(2025)와 김혜림(2025)의 내용(Content) 항목, 그리고 두와 루(2024) 및 배문정(2025)이 활용한 FAR 모델의 기능적 동가성(Functional Equivalence) 항목과 상응하는 범주로 볼 수 있다. 또한 판티누올리와 프란디(2021)의 이해 용이성(intelligibility)은 목표 언어 산출물이 수용자에게 얼마나 쉽게 이해되는지를 판단한다는 점에서 WHO(2025)와 김혜림(2025)의 표현력(Expression), FAR 모델의 수용성(Acceptability)과 개념적으로 연결된다. 이처럼 기존 연구의 평가 방법은 명칭과 세부 절차에는 차이가 있으나, 궁극적으로는 인간통역의 수행을 평가하기 위해 발전해 온 의미 전달, 목표 언어 표현, 수용자 이해라는 기준을 조정하여 적용한 것으로 이해할 수 있다(Pöchhacker, 2001).

그러나 이러한 기존 평가 틀은 AI 음성 통번역이 산출하는 오류의 특성과 오류 발생의 과정을 충분히 설명하는 데에는 한계를 보인다. AI 음성 통번역에서는 음성 인식 오류가 기계번역 단계로 전이되고, 발화의 분절 방식, 화자의 발음, 억양, 의미상 불필요한 채움말(filler), 자기 수정 등이 최종 산출물의 품질에 복합적으로 영향을 미친다. 실제로 선행연구들은 점수화된 평가 결과 외에 음성 인식 실패, 부적절한 분절, 환각, 문맥 의존적 표현의 오역 등을 별도의 분석을 통해 기술하거나, 이러한 오류 유형을 반영하기 위한 세부 기준을 추가하는 방식으로 기존의 평가틀을 조정하기도 하였다. 이는 기존의 인간통역 기반의 평가 틀만으로는 AI 음성 통번역의 오류 특성을 충분히 설명하기 어렵다는 점을 방증한다. 또한 배문정(2025)이 지적했듯이, 모든 오류가 결국에는 ‘의미’ 오류로 귀결되어 한 가지 오류가 여러 평가 항목에 중복 반영될 수 있다는 한계도 있다. 이러한 한계점은 ‘맥락화된 강건성(contextualized robustness)’ 개념이 AI 음성 통번역 서비스 평가에 반영될 필요가 있다는 최문선(2026)의 주장을 뒷받침한다.

요약하자면, 선행연구들은 AI 음성 통번역의 성과와 한계를 밝혀내는 데

기여한 동시에, 기존의 인간통역 기반 평가 틀이 AI 음성 통번역 고유의 오류 양상을 충분히 포착하지 못한다는 점도 드러내 주었다. 해당 오류들은 사전에 설정된 항목만으로는 그 발생 조건과 양상을 온전히 설명하기 어려우므로, 실제 오류 데이터로부터 범주를 귀납적으로 도출하는 접근이 선행될 필요가 있다. 이러한 분석 결과는 향후 기술적, 화용적 변수를 반영한 AI 음성 통번역 평가 틀 구축을 위한 기초자료가 될 수 있을 것이다.

3. 연구 방법

3.1 연구 대상

본 연구의 분석 대상은 2026년 6월 국내 한 대학교에 방문한 외국인 견학생들을 대상으로 진행된 실제 강연 코퍼스이다. 한국어로 진행된 해당 강연 현장에는 국내 에듀테크 기업 자이닉스(Xinics)의 학습관리시스템(LMS) 기반 실시간 자막 및 번역 서비스인 트랜스라이브(TransLive)가 도입되어, 한국어 음성 자동 전사본과 영어 번역본이 화면에 실시간으로 제공되었다. 시스템의 세부 기술 사양은 공개되어 있지 않으나, 일반적인 음성번역 시스템과 마찬가지로 (1) 음성 인식, (2) 전사된 텍스트 기반 기계번역, (3) 자막 출력의 단계로 구성된 파이프라인 구조를 따르는 것으로 확인된다.¹⁾ 다만, 현시점 실시간 AI 음성번역이 아직 완벽하지 않다는 점을 감안하여, 현장에는 전문 인간통역사가 배치되어 문장 단위의 순차통역을 병행 제공하였다. 이때 순차통역은 본 연구자가 직접 수행하였다. 실시간 AI 음성번역은 인간통역의 유형 중 동시통역에 가깝다 할 수 있으나 주최 측의 요청에 따라 순차통역이 수행되었다. 연사가 주로 한 문장 단위의 짧은 발화를 이어감에

1) 본 연구는 사용자 인터페이스를 통해 획득한 번역 결과를 대상으로 하며, 특정 기업 또는 제품의 성능을 평가하기 위한 목적을 갖지 않는다. 이에 본 연구의 결과는 특정 시스템의 절대적 성능을 단정적으로 판단하기 위한 것이 아니라, AI 음성번역 시스템이 고맥락 담화 환경에서 작동할 때 나타나는 처리 특성과 오류 양상을 학술적으로 탐색하는 데 초점을 둔다.

따라, 인간통역사의 통역 단위와 AI 시스템의 전사 및 번역 처리 단위가 사실상 거의 일치하는 환경이 조성되었다. 또한 AI 시스템 역시 음성 인식과 기계번역을 거쳐 자막을 출력하기까지의 지연(latency)이 있어, 현장에서는 AI 자막이 인간통역사의 통역 발화보다 늦게 화면에 제시되는 현상이 관찰되기도 하여, 결과적으로 AI 음성번역과 인간통역 간 시간적 제약 조건이 유사한 환경에서 비교가 이루어질 수 있었다.

한편, 본 연구는 고맥락 상황의 강연 코퍼스에 한정된다는 점을 명시하고자 한다. 기본적으로 한국어 담화가 지닌 고맥락적 특성에 더해, 본 강연은 연사의 단독 발화로만 진행되지 않고 청중과의 즉흥적인 상호작용, 시청각 자료 활용, 억양 및 제스처와 같은 비언어적 단서, 그리고 선행 담화에 대한 의존도가 높게 나타난 환경이었다. 따라서 본 연구의 분석 결과 및 AI 음성번역의 오류 양상은 AI 음성번역 일반의 본질적 성능 한계로 일반화하기보다는, 본 연구 대상 강연의 고맥락 조건 하에서 관찰된 특성으로 한정하여 해석하는 것이 타당하다.

전체 강연 시간은 순차통역을 포함하여 총 1시간 40분이나, 본 연구에서는 강연의 처음부터 35분까지(한국어 원문 총 1,502단어)를 분석 대상으로 하였다. 이는 강연자의 익명성을 보장하기 위해 본격적인 주제로 들어가기 전 일반적인 내용을 다루는 부분을 분석 대상으로 하기 위함이다. 아래 결과 분석 부분에서도 강연자 고유의 연구 내용이나 신원이 노출될 우려가 있는 고유명사 등은 언어적 오류 양상과 통역 전략의 특성이 그대로 보존되는 범위 내에서 변경하여 제시하고자 한다. 강연자는 대체로 짧고 간결한 문장을 구사하여 문장당 평균 단어 수는 6.02개, 가장 긴 문장의 단어 수는 20개였다.

3.2 분석 방법

본 연구는 기존의 평가 틀을 적용하여 오류를 미리 정해진 기준이나 범주에 따라 분류하는 접근방식을 따르지 않기로 한다. 상기 ‘문헌 고찰’에서 기술한 바와 같이, 선행연구에서 사용된 기존 평가 틀은 기본적으로 인간통역 평가 기준에 기반하고 있다. 이 때문에 실시간 강연을 문자로 번역하는 환경으로 인해 발생하는 유형의 오역은 충분히 반영하기 어렵다. 또한 기존

의 평가 틀은 오류 범주를 구분하고 있지만, 실제로는 많은 오류가 결과적으로 ‘의미 전달’ 오류로 수렴되어 하나의 오류가 2개 이상의 범주에 중복으로 반영될 여지가 많다는 한계도 있다.

이에 본 연구에서는 수집한 데이터에서 관찰된 현상으로부터 범주를 도출하는 귀납적 접근방식을 채택하였다. 구체적인 분석 절차는 다음과 같다.

첫째, 데이터 코퍼스를 구축하였다. 우선, 실시간 강의 번역 서비스의 시스템 로그를 통해 기록된 한국어 원문 자동 전사본과 실시간 영어 번역본(음성 전사 번역본)을 확보하였다. 그다음 연사의 발화와 인간통역사가 수행한 순차통역 음성 녹음본을 문자로 전사하였다. 최종적으로 발화 원문(Source Speech, SS), 원문 자동 전사본(Transcription, TR), 음성 전사 번역본(Target Text, TT)²⁾, 인간통역본(Human Interpreting, HI)를 구간 단위로 정렬하여 코퍼스를 구성하였다. 여기서 ‘구간’은 한 번에 화면에 표시되는 단위를 의미한다.

둘째, 정렬된 SS-TR-TT-HI 코퍼스의 246개 구간을 모두 검토하여, TT에서 번역 결과가 원문의 의미를 정확하게 전달하지 못하거나 청자의 이해에 혼란을 초래할 가능성이 있는 구간을 추출하였다. 추출된 개별 오류 사례에 대해 오류의 현상적 특성(예: 음성 인식 영향 여부, 지나친 직역, 주어 선택 오류, 목표어 부적합성 등)을 기록하는 1차 개방 코딩을 수행하였다.

셋째, 1차 코딩된 오류 사례들을 대상으로 오류가 발생한 원인을 바탕으로 유사한 양상끼리 범주화하여 중분류 범주를 도출하였으며, 해당 중분류들을 다시 유사한 범주끼리 묶어 대분류로 정리하였다. 기본적으로 모든 오류는 의미 전달 실패로 이어진다는 것을 전제하였기 때문에 기존 평가 틀의 ‘정확성’, ‘정보 충실성’ 등에 해당하는 범주는 설정하지 않았고 의미 전달 실패의 원인에 따라 범주가 분류된 것이다. 또한 처음에는 ‘직역’을 별도의 범주로 두었으나, 직역으로 인한 오류 양상이 다양할 뿐만 아니라 다른 범

2) 본 연구에서 ‘Target Text(TT)’는 음성 전사 이후 생성된 실시간 번역 텍스트를 지칭한다. 일반적으로 이 단계는 ‘기계번역(MT)’으로 표기되기도 하고 ‘음성 인식 기반 번역(STT-based Translation)’으로 설명될 수 있으나, 본 연구는 특정 처리 단계가 아닌 도착어(Target Text)로 된 최종 산출물로서의 번역 텍스트 자체를 분석 대상으로 삼기 때문에 TT라는 용어를 사용한다.

주와도 중첩되는 문제가 있었다. 이에 따라 직역 오류들을 세부적인 성격에 따라 각 중분류 및 대분류에 분산하여 귀속시키기로 최종 결정하였다. 이렇게 함으로써 오류가 2개 이상의 범주에 중복 반영되는 것을 방지하였다. 이러한 과정을 거쳐 최종적으로 4개의 대분류가 확정되었다(<표 1> 참조).

넷째, 도출된 TT의 오류 지점을 중심으로 인간통역사의 번역 변환(Translation Shift)³⁾ 및 대처 양상을 대조 분석하였다. 즉, 인간통역사가 어떠한 맥락적 단서와 전략을 활용하여 AI 음성번역에서 나타난 오류를 회피하였는지 분석하였다.

4. 분석 결과

4.1 TT 오류 분류

본 연구에서 수집된 TT 오류 데이터를 분석한 결과, 오류 유형을 4개의 대분류, 9개의 중분류로 구분할 수 있었다(표 1 참조). 분석 대상 총 246개 구간 중 126개 구간에 오류가 발견되었다. 연사가 대체로 짧은 문장을 구사했기 때문에 대부분 한 구간이 한 문장이었지만 한 문장이 2개 이상의 구간으로 나누어진 경우도 있었다. 한 구간 내에서도 여러 부분에 오류가 있을 수 있기 때문에 오류의 총횟수는 구간의 수보다 많은 139회였다.

전체 오류 발생 양상을 대분류 기준으로 살펴보면, ‘담화 맥락 해석 실패’가 총 55회(39.6%)로 가장 높은 비중을 차지하였다. 뒤를 이어 ‘음성 및 구어체 처리 오류’가 43회(30.9%), ‘정보 누락 및 과생성’이 20회(14.4%), ‘목표어 부적합성’이 19회(13.7%) 순으로 나타났다.

이하 본고에서는 이 4개의 대분류를 중심으로 세부적인 TT 오류 양상을 구체적 사례와 함께 논의하고, 동일한 발화 구간에 대한 인간통역사(HI)의 처리 양상을 비교하고자 한다.

3) 번역 변환(Translation Shift)은 원문을 목표어로 옮길 때 형식적 대응에서 벗어나 형태나 구조가 달라지는 현상을 말한다(Catford, 1965, 이근희, 2004에서 인용됨).

표 1
MT 오류 유형

대분류	중분류	세부 유형(횟수)	횟수
1. 음성 및 구어체 처리 오류	ASR 오류 및 음성 노이즈 ⁴⁾	- ASR 전사 오류(33) - 채움말(filler) 전사로 인한 오역(2) - 연사 말 실수(6)	43
	구어체 및 노이즈 직역	- 반복 및 불필요한 말 그대로 직역(1) - 부정확한 구어체 발화 직역(1)	
2. 목표어 부적합성	어휘 및 어휘 결합 오류	- 어휘 다의성 해석 실패(6) - 단순 오역(1) - 어휘 및 구조 직역 (부자연스러운 영어 표현)(9)	19
	정보 구조 및 문법 오류	- 직역으로 인한 문법 오류(1) - 주어, 목적어가 드러나지 않는 직역(1) - 직역으로 인한 논리 비약(1)	
3. 정보 누락 및 과생성	정보 누락	- 번역 누락(2)	20
	정보 과생성 및 환각	- 없는 말 삽입(3) - 환각 현상(15)	
4. 담화 맥락 해석 실패	생략 성분 자의적 결정	- 생략된 성분(주어, 목적어) 복원 오류(21)	55
	화용(맥락) 및 의도 미반영 직역	- 유머 이해 실패(1) - 연사 의도와 불일치(5) - 지나치게 단정적인 표현으로 직역(1) - 담화 맥락 반영 실패(10) - 비언어적 맥락(멀티모달 통합 실패, 성별 구분 실패, 운율(억양, 휴지 등) 정보) 미반영(11) - 지시어 해석 오류(2)	
	담화 관습 오류	- 문장 유형 오해(1) - 담화 표지 직역(3)	

4) ‘노이즈’는 ‘불순물’을 뜻하며, 본 연구에서는 채움말, 말실수, 중복 반복 발화와 같이 실질적인 의미를 전달하지 않는 표현이 그대로 전사되어 번역에 반영됨으로써 번역 품질을 저해하는 구어체 음성 불순물 전반을 포괄하는 용어로 사용된다.

4.2 TT 오류 분석

4.2.1 음성 및 구어체 처리 오류

음성 및 구어체 처리 오류는 입력 단계에서 발생하는 오류로, 주로 음성 인식(ASR) 오류나 연사의 발화 중 채움말이나 불필요한 반복 및 말 실수, 부정확한 구어체 발화가 그대로 전사되어 이후 번역에 영향을 미치는 유형이다. 이러한 오류는 번역 자체의 문제라기보다 입력 단계의 결함에 기인한 연쇄적 의미 왜곡에 해당한다. ASR 전사 오류는 33건으로 단일 세부 유형으로는 가장 많았다.

<예시 1>의 경우, 연사가 ‘내가’라고 한 것이 ‘내일가’로 잘못 전사되면서 TT에 원문에 없는 ‘tomorrow’가 들어가고 논리적으로도 맞지 않는 문장이 만들어졌다. <예시 2>는 연사가 사실관계를 부정확하게 발화한 사례이다. TT는 연사의 발화를 그대로 직역하여 ‘한국이라는 국가가 건물을 지었다’고 번역한 반면 HI는 연사의 의도가 ‘한국 기업이 지었다’는 것임을 파악하고 표현을 적절히 변환하여 통역하였다.

<예시 1>

SS: 내가 설명하는 것이 이해가 됐습니까?

TR: 내일가 설명하는 것이 이해가 됐습니까.

TT: Did you understand the explanation given tomorrow?

HI: Do you understand what I mean?

<예시 2>5)

SS: 그 건물을 지은 나라는 대한민국입니다.

TT: The country that built the building is South Korea.

HI: That building was built by a Korean company.

4.2.2 목표어 부적합성

목표어 부적합성 오류는 번역 결과가 목표어의 어휘 및 문법 관습에 부합하지 않는 경우를 의미한다. 여기에는 어휘 선택의 부적절성, 다의성 해석 실패, 문법 오류 등이 포함된다. 또한 단어 대 단어 대응에 치우친 번역으로

5) 예시 중 전사가 올바르게 된 경우에는 TR을 별도로 표시하지 않았다.

인해 문장의 정보 구조가 불명확해지거나 논리적 연결이 약화되는 경우도 이에 해당한다. 이 범주는 목표어에서 자연스럽게 적절한 표현을 생성하지 못한 문제, 즉 소위 '어색한 번역'을 전반적으로 포괄한다.

<예시 3>과 <예시 4>는 어휘의 다의성 해석 실패 사례이다. 한국어에서 '친구'는 'friend'의 의미도 있지만 발화자보다 나이가 더 어린 사람을 친근하게 부를 때도 쓰이는 말인데 <예시 3>의 TT는 후자의 의미를 반영하지 못하고 '친구'의 1차적인 뜻으로 번역하였다. <예시 4>에서도 연사는 '질문에 답을 잘하면 상(賞)을 주겠다'는 의미로 말한 것인데 TT는 '소비자에게 팔리는 물건'이라는 뜻의 '상품(商品)'으로 잘못 해석했다. 또한 '다 있다'는 것은 '질문이 있을 때마다 상품이 있다'는 뜻이었는데 '다'를 '모든 종류'라는 뜻으로 해석해 'all kinds of'로 오역했다. <예시 5>는 직역으로 인해 TT에서 부자연스러운 영어 문장이 만들어진 경우이다. TT는 문법적으로는 문제가 없지만, '협력으로'와 '있다'를 그대로 직역함으로써 맥락을 반영하지 못했을 뿐만 아니라 영어 원어민이라면 쓰지 않을 문장을 만들어냈다. HI는 'as a result'를 사용하여 인과관계를 더 명확하게 해주었고 '오늘 프로그램이 있다'는 표현이 '오늘 이 행사를 하고 있다'는 의미임을 파악하여 통역했다.

<예시 3>

SS: 이 친구는 졸업했습니다.

TT: This friend has graduated.

HI: This student has already graduated.

<예시 4>

SS: 내가 이렇게 질문할 때는 상품이 다 있습니다.

TT: When I ask this kind of question, there are all kinds of products available.

HI: Whenever I ask you a question, I have a prize for you.

<예시 5>

SS: 이런 협력으로 오늘 프로그램이 있습니다.

TT: There is a program today through this cooperation.

HI: As a result of this kind of cooperation, we are having this event today.

4.2.3 정보 누락 및 과생성

정보 누락 및 과생성 오류는 번역 과정에서 필수 정보가 생략되거나, 반대로 원문에 없는 내용이 추가되는 경우를 의미한다. 특히 본 연구 대상 실시간 강의 번역 시스템에서는 ‘안녕’ 또는 ‘안녕하세요’가 실제 발화에 없는 데도 15차례나 반복적으로 생성되면서 계속해서 ‘hello’라는 번역이 화면에 나타나는 해당 시스템 고유의 취약점을 보여주었다. 그리고 가끔 ‘야’, ‘있어요’ 같은 말이 원문과 관계없이 화면에 나오면서 그에 대해 ‘Hey’, ‘There is’와 같은 번역문이 제시되었다. <예시 6>은 원문의 일부가 번역되지 않은 누락 사례이다. TT에는 ‘이분이’에 대한 번역이 누락되어 원문의 의미가 완전하게 전달되지 않았다. HI는 연사가 슬라이드에 있는 특정인에 대해 한 말이라는 것을 알았기 때문에 풀어서 통역했다.

<예시 6>

SS: 여러분은 이분이 궁금할 겁니다.

TT: You must be curious.

HI: Maybe you are curious about who he is.

4.2.4 담화 맥락 해석 실패

담화 맥락 해석 실패는 문장 단위를 넘어선 담화적 차원에서의 해석 실패를 의미한다. 대분류 기준으로 가장 큰 비중을 차지한 오류 유형이었고 TT와 HI의 차이가 가장 크게 드러난 오류 유형이기도 하기 때문에 각 중분류에 대해 상세하게 살펴보기로 한다.

4.2.4.1 생략 성분의 자의적 설정

중분류 중 ‘생략 성분의 자의적 설정’은 한국어 구어체에서 흔히 생략되는 주어, 목적어(목적어 생략도 횡수는 주어 생략보다 적지만 빈번함)를 TT가 전후 담화 맥락에서 올바르게 추론하지 못하고 임의로 잘못 설정하는 경우를 가리키며, 단일 세부 유형임에도 21건이 확인되어 비중이 컸다. 이는 처음에는 ‘음성 및 구어체 처리 오류’로 분류할 것인지 검토되었으나, 주어 및 목적어 생략이 연사의 발화 실수가 아니라 한국어의 통상적 화법이며 인간통역사라면 맥락을 통해 무리 없이 추론 가능한 영역이라는 점에서 최종

적으로 ‘답화 맥락 해석 실패’로 재분류하였다. 간단한 예를 들자면, <예시 7>에서 연사는 이전에 재생했던 동영상을 다시 재생하면서 ‘다시 한번 보겠습니다’라고 했고, 이것은 참석자들에게 같은 영상을 다시 한번 보자고 권유하는 상황이었다. 그런데 주어가 없었기 때문에 TT에서는 임의로 주어가 ‘I’로 설정되어 마치 연사 본인만 무엇인가를 다시 한번 보겠다는 의미인 것처럼 번역되었다. HI는 동영상을 본다는 것을 알고 있었기 때문에 ‘watch’라는 동사를 선택한 것을 알 수 있다. <예시 8>에서도 원문에 주어가 생략된 상태에서 시각 정보도 활용할 수 없는 TT는 마치 연사 자신이 장학금을 많이 받은 것처럼 오역한 반면 HI는 장학금을 받은 사람이 슬라이드에 있는 여학생임을 알고 있었기 때문에 주어를 ‘She’로 처리했다.

<예시 7>

SS: 다시 한번 보겠습니다.

TT: I will take another look.

HI: Let's watch it again.

<예시 8>

SS: 장학금도 많이 받았어요.

TT: I also received a lot of scholarships.

HI: She also received a lot of scholarships.

4.2.4.2 화용(맥락) 및 의도 미반영 직역

‘화용(맥락) 및 의도 미반영 직역’은 문장 자체는 문법적으로 성립하나 맥락을 반영하지 않은 직역으로 인해 화자의 실제 의도, 유머, 어조, 문장 유형이 왜곡되거나 지나치게 단정적으로 번역되는 경우이다. 이 중에서도 주목할 만한 것이 ‘비언어적 맥락 정보 미반영으로 인한 의미 왜곡’이다. 연사는 PPT 슬라이드를 보고 가리키거나 참석자들과 소통도 하면서 강연을 진행하는데 AI 음성번역 시스템은 그것을 볼 수 없기 때문에 시각 정보 부재로 인한 오류가 심심찮게 발생했고, 연사의 억양 및 휴지와 같은 운율 정보를 통해 파악할 수 있는 뉘앙스를 파악하지 못해 오류가 발생하는 경우도 있었다.

<예시 9>는 연사가 PPT 슬라이드에 있는 한 여학생의 사진을 가리키며

한 말인데 한국어만 봤을 때는 성별 구분이 되지 않기 때문에 TT에서는 성별이 드러나는 대명사가 잘못 선택되었다. 반면 HI는 연사의 행동과 슬라이드를 모두 보면서 통역했기 때문에 쉽게 지시 대상의 성별을 파악할 수 있었다. <예시 10>은 연사가 강연 중에 더워서 겉옷을 벗겠다고 양해를 구하는 상황이었는데 TT의 번역만 보면 겉옷뿐이 아니라 여러 개의 옷을 벗는 느낌으로도 읽힐 수 있다. 한국어에서는 통상적으로 쓰이는 표현이지만, 이를 그대로 직역하다 보니 연사의 의도가 제대로 전달되지 않았다. 반면 HI는 상황에 따른 연사의 의도를 파악해 ‘jacket’이라는 단어를 선택했다.

<예시 11>은 연사가 ‘원장’이라는 직책의 이씨 성을 가진 사람을 부르며, 사진에 있는 학생이 상을 받았는지를 물어보는 상황이다. 그런데 전사본에는 ‘이 원장’ 다음에 쉼표가 없어서 TT는 ‘이’를 성씨가 아닌 지시사 ‘this’로 잘못 처리했고, 전사본에 물음표가 없어서 번역본도 의문문이 아닌 평서문이 되었다. 또한 ‘애’는 사진 속 학생을 의미하는데 TT에서는 ‘이원장’ 자신으로 번역이 되었다. 반면 인간통역사는 연사가 ‘이 원장’을 불러서 물어보는 상황이라는 것을 연사의 어조와 손짓을 통해 쉽게 알 수 있었다. 또한, 연사가 ‘상을’이라는 목적어를 말하지 않아서 TT에서는 무엇을 받았는지가 특정되지 않아 임의로 ‘it’을 목적어로 처리했다. 반면 HI는 연사가 목적어를 말하지 않았더라도 앞에서 학생들이 상을 받았다는 이야기가 나온 것을 기억하고 ‘the award’를 목적어로 두었다. 이 예시를 통해 아무리 짧은 문장이라도 시각 정보, 어조 억양과 같은 운율 정보 그리고 상황 정보가 없는 AI 음성번역은 많은 오류를 생성할 수 있음을 알 수 있다.

<예시 12>도 AI 음성번역 시스템이 시각 정보를 활용하지 못해 발생한 오류이다. 연사는 글로벌 기업들의 순위 변화 추이를 보여주며 과거 최상위에 있던 기업의 순위가 최근 크게 하락했음을 전달하고자 했다. 그런데 TT에서는 마치 ‘이 회사가 몰락하는 것을 지켜보라’라는 뉘앙스로 번역되었으며, ‘한번’의 용례 또한 잘못 파악되었다. 한국어의 ‘한번’은 횟수로서의 ‘한 차례’를 뜻하기도 하지만 맥락상 제안이나 권유를 나타내기도 한다. 여기서는 후자의 의미인데 전자로 오역한 것이다. 이에 비해 HI는 표현상에 개선의 여지가 있기는 하지만 연사의 슬라이드를 함께 보고 있었기 때문에 기업의 순위 하락이라는 맥락을 인지하고 통역에 반영했다.

<예시 9>

SS: 중국 학생입니다.

TT: He is a Chinese student.

HI: She is from China.

<예시 10>

SS: 옷 좀 벗겠습니다.

TT: I will take off some clothes.

HI: Let me take off my jacket.

<예시 11>

SS: 이원장, 애가 받았어요?

TR: 이 원장 애가 받았어요.

TT: This director received it.

HI: Director Lee, did this student receive the award?

<예시 12>

SS: 이 회사가 떨어지는 걸 한번 보십시오.

TT: Watch this company fall once.

HI: Look at how this company is going down on the list.

4.2.4.3 담화 관습 오류

‘담화 관습 오류’ 중분류에서 가장 주목할 만한 세부 오류 유형은 ‘담화 표지 직역’이다. 한국어에는 구체적인 의미는 없지만 대화를 부드럽게 만들기 위해 관습적으로 쓰이는 담화 표지들이 있다. <예시 13>의 ‘그렇게 하겠습니다’가 그에 해당하는데 TT에서는 이러한 담화 표지의 기능이 인식되지 못하고 그대로 직역되었다. 반면 HI는 ‘그렇게 하겠습니다’에 의미 있는 정보가 없다는 것을 파악하고 이를 생략했다. <예시 14>에도 ‘그렇게’라는 담화 표지가 포함되어 있는데 TT는 그대로 직역해 ‘like that’이라고 번역한 반면 HI는 ‘그렇게’를 생략하고 문장을 재구성했다. 또한 원문에 ‘시간’이라는 단어가 두 번 등장하는데 TT는 두 번 모두 ‘시간’을 ‘time’으로 직역하여 전체적으로 부자연스럽고 경직된 문장이 만들어졌다. 이에 비해 HI는 ‘이 짧은 시간이’의 ‘시간’은 ‘해당 시간에 이루어지는 활동 즉, 교육’으로 그리고

‘생각하는 시간’은 ‘생각하는 기회’로 해석하여 각각 ‘training’과 ‘opportunity’로 번역했다.

<예시 13>

SS: 제가 준비한 자료를 말씀드리면서 저도 소개하면서 그렇게 하겠습니다.

TT: I will introduce myself while presenting the materials I have prepared, and I will proceed that way.

HI: I'd like to show you the slides I prepared, and at the same time I also would like to introduce myself.

<예시 14>

SS: 이 짧은 시간이 그냥 편안하게 그렇게 한번 같이 생각하는 시간이면 좋겠습니다.

TT: I hope this short time can simply be a comfortable time to think together like that.

HI: I hope that this lecture gives you an opportunity to think about things together, even though it's a short training.

4.3 TT 대비 HI 특성 분석

본 연구에서 원문(SS)과 전사본(TR), 음성 전사 번역본(TT), 인간통역사의 통역(HI)을 분석한 결과, 전반적으로 HI는 어휘 층위의 등가성보다 발화 전체의 의사소통적 의도를 우선시하는 전략을 취하고 있음을 알 수 있었다. TT와 구분되는 HI의 대표적인 처리 전략을 정리해 보면 다음과 같다.

첫째, HI는 담화 맥락을 적극적으로 반영하여 문장을 재구성하는 경향을 보인다. 원문에서 명시되지 않은 정보라도 발화 상황, 청중, 화자의 의도를 고려하여 의미를 명시화하는 방식이다. 상기 <예시 14>에서 ‘시간’을 통역한 방식이 대표적인 예가 될 수 있다. 아래 <예시 15>에서도 TT는 SS의 표면적인 의미만 전달했지만, HI는 당시 상황과 연사의 의도를 정확히 파악하여 통역했다. 즉, 학교 측이 손님들을 특정 강의실로 초대한 상황이라는 점과 해당 강의실의 기능을 강조하려는 연사의 의도를 반영해, 원문에는 없는 ‘we invited you’를 추가하여 통역한 것이다.

<예시 15>

SS: 이 강의실로 오게 된 이유가 있습니다.

TT: There is a reason why you came to the lecture room.

HI: There is a reason that we invited you to this lecture room.

둘째, HI는 단순 음성 텍스트 처리에 국한되지 않고, PPT 자료, 연사의 제스처, 발화의 운율, 현장 분위기 등 멀티모달(multimodal) 정보 통합 능력을 발휘한다. TT는 텍스트 이외의 시각적, 운율적, 상황적 정보를 인식하지 못해 오역을 발생시키는 반면 HI는 이러한 모든 비언어적 정보를 통합하여 즉각적으로 통역에 반영한다. 상기 4.2.4.2의 <예시 9>, <예시 10>, <예시 11>, <예시 12>가 대표적인 예이다.

셋째, HI는 원문에서 생략된 주어, 목적어, 지시어 등을 맥락에 따라 적절하게 복원한다. 특히 한국어 발화에서 빈번히 나타나는 주어 생략은 TT에서 대명사 오류로 이어지는 경우가 많았으나, HI는 상황을 기반으로 적절한 주체를 복원하였다(상기 <예시 7>, <예시 8>, <예시 9> 참조).

넷째, HI는 의미 없는 채움말이나 담화 표지는 생략하거나 축약한다. 예를 들어, 상기 <예시 13>, <예시 14>에서와 같이 TT에서는 ‘그렇게 하겠습니다’, ‘그렇게’와 같은 표현들이 그대로 직역되어 부자연스러운 문장이 생성된 반면, HI는 이를 제거하거나 영어적으로 자연스럽게 문장을 재구성하였다.

다섯째, 연사의 발화 과정에서 발생하는 말실수, 불완전한 문장, 논리적 비약에 대해 HI는 연사의 진의를 해석하여 보정된 통역을 한다. 연사가 잘못 말하거나 문장을 완결하지 않은 경우에도 HI는 전체 맥락을 기반으로 의도된 의미를 추론하여 전달하는 것이다. 예를 들어, 아래 <예시 16>에서 연사는 ‘세계에서 가장 잘사는 사람 A’를 실수로 ‘세계에서 가장 잘사는 나라 사람 A’라고 잘못 발화했다. TT는 이를 그대로 직역한 반면, 앞서 A라는 인물이 언급된 맥락을 기억하고 있던 인간통역사는 연사의 원래 의도를 파악하여 ‘세계에서 가장 잘사는 사람 A’로 수정하여 통역했다.

<예시 16>

SS: 세계에서 가장 잘 사는 나라 사람 A.

TT: A, a person from the richest country in the world.

HI: The richest man in the world, A.

이와 같이 본 연구에서 분석한 TT와 HI의 주요 처리 방식 및 특성을 비교하여 정리하면 <표 2>와 같다.

표 2

TT와 HI 처리 방식 비교

구분	TT	HI
처리 단위 및 방식	표면 구조에 따른 기계적 직역	담화 맥락에 따른 문장 재구성
멀티모달 정보 통합	ASR 기반 텍스트에만 의존	시각, 청각, 상황 등 비언어적 맥락을 통합적으로 판단
구어체 처리	채움말, 담화 표지 문자 그대로 직역	불필요한 입력을 필터링하고, 목표어 관습상 자연스러운 표현으로 전환
생략 성분 복원	문맥 파악 불가로 잘못된 대명사 선택	맥락에 맞는 주어, 목적어 및 기타 지시어 복원
발화 오류 대응	ASR 전사 오류 및 연사의 말실수를 그대로 번역	연사의 진의를 파악하여 수정

5. 논의

5.1 TT 오류의 특성과 함의

본 연구에서 도출된 오류 유형은 실시간 강연에 대한 AI 음성번역이 텍스트 기반 기계번역뿐 아니라 인간통역과도 다른 조건에서 작동한다는 점을 보여준다. 특히 4개의 대분류 중 ‘담화 맥락 해석 실패’가 55회로 가장 큰 비중을 차지했다는 결과는 AI 음성번역 시스템의 한계가 어휘 문법 수준의 정확성 문제보다 담화 및 화용 층위의 정보 처리 실패에서 더 두드러지게 나타남을 시사한다. 나아가 이러한 결과는 실시간 AI 음성 통번역이 인

간통역과 달리 제스처, 시선, 시각 자료 등 비언어적 단서와 선행 발화와의 연계성을 충분히 통합하지 못한다는 점에서, 단순히 ‘자동화된 통역’으로 간주되기 어렵다는 사실을 드러낸다. 그런데 기존 평가 틀은 이러한 멀티모달 상호작용적 특성에서 발생하는 오류를 충분히 포착하기 어렵다는 한계가 있다. 이는 실시간 AI 음성 통번역의 특성을 반영한 새로운 평가 기준이 필요함을 시사한다.

여기서 추가로 짚어야 할 문제는 오류의 ‘개수’와 오류의 ‘무게’가 반드시 비례하지 않는다는 점이다. 예컨대 <예시 8>에서처럼 ‘She received a lot of scholarships’이어야 할 문장이 주어 하나만 잘못 대응되어 ‘I received a lot of scholarships’로 번역된 경우, 단어 단위 오역은 단 하나에 불과하다. 그러나 이 한 단어로 인해 지시 대상이 완전히 달라지면서, 청중은 실제로는 슬라이드 속 학생에 관한 정보를 연사 자신에 관한 정보로 오해하게 된다. 결국 BLEU나 TER 같은 표면적 텍스트 유사도 기반 평가에서는 90% 이상의 높은 점수로 산출될 수 있으나, 작은 어휘 오류 하나가 텍스트 전체의 명제적 의미를 심각하게 왜곡할 수 있음을 보여준다.

한편, 자동 음성 인식(ASR) 단계에서의 전사 오류가 이후 번역 단계로 연쇄적으로 이어져 오역을 유발하는 ‘ASR 전사 오류’는 총 33회로 단일 세부 유형 중 가장 높은 빈도를 보였다. 이는 음성 인식 기술이 지속적으로 발전하고 있음에도 불구하고 실시간 음성번역 환경에서는 여전히 의미 왜곡의 주요 원인으로 작용하고 있음을 보여준다. 최근 등장하고 있는 음성 대 음성(speech-to-speech) 통번역 시스템에서는 이러한 연쇄 오류가 어느 정도 완화될 수 있을지, 혹은 다른 형태로 나타날지에 대해 살펴보는 것도 흥미로운 연구 주제가 될 수 있겠다.

또한 ‘생략 성분의 자의적 결정’이 단일 세부 유형만으로 21건이 확인되었다는 점은 한국어의 화용론적 특성, 즉 주어, 목적어의 빈번한 생략이 실시간 AI 음성 통번역의 취약점으로 직결됨을 보여준다. 이는 한국어와 영어의 언어 계통적 거리가 실시간 번역 환경에서 오류를 증폭시키는 요인으로 작동함을 의미한다.

5.2 HT의 대응 전략이 갖는 함의

분석 결과에서 드러난 HI의 다섯 가지 대표적인 처리 전략 - 담화 맥락 기반 문장 재구성, 멀티모달 정보 통합, 생략 성분 복원, 담화 표지 생략, 발화 오류 보정 - 은 공통적으로 ‘텍스트 표면을 넘어선 의미 추론’이라는 특성으로 수렴한다. 이는 통역이 단순한 언어 치환이 아닌, 맥락과 화자의 의도에 대한 해석에 기반한 의미 전달 행위임을 재확인해 준다. 특히 <예시 14>, <예시 15>와 같이 원문에 명시되지 않은 정보를 능동적으로 추가하는 HI의 전략은 원문 자체에 대한 충실성보다 의사소통적 충실성을 지향하고 있음을 보여준다.

다만 본 연구의 목적은 인간통역이 기계번역보다 전반적으로 우수하다는 점을 입증하는 데 있지 않다는 점을 분명히 할 필요가 있다. 실제로 본 연구의 HI 데이터에서도 문장이 다소 장황하거나, 사소한 문법적 오류 또는 일부 정보의 누락과 같은 크고 작은 실수가 관찰되었다. 그러나 이러한 HI의 실수는 TT에서 나타난 오류와는 그 성격이 근본적으로 다르다. TT의 오류는 대부분 전체적인 맥락을 비롯한 비언어적 정보에 대한 이해의 부재에서 비롯되어 지시 대상 오인 및 의미의 왜곡으로 이어지는 반면, HI의 실수는 맥락을 정확히 파악한 상태에서 발생하는 표현상의 미숙함이 대부분이다. 즉 두 오류는 발생 층위 자체가 다르며, 이 차이야말로 실시간 멀티모달 환경에서 인간통역사가 수행하는 고유한 역할이 무엇인지를 보여주는 지점이라 할 수 있다.

6. 결론

6.1 연구 요약

본 연구는 실제 한국어 강연 현장에서 산출된 실시간 AI 음성번역과 전문 인간통역사의 순차통역을 비교하여, AI 음성번역의 오류 양상과 인간통역사의 대응 전략을 분석하였다. 이를 위해 한국어 원문(SS), 원문 전사본

(TR), 음성 전사 번역본(TT), 인간통역 전사본(HI)을 문장 단위로 정렬한 코퍼스를 구축하였다. 특히 인간통역에 대해서는 전체 산출물을 동일한 기준으로 전수 분석하기보다는, TT에서 오류가 나타난 구간을 중심으로 해당 지점에서 인간통역사가 어떠한 번역 변환과 담화적 조정을 수행하는지를 대조적으로 분석하였다.

TT 오류 분석 결과, 총 246개 분석 구간 가운데 126개 구간에서 오류가 확인되었으며, 오류 발생 횟수는 총 139회였다. 도출된 오류는 크게 음성 및 구어체 처리 오류, 목표어 부적합성, 정보 누락 및 과생성, 담화 맥락 해석 실패의 네 범주로 구분되었다. 이 가운데 담화 맥락 해석 실패는 총 55회로 전체 오류의 약 40%를 차지하여 가장 높은 비중을 보였다.

한편 인간통역은 담화 맥락에 따라 문장을 재구성하고, 멀티모달 정보를 통합하며, 생략된 성분을 복원하고, 채움말과 담화 표지를 선택적으로 생략하며, 연사의 말실수에 대해서도 발화의 진의를 추론하여 전달하는 등 AI와 뚜렷이 구별되는 처리 방식을 보였다.

6.2 연구의 의의

본 연구의 의의는 다음과 같다. 첫째, 본 연구는 실험실에서 통제된 발화나 유튜브 영상이 아니라 강연자가 PPT 자료를 활용하고 청중과 질문과 응답을 주고받는 실제 의사소통 상황에서 산출된 데이터를 분석 대상으로 삼았다는 점에서 의의가 있다. 동일한 현장에서 AI 기반 실시간 음성번역과 전문 인간통역사의 순차통역이 병행됨에 따라, 동일 발화에 대해 기계와 인간의 처리 방식이 어떻게 다른지 직접적으로 비교할 수 있었다.

둘째, 한국어-영어 단일 언어 방향과 강연 일부 구간만을 대상으로 삼았고 연사의 발화가 문장 구조와 내용 면에서 비교적 단순했던 만큼, 복잡하고 전문적인 발화에 대한 분석까지는 이루어지지 못했다는 한계가 있다. 하지만 단순한 발화 데이터를 활용함으로써 역설적으로 실시간 음성번역이 지닌 전형적인 한계점들을 명확하게 드러내는 데 기여했다고 볼 수 있다.

셋째, 연구자가 인간통역을 직접 수행함으로써 그 처리 과정을 보다 면밀히 해석할 수 있었다는 점도 본 연구의 강점이다. 통역 당시 화자의 발화 의도, 강연 자료, 제스처와 억양, 청중의 반응 등 통역에 영향을 미친 상황

적 요인을 직접 경험한 덕분에, 인간통역 산출물에 나타난 정보의 명시화, 생략, 재구성, 오류 보정 등이 어떠한 의도와 판단에 기초했는지 직접적으로 분석할 수 있었다.

넷째, 본 연구에서 도출된 오류 유형과 발생 조건은 향후 실시간 멀티모달 환경의 특성을 충실히 반영하는 새로운 AI 음성 통번역 품질 평가 기준 및 프레임워크를 수립하는 데 실질적인 자료로 활용될 수 있을 것이다.

6.3 연구의 한계 및 후속 연구 제언

다만 본 연구에는 몇 가지 한계가 있다. 우선 앞서 언급했듯이, 분석 대상이 단일 대학 강연의 한국어-영어 방향 통번역 결과물에 한정되어 있다. 또한 특정 음성번역 서비스 하나만을 활용했고, 비교적 짧고 간결한 문장을 구사하는 강연자 1인의 발화만을 다루었으므로 결과를 일반화하는 데에는 신중할 필요가 있다.

또한 본 연구는 AI 음성번역(TT)의 오류가 나타난 구간을 중심으로 인간통역(HI)의 대응 양상을 대조하여 분석하는 방식으로 이루어졌기 때문에, 인간통역 전체 산출물에 대한 체계적인 분석이 이루어지지 않았다는 점에서 한계를 지닌다. 따라서 후속 연구에서는 동일한 원문 데이터에 대해 인간통역 산출물 역시 오류 유형, 처리 전략, 담화 구성 방식 등을 보다 정밀하게 분석하고, 양측 결과를 균형 있게 비교할 필요가 있다. 이를 통해 AI 음성번역과 인간통역 간의 차이를 보다 종합적으로 규명할 수 있을 것이다.

더불어 본 연구의 오류 판정, 범주화, 인간통역 분석 과정은 연구자에 의해 단독으로 수행되었으며, 제2 평가자 또는 전문가 검토를 통한 상호 검증 절차를 포함하지 못했다는 점에서도 일정한 한계를 가진다. 향후 연구에서는 다수의 평가자를 통한 교차 검증이나 평가자 간 신뢰도를 확보함으로써 분석의 객관성과 재현성을 더욱 강화할 필요가 있다.

그리고 연구자가 인간통역을 직접 수행함으로써 통역 당시의 맥락과 전략을 구체적으로 분석할 수 있었다는 장점이 있는 반면, 그 해석이 연구자 본인의 자기 보고에 의존했다는 한계가 있다. 따라서 향후 언어 방향, 음성번역 서비스 종류, 인간통역 코퍼스 측면에서 연구 대상을 확대함으로써 본 연구에서 도출된 결과의 타당성을 검증할 필요가 있다.

또한 본 연구에는 사용자 수용성 측면의 평가는 포함되어 있지 않다. 예를 들어, 번역 결과가 제시되기까지의 지연 시간(latency), 화면상 텍스트의 제시 방식과 가독성, 자막의 정보 밀도, 청중의 인지적 부담과 이해도 등이 모두 청중의 수용성에 영향을 미칠 것이다. 이러한 요소를 고려한 사용자의 전반적인 만족도 평가도 병행되어야 AI 음성 통번역 서비스 품질에 대한 종합적인 평가라 할 수 있을 것이다.

선행연구와 본 연구에서 다룬 실시간 AI 음성 통번역 시스템은 대체로 음성인식(ASR), 기계번역(MT), 그리고 일부는 음성합성(TTS)까지 순차적으로 거치는 단계별(cascaded) 방식에 기반한다. 최근에는 중간 전사 과정 없이 음성 입력을 목표 언어 음성 또는 텍스트로 직접 변환하는 종단간(end-to-end) 음성번역 방식도 발전하고 있으므로, 향후 연구에서는 이러한 종단간 방식과 단계별 방식의 통번역 양상과 품질의 차이를 연구하는 것도 의미가 있을 것이다.

참고문헌

- 권용만. (2025. 6. 18.). “AI통역에 청중도 놀랐다” 플리토 ‘라이브 트랜스레이션’ 서비스 써보니. IT조선. <https://it.chosun.com/news/articleView.html?idxno=2023092142684>
- 김혜림. (2025). 중한 AI통역의 위험요인 분석: 讯飞同传(iFlytek Simultaneous interpretation)을 중심으로. 통번역교육연구, 23(2), 57-87.
- 노인호. (2025. 9. 8.). 경북대, 모든 정규 과목 'AI기반 실시간 통번역 서비스' 제공. 조선일보. <https://www.chosun.com/national/regional/2025/09/08/MSJY2TIMOBCIVNHZAAQSSOPEYU/>
- 배문정. (2025). 온라인 회의 환경에서 실시간 음성 번역 성능 분석: FAR 모델 기반 EventCat 사례 연구. 번역학연구, 26(3), 395-423.
- 이근희. (2004). 번역의 변환(Shift) 현상에 관한 소고: 정보제공 목적의 텍스트 중심. 번역학연구, 5(2), 113-144.
- 최문선. (2026). 동시음성번역(SimulST) 서비스의 평가: 실사용 환경을 고려

- 한 평가 접근법 제안. 통역과 번역, 28(1), 113-141.
- 한혜선. (2026. 8. 3.). AI 통번역으로 20개국 석학 연결...엑스엘에이트, 'DH2 026' 실시간 번역 지원. Venturesquare.
<https://www.venturesquare.net/1102940/>
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72).
<https://aclanthology.org/volumes/W05-09/>
- Du, J., & Lu, J. (2024). A comparative study on the translation quality between human and machine-generated subtitles. *2024 6th International Conference on Natural Language Processing* (pp. 62-66). <https://ieeexplore.ieee.org/document/10692675>
- Fantinuoli, C., & Prandi, B. (2021). Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)* (pp. 245-254).
<https://aclanthology.org/2021.iwslt-1.29/>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). <https://aclanthology.org/P02-1040/>
- Pedersen, J. (2017). The FAR model: Assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, 28, 210-229.
- Pöschhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta*, 46(2), 410-425.
- Snoover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223-231).

<https://aclanthology.org/2006.amta-papers.25/>

Tiselius, E. (2009). Revisiting Carroll's scales. In C. V. Angelelli & H. E. Jacobson (Eds.), *American Translators Association Scholarly Monograph Series* (Vol. 14, pp. 95-121). John Benjamins.

WHO. (2025). *Report on WORDLY AI interpretation*. <https://aiic.fr/wp-content/uploads/2025/06/rapport-oms.pdf>

Error patterns in real-time AI speech translation in an authentic lecture setting: Implications for evaluation frameworks

Munjung Bae

School of Global Interpretation and Translation, Yeungnam University

Abstract

This study investigates error patterns in AI-based real-time speech translation by comparing AI outputs with human interpreting in a live Korean-to-English lecture setting. The analysis draws on a corpus comprising source speech (SS), automatic speech recognition transcripts (TR), machine-translated outputs (TT), and human interpreter renditions (HI). Rather than applying a predefined evaluation framework, the study inductively categorizes TT errors according to their underlying causes. The analysis identifies four major categories of errors: errors in processing speech and colloquial language, target-language inadequacy, information omission and over-generation, and discourse-context errors. Of these, discourse-context errors account for the largest proportion, closely followed by errors related to speech and colloquial language processing. This highlights the limitations of current AI speech translation systems in handling spoken-language variability and integrating multimodal and contextual cues. In contrast, human interpreting performs effectively in these areas through strategies such as compensating for disfluencies, context-sensitive reformulation, multimodal integration, and pragmatic adjustment. The findings suggest that evaluation frameworks for AI speech translation that draw on human interpreting assessment may fail to capture the distinctive error profile of AI speech translation. They therefore point to the need for frameworks that incorporate AI-specific error sources such as ASR errors (both speaker disfluency-induced and system-induced) and discourse-level misinterpretation.

Keywords: AI-based speech translation; error patterns; evaluation frameworks; multimodal communication; human interpreter strategies

키워드: AI 음성번역, 오류 양상, 평가 틀, 멀티모달 커뮤니케이션, 인간 통역사 전략

배문정(<https://orcid.org/0000-0002-8847-4176>)

영남대학교 글로벌통번역학부 영어통번역전공 조교수

m.bae7811@gmail.com

논문 투고일: 2026년 8월 12일

1차 심사 완료일: 2026년 9월 1일

2차 심사 완료일: 2026년 9월 8일

게재 확정일: 2026년 9월 15일