대화형 AI 기반 다국어 자동통역 서비스 현황과 핵심 기술 동향

김상훈 한국전자통신연구원 책임연구원

1. 머리말

최근 K-팝, K-드라마 등 한류 확산으로 외국인 방문이 급증하고 있다. 과거 미국, 중국, 일본이 주류였던 관광객 국적도 이젠 유럽, 남미, 동남아, 구소련 국가까지 매우 다양해지고 있다. 산업적으론 코로나 팬데믹 이후 메타버스, 가상현실 기반 원격 회의, 원격협업, 원격 교육 등을 통해 물리적 제약이 없는 가상공간에서 외국인과의 비대면 의사소통이 빈번해지고 있다. 이에 따라 해외여행자 대상 긴급상황에서의 간단한 언어소통에서부터 난이도가 높은 비즈니스 회의 통역까지, 다양한 분야에서 자동통역 서비스 수요가 증가하고 있다.

자동통역은 모국어로 외국인과 의사소통을 가능하게 해주는 기술로, 인간 음성언어와 외국어 지식을 모사하는 대표적 AI 기술이다. 이에 이번 원고에선 최근 사회적 현안이 되고 있는 언어장벽 해소를 위해, 대화형 AI 기반 다국어 자동통역 기술의 국내외 서비스 동향을 살펴보고자 한다. 더불어 최근 서울지하철 11개 주요 역사에 실시 중인 다국어 자동통역 서비스에 적용된 핵심 기술도 살펴본다.

2. 국내외 동향

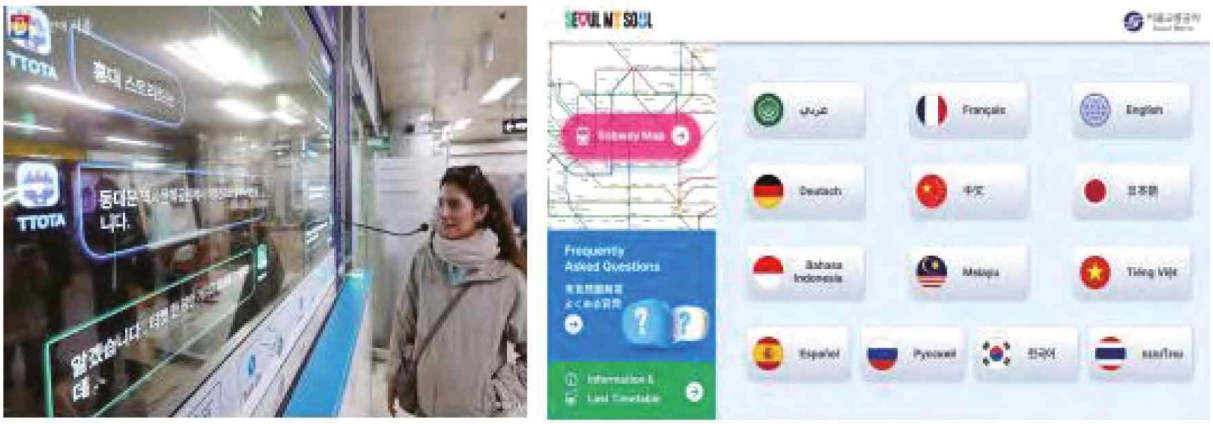
2.1 국내외 서비스 현황

자동통역 기술 연구개발은 1990년대부터 주로 국가 연구기관, 글로벌 다국적 기업을 통해 진행돼 왔다. 다만 당시 시행됐던, 휴리스틱에 의존한 규칙 기반이나 확률/통계적 방식은 음성·언어와 같은 비정형 데이터 모델링에 한계가 있었다[1]. 2010년대 초 딥러닝 방식이 패턴인식에 뛰어난 능력을 보임에 따라, AI 기술이 비약적으로 발전했다. 2020년대엔 비정형 데이터의 특성을 매우 잘 학습할 수 있는 트랜스포머 AI 모델이 나오면서, 통역 체감성능을 일상생활에 적용 가능한 수준까지 개선할 수 있게 됐다.

최근 출시된 갤럭시 S24 스마트폰에는 온디바이스 자동통역이 탑재돼 있다. 이는 전화를 통해 외국인과 실시간 의사소통은 물론, 여행하는 동안 발생하는 긴급상황, 식당 이용, 호텔 예약, 관광정보 안내, 비즈니스 대화 등 다양한 상황에서의 통역을 제공한다. 네이버는 축적된 음성 검색 기술과 번역 기술을 점차 통역으로 확대하고 있다. 현재 한국어·영어·일본어·중국어 등 총 13개 언어 간 번역과 함께, 스페인어·불어 등 5개 언어 통역이 지원되고 있다.

2018년 평창 동계 올림픽에서 세계 최초 자동통역 공식서비스를 실시했던 ETRI(한국전자통신연구원, Electronics and Telecommunications Research Institute)는 2024년 서울 명동 지하철역에

서 13개 언어 자동통역 시스템을 시범 운영하고 있다. 해당 시스템은 지하철 역사 내 개찰구 기기 작동 소리, 스피커 안내음성, 주변 대화 잡음 등 고소음 환경에서 인식·번역이 가능하도록 잡음을 증강한 AI 모델링 기술을 적용했다. 현재 한국어·영어·일본어·중국어·프랑스어·스페인어·독일어·러시아어·태국어·베트남어·말레이시아어·인도네시아어·아랍어 등 13개 언어가 지원된다.



출처: 서울특별시, '자동 통역되는 투명 스크린, 서울지하철 11개 역으로 확대된다', 2024. 3.

[그림 1] 서울지하철 역에 설치된 13개 언어 다국어 자동통역 서비스

서울지하철은 미국 뉴욕, 일본 동경과 함께 세계 3대 메트로 시스템으로 꼽힌다. 그중에서도 외국인이 가장 많이 이용하는 명동을 비롯해 광화문, 강남 등에 국내 자체 개발한 자동통역 기술이 활용되고 있다. 이러한 사례는 외국인에게 한국 관광의 편의성을 제공하고 K-테크의 우수성을 알린다는 큰 의미를 가진다[2].

국외에선 Chat-GPT를 개발한 미국오픈 AI(Open AI)가 60개 언어를 인식할 수 있는 '위스퍼(Whisper)' 모델을 공개했다. 오픈 AI는 또한 다국어 번역 기능도 제공해 국내 사용자에게 유료로 서비스하고 있다.

메타(Meta)는 53개 다국어 사전학습 모델을 공개했으며, 100여 개 언어에 대한 다국어 통역 서비스를 제공하고 있다. 더불어 메타는 2022년 "메타버스 환경에서의 글로벌 언어소통을 위한 '바벨 피쉬(Babel Fish)' 전략을 수립, 세계 100여 개 언어 간 자동통역 기술을 제공한다"고 언론에 공개한 바 있다.

마이크로소프트(Microsoft)는 스카이프 영상통화로 실시간 대화형 통역을 시연했으며, 통역 앱인 '마이크로소프트 트랜스레이터(Microsoft Translator)'를 통해 70개국 자동통역을 제공하고 있다. 특히 영어-중국어 뉴스 통역 서비스는 인간 통역사와 거의 동등한 수준의 성능을 갖고 있다.

아마존(Amazon)은 AI 비서 '알렉사(Alexa)'를 통해 영어에서 다른 48개 언어로 자동통역을 제공하고 있다. 이는 쇼핑몰 아마존닷컴의 외국인 고객에게 편의를 제공하기 위함이다. 아마존은 이에 더해 '아마존 트랜스레이트(Amazon Translate)'를 통해 55개 언어 간 통역 기능을 유료로 제공하고 있다. 중국에선 최대 전자상거래 업체 알리바바(Alibaba)가 세계 최초 라이브 커머스용 실시간 자동통역을 시연한 바 있다. 알리바바는 2020년부터 중국어를 영어·러시아·스페인어로 통역하는 라이브 커머스용 실시간 자동통역 방송을 제공하고 있다.

텐센트(Tencent)는 2020년 유엔, 국제회의용 자동통역 솔루션 공급업체로 선정돼 15개 언어, 83

종 언어의 번역을 지원하고 있다. 이 중 중국어-영어 통역은 이미 뉴스, 학습, 일부 과학기술 영역에서 세계 최고 수준으로 알려져 있다.

바이두(Baidu)는 지난 7년간 AI에 적극적으로 투자해 현재 영어-중국어, 영어-독일어 자동통역 서비스를 지원하고 있다. 특히 'STACL(Simultaneous Translation with Integrated Anticipation and Controllable Latency)' 기술은 말을 끝내기도 전에 단어를 미리 예측하는 방법으로 실시간 자동통역을 가능케 했는데, 이는 두 언어 간 상이한 문장구조로 인해 실시간 자동통역이 사실상 불가능하다는 한계를 극복한 것이다.

한편 일본 보이스핑(VoicePing)은 회의 통역 상황에서 최대 45개 언어가 가능한 실시간 자막을 제공한다고 밝혔다. 보이스핑은 아시아 언어의 특성을 고려한 음성인식 및 번역 모델을 개발해 기술의 우수성을 강조하고 있다[3].

2.2 자동통역 기술 진화

자동통역 기술은 플랫폼 종류, 통역 범위, 사용성 등 크게 3가지 차별화 요소에 따라 <표 1>과 같이 1~3세대로 구분할 수 있다. 1세대 통역은 원격으로 언어소통이 가능한 전화망 기반 시스템이며, 호텔·항공기 예약 등 협소한 분야에서 정해진 문형으로만 통역이 가능했다. 2세대부턴 스마트폰이 대중화되면서 현지 대면 상황 통역 수요가 증가했고, 통역 범위도 몇백 문장 수준의 제한 영역에서 여행 영역 전체로 넓어졌다. 그러나 기술적 한계로 여전히 단문 수준을 벗어나지 못했다. 2세대 자동통역의 경우, 스마트폰 터치스크린을 통해 마이크 활성 버튼을 누르고 음성을 입력한 다음, 자동으로 번역된 결과를 화면으로 상대방에게 보여주거나 음성합성으로 들려주는 방식이 일반적이었다. 이 때문에 발화할 때마다 통역 시스템 제어를 위한 디바이스 터치 혹은 Wake-up Call을 통한 부자연스러운 인터페이스가 필요했다.

최근 딥러닝 기반 통역 성능이 대폭 향상됨에 따라 3세대부터는 강연, 회의 등 난이도가 매우 높은 자유발화 동시통역 수준으로 적용 범위가 확대되고 있다. 스마트폰 화면 터치로 통역하는 2세대 통역 방식이 사용성 한계에 다다른 시점에서, 3세대 핸즈프리 형태 자동통역으로 연구 개발이 진화하고 있는 것이다. 3세대 Non PTT(Push-to-Talk) 통역은 말하는 도중 시스템 제어 없이 대면 상황에서 자유발화로 실시간 통역이 가능하며, 이를 바탕으로 사용성이 획기적으로 개선된 시스템을 말한다.

<표 1> 자동통역 기술의 진화

구분	1세대(1990년대)	2세대(2010년대)	3세대(2020년대)
플랫폼	전화망	스마트폰	웨어러블
통역 범위	호텔 예약, 항공 예약 등 제한 영역	여행, 민원 등 일상 영역	강연, 회의 등 비즈니스 영역
사용성	정해진 패턴 통역	단문 수준 통역	자유발화 통역

출처: ETRI

최근 [그림 2]과 같이 LCD 또는 OLED 기반 투명 디스플레이가 포함된 고품질 소형 제품이 출시됐다. 이를 통해 외국인과 얼굴을 마주 보는 자연스러운 자동통역 인터페이스가 떠오르고 있다. 또한 향후 AR/VR 스마트 글라스 보급이 점점 늘어남에 따라, 웨어러블 기반 Non PTT 자동통역

도 곧 출현할 것으로 기대된다[4].

출처: ETRI

[그림 2] 투명디스플레이 기반 자연스러운 대화형 통역

3. 다국어 자동통역 핵심기술 현황

자동통역은 크게 음성인식, 기계번역, 음성합성 등 3가지 핵심기술로 구성된다. 이에 더해 통역 사용 환경이나 입력 모달리티 구성 여부, 딥러닝 AI 모델, 실시간 처리 여부, 다국어 번역, 전문 용어 처리처럼 실제 사용 현장에 적용하기 위한 추가적인 요소기술이 필요하다.

3.1 멀티모달 통역

지하철역, 길거리 등 주 변 소음이 큰 상 황에서 Non PTT 통역을 사용할 경우 음성 검출이 어려워 실패할 가능성이 높아진다. 이에 멀티모달 기반 음성발화 검출은 음성-영상 정보를 동시에 이용해 입술의 움직임을 동적으로 추적하고, 주변 잡음 환경에 강인하게 발화검출을 가능하게 한다.

최근 연구결과에 따르면, 여러 명의 음성이 섞여 있거나 매우 시끄러운 환경에서도 음성과 영상 정보를 이용해 타깃 음성을 정밀하게 추출할 수 있을 정도로 성능이 고도화되고 있어, 멀티모달이 고소음 환경에서 매우 효과적인 방안임을 알 수 있다. 다만 발화자의 얼굴을 계속 추적해야 하므로 여러 명이 대화 통역을 하는 상황에서 동시에 모든 사람을 추적할 순 없다. 때문에 보이스 필터 기술과 보완하는 방식으로 시스템이 구현돼야 할 것으로 보인다[5].

3.2 실시간 종단형 통역

일반적으로 자동통역은 음성인식, 자동번역, 음성합성, 자연어처리 모듈을 각각 학습해 각 모듈

의 입출력을 연결해 수행하는 다단계형(Cascade) 방식이 주류를 이루고 있다. 그런데 다단계형 통역은 최초 단계에서 오류가 발생할 경우, 오류가 단계별 모듈을 지나면서 증폭되는 효과가 있기 때문에 성능에 치명적인 영향을 미친다. 이를 극복하기 위해 최근 종단형(End-to-End) 통역이 제안되고 있다[6].

종단형 통역은 모국어 음성이 입력되면 단일 모델로부터 외국어 번역이나 합성음성이 직접 추론되도록 학습한다. 덕분에 기존 다단계 통역 방식의 오류전파 문제를 해결할 수 있다. 특히 모국어 음성에 포함된 음색이나 감정을 그대로 유지한 채 외국어로 변환할 수 있는데, 이를 바탕으로 인간을 모사하는 AI 서비스에 매우 중요한 기술이 될 수 있다. 다만 모국어 음성과 외국어 번역쌍 학습데이터를 대량으로 확보하기가 매우 어렵기에, 아직 실험실 수준에 머물고 있는 상태다.

한편 자동통역은 실시간성이 매우 중요하다. 따라서 음성입력 후 2~3초 내로 통역 결과가 나와야 하며, 이를 위해 종단형 구조에서 스트리밍 구조로의 변경이 필요하다. 또한, 통역 대상 언어 간 어순(예: 한국어와 영어의 어순 차이)이 다르기에, 통역 품질을 유지하면서 실시간 내 처리가 가능해야 한다. 따라서 통역을 구성하고 있는 음성인식, 번역, 합성 시스템이 시간에 따라 순차적으로 입력되는 데이터에 대해 즉시 출력 결과를 내는, 스트리밍 방식으로 구현돼야한다.

통상 음성인식은 스트리밍 구현이 일반적이나 자동번역이나 음성합성의 경우, 문장 전체가 입력됐을 때 처리가 시작되고 문장 단위로 출력을 내는 구조로 돼 있다. 따라서 실시간 자동통역이 가능하기 위해선 자동번역, 음성합성 시스템도 스트리밍 구조로 변경해야 한다. 각 시스템이 스트리밍 구조로 구현됐을 경우, 모국어 사용자가 발성한 시점부터 외국인 청자가 통역 결과를 듣기까지 시간적 지연이 생기는데, 이 지연시간을 어떻게 줄이느냐가 매우 중요하다. 또한 해당 지연시간을 줄임에 따라 전체 시스템 성능도 떨어지는데, 성능을 최대한 유지하면서 지연시간을 줄이는 기술이 매우 중요하다고 할 수 있다.

3.3 다국어 확장 및 언어 통합

다국어 자동통역 기술 개발에서 가장 어려운 부분은 AI 학습용 다국어 음성언어 및 번역 데이터 확보다. 전 세계 주요 언어인 영어·중국어·일본어는 확보가 다소 용이하나 동남아 언어·유럽어·아랍어·인도어 등 희소한 언어의 경우, 학습 데이터 확보가 매우 어렵다. 더구나 최근엔 언어별로 수만 시간 이상 음성과 수백만 규모의 대역 문장이 필요하다. 때문에 다국어로 확장하기엔 구축 비용이나 시간이 많이 필요하고, 대용량 데이터의 일관성 있는 품질 유지도 쉽지 않다.

최근 국내외에서 희소 언어에 대한 저용량 데이터 문제를 해결하기 위한, 학습 방법에 대한 연구가 활발히 진행되고 있다. ETRI 역시 다국어 확장을 용이하게 하기 위해 저용량 데이터를 가진 언어에 대해서도 일정 수준 성능을 확보할 수 있도록 노력하고 있다. 어원이 가까운 언어 간 단일 모델로 통합하거나 전사 레이블이 없는 음성 데이터로 학습하는 방법, 음성합성기를 이용한 인공음성 데이터 생성 등 다양한 기술을 도입하고 있다.

현재 서울지하철에 적용된 다국어 자동통역 서비스는 13개 언어별 음성인식 엔진과 24개 자동번역 엔진 등 총 36개 엔진이 하나의 서버에 가동 중이다. ETRI는 서울지하철 11개 역에서 동시에 끊김 없는 서비스를 제공하기 위해 각 엔진 당 평균 2~3개 동시접속이 가능하도록 운영하고

있다. 그런데 이와 같이 13개 언어에 대한 다국어 자동통역 서비스를 실시간으로 제공하는 것은 서버 부담을 매우 가중시킬 수 있다. 이와 같은 상황에서 백화점이나 호텔 등으로 서비스가 확대하기 위해선 시스템을 효율적으로 운용하는 방안이 필요하다. 개별 언어를 통합해 하나의 AI 모델로 여러 개 언어를 인식·번역하는 통합 AI 모델 개발의 필요성이 커지고 있는 것이다.
번역의 경우, 언어를 모두 통합해 하나의 모델로 번역해도 큰 문제는 없을 것으로 예상된다. 다만 음성인식의 경우, 개별 언어를 하나의 모델로 통합했을 때 음성인식률 저하가 없도록 해야 하며, 말하는 언어가 무슨 언어인지 혼돈되지 않도록 신뢰성을 보장하는 것이 매우 중요하다. 실제 현장에선 다양한 소음이 발생하며 소음 레벨도 높아 언어 식별에 문제가 일어날 가능성이 높다. 이에 다국어 음성인식 AI 통합 모델에 대해, 소음에 대한 강인성을 높이는 연구도 진행 되고 있다.

3.4 구어체 번역 및 최적화

자동번역의 경우, 외국인과 대화형으로 자연스럽게 언어소통이 가능하도록 구어체 문장에 대한 번역이 가능해야 한다. 구어체 번역은 문어체와 달리 의역이 많기 때문에 직역을 하게 되면 의미 전달에 오류가 발생한다. 이에 구어체 번역은 두 사람이 서로 주고받는 대화 문장을 위주로 학습하며, 의역이 가능하도록 문맥을 가급적 길게 볼 수 있도록 학습이 필요하다.
그리고 아시아권 언어나 유럽어 등 언어 간 특성이 다르기에, AI 모델의 파라미터 설정도 달리해야 한다. 이를 위해 번역 시스템을 학습 단계와 추론 단계로 나눠 '주요 파라미터들이 개별 언어별 성능에 어떻게 영향을 미치는지'가 연구되고 있다. 학습 단계에선 어휘 통합 여부, 어휘 크기, 임베딩 가중치 공유(Embedding weight sharing) 등의 파라미터를 바꿔가며 성능에 미치는 영향을 분석하고, 추론 단계에서는 빔 크기(Beam size), 출력 길이 비율(Output length ratio) 등을 다양하게 변화시키고 있다. ETRI는 현재 한국어를 중심으로 영어, 일본어, 중국어로 구성된 총 200만 문장 데이터셋을 활용, 다양한 파라미터 값 변화에 따른 성능 변화를 체계적으로 분석하고 있다.

3.5 전문용어 통역

외국인이 한국을 방문하는 경우, 한국어 지명, 지하철 역명 등 고유명사가 포함된 질문을 많이 하게 된다. 이런 경우, 한국어 고유명사에 대한 발음이 부정확하고 음운 변이 또한 다양해 음성인식 오류가 일어날 가능성이 높아진다.
일반적으로 지명, 사람 이름, 숫자, 전문용어 등은통상 모국어 음성인식이라 하더라도 오류율이 높은데, 외국인이 발성하는 고유명사나 전문용어는 더욱 어렵다. 특히 외국어 문장 내 한국어 고유명사가 포함되는 경우가 극히 드물어 외국어 음성인식용 모델은 문맥 측면에서도 한국어 고유명사 음성인식에 매우 취약하다. 외국인을 대상으로 튜닝용 데이터를 수집한다고 해도 외국인이 한국어 고유명사를 발성하기 힘들어 한다. 또한 외국인 수십 명을 대상으로 여러 언어에 대한 데이터를 도메인마다 수집하는 것도 사실상 어려워 해결가능한 방안이 아직 명확하지 않다.
이에 외국인이 발성한 문장 내 포함된 한국어 고유명사를 좀 더 잘 인식할 수 있도록, 발음 변이 현상에 대한 음향적 분석과 함께 한국어 고유명사가 포함된 문맥을 AI 모델에 강화하는 연구

가 필요하다. 공학, 언어학, 음성학, 외국어 교육학 등 다학제 간 연구를 통한 해결이 가장 바람직해 보인다.

4. 맺음말

최근 AI 학습 능력이 비약적으로 발전함에 따라 20여 년간 정체돼 왔던 자동통역의 사용성이 대폭 향상됐고, 일상생활 언어장벽 해소에서부터 비즈니스 동시통역 등 파급효과가 큰 영역으로 적용이 점차 확산되고 있다. 자동통역을 위한 양방향 인터페이스도 스마트폰에서 투명 디스플레이, XR/AR 글라스까지 다양해지고 있다. 통역 가능한 언어의 경우 100여 개에 가까운 다국어 서비스까지 지원이 가능해지고 있어, 이제 자동통역은 일상생활이나 업무에서 없어서는 안 될 매우 중요한 기능으로 자리잡고 있다[5].

최근엔 외국인 근로자들의 대량 입국으로 국내 체류 외국인이 증가하며 민원이나 범죄 등이 증가하고 있다. 이에 신속하고 정확하게 해결해야 할 사건 현장이 늘어나고, 검·경찰청, 법원에선 동시통역사가 부족해 피의자 인권침해 등이 우려되고 있다. 법률적 언 어소통을 제공하는 동시통역사도 대부분 서울지역에분포하거나, 법률 관련 통역엔 한계가 있어 AI 기반 사법통역이 동시통역사를 지원하는 기능이 절실히 요구되고 있다.

특히 외국인 범죄자 대부분이 희소 언어를 사용하므로 지금부터라도 지속적인 학습데이터를 구축하고, 지역 현장에서 통역기술 실증을 통한 문제 보완에 나서야 한다. 더불어 동시통역사 지원 장치로 통역기를 허용하는 제도를 마련하는 데 정부와 국회의 지원도 필요한 시점이다.

[참고문헌]

[1] 김상훈, "Conversational AI 기반 자동통역 기술 동향", IITP 주간기술동향 1982호, 2021.

[2] 조선일보, "AI가 자동통역...명동역 가면 13개국어 통한다", 2023.

[3] 김상훈, "언어장벽 없는 세상 실현을 위한 AI 자동통역 발전 동향", KISDI AI TREND WATCH, 2021-20호, 2021..

[4] Philipp A. Rauschnabel, et al., "What is XR? Towards a Framework for Augmented and Virtual Reality", Computers in Human Behavior, Aug. 2022.

[5] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," ICLR 2022.

[6] Jeong-Uk Bang, Min-Kyu Lee, Seung Yun, Sang-Hun Kim, "Improving End-To-End Speech Translation Model with Bert-Based Contextual Information," ICASSP 2022.

※ 출처: TTA 저널 제214호