정비 교육 및 훈련을 지원하는 AI 디지털 튜터용 음성인식 모듈 구현

임정현, 노병희*, 강경민
*아주대학교 AI융합네트워크학과, ㈜엠아르오디펜스 AI/IoT 연구개발팀
{wjdguszoqt, *bhroh}@ajou.ac.kr, kmkang@mrodefence.com

Implementation of Voice Recognition Module for AI Digital Tutors to Support Maintenance Education and Training

Junghyun Lim, Byeong-hee Roh*, Kyeong-min Kang
*Dept. of AI Convergence Network, Ajou University, MRODefence Co.Ltd.

요 약

본 논문은 장비 정비와 같은 복잡한 산업 훈련 환경에서 필수적인 핸즈프리(Hands-free) 사용자 인터페이스 지원을 위한 음성인식 모듈을 제안하고 그 성능을 비교 분석한다. 현대의 고정확도 음성-텍스트 변환(Speech-to-Text, STT) 모델은 연산량이 많아 자원이 제한된 엣지 디바이스 장치에서 실시간으로 구동하기 어려운 기술적 과제를 안고 있다. 이러한 문제를 해결하기 위해, 본 연구에서는 OpenAI의 Whisper 모델을 고속 추론 엔진 CTranslate2 기반으로 재구현한 Faster Whisper를 사용하여 온디바이스(on-device) 음성인식 파이프라인을 설계 및 구현하였다. 제안된 모듈을 범용 싱글 보드 컴퓨터(Single-Board Computer, SBC)인 Raspberry Pi 5와 AI 가속에 특화된 NVIDIA Jetson Orin Nano에서 실험한 결과, 모든 하드웨어에서 0.1초 미만의 평균 추론 속도를 보이며, 실시간에 가까운 음성 처리가 가능함을 확인하였다. 이는 제안된 모듈을 엣지 디바이스와 함께 사용하는 경우 최신 STT 모델을 현장에 효과적으로 배포할 수 있음을 입증하며, 실용적인 AI 기반 디지털 튜터 시스템 개발이 가능함을 나타낸다.

Ⅰ. 서론

4차 산업혁명 시대가 도래함에 따라 산업 현장의 기계 및 장비는 점차 고도화되고 있으며, 이에 따라 정비 및 운영 인력을 위한 효과적인 교육 훈련 솔루션의 필요성이 증대되고 있다. 특히, 정비 작업과 같이 작업자의 양손이 자유롭지 못한 환경에서는 음성 기반의 지능형 사용자 인터페이스가 단순한 편의 기능을 넘어 안전과 작업 효율성을 위한 핵심 요구사항으로 부상하고 있다. AI 디지털 튜터는 이러한 요구에 부응하여 작업자에게 실시간으로 절차를 안내하고 상호작용하는 차세대 훈련 시스템으로 주목받고 있다.

최근 OpenAI의 Whisper와 같은 대규모 트랜스포머 기반의 자동 음성 인식(Automatic Speech Recognition, ASR) 모델은 방대한 데이터셋(약 68만 시간) 학습을 통해 기존 모델을 뛰어넘는 정확도와 잡음 환경에 대한 강인성을 보여주었다 [1]. 이러한 특성은 소음이 많은 정비 현장 환경에 매우 적합하다. 그러나 이 모델들은 막대한 연산량을 요구하기 때문에, 저전력, 저비용의 일반적인 엣지 디바이스에서는 실시간 처리가 거의 불가능하다는 본질적인 한계를 가진다.

본 논문은 최적화된 추론 모델과 엣지 디바이스를 결합함으로써 이러한 성능 격차를 해소할 수 있다는 가설을 검증하고자 한다. 본 연구의 기여는 다음과 같다.

- AI 디지털 튜터를 위한 독립형 온디바이스 STT 처리 파이프라인의 설계 및 구현
- 범용 SBC인 Raspberry Pi 5와 AI 특화 디바이스 NVIDIA Jetson Orin Nano에서의 구현 모듈의 성능 비교 분석
- 실험 결과를 바탕으로 실제 산업 환경에서의 활용 가능성을 검증, 유사 응용 분야에서의 하드웨어 선정 가이드를 제공

Ⅱ. 관련 연구

초기 음성인식 시스템은 Kaldi 기반의 은닉 마르코프 모델(HMM)과 심층 신경망(DNN)을 결합한 하이브리드 방식이 주를 이루었으나, 최근에는 전체 시스템을 단일 신경망으로 구성하는 종단간(End-to-End) 모델이 대세로 자리 잡았다 [2]. 대표적으로 Meta의 Wav2Vec2는 레이블이 없는 대규모 음성 데이터로 사전 학습하는 자기지도학습(Self-supervised learning) 방식을 통해 높은 성능을 달성했다. OpenAI의 Whisper는 여기서 한 걸음 더 나아가, 웹에서 수집한 68만 시간 이상의 다국어 및 다중 도메인 음성 데이터를 활용하여 별도의 파인튜닝 없이도 다양한 환경에서 높은 정확도를 보이는 제로샷(zero-shot) 성능을 입증하였다.

Whisper와 같은 대형 모델을 엣지 디바이스에 배포하기 위해서는 반드시 모델 최적화 과정이 필요하다. 본 연구에서 채택한 Faster Whisper는 Whisper 모델 아키텍처를 CTranslate2 추론 엔진 위에서 재구현한 것이다. CTranslate2는 Layer Fusion, Padding 제거, INT8 양자화(Quantization)와 같은 기법을 통해 CPU와 GPU 환경 모두에서 추론 속도를 극대화하는 고성능 엔진이다. 이러한 최적화를 통해 Faster Whisper는 원본 Whisper 대비 상당한 속도 향상을 이루어, 엣지 환경에서의 실시간성에 대한 가능성을 열었다.

엣지 AI를 구현하는 하드웨어는 크게 두 가지 접근 방식으로 나뉜다. 첫째는 Raspberry Pi 시리즈의 ARM Cortex 계열 CPU와 같이 범용 컴퓨팅 성능을 강화하는 방식이다. 둘째는 NVIDIA Jetson 시리즈처럼 CUDA 코어나 텐서 코어(Tensor Core)와 같은 신경망 처리 가속기(NPU 또는 GPU)를 내장하여 AI 연산에 특화된 성능을 제공하는 방식이다. GPU를 내장한 하드웨어의 성능이 더 뛰어나지만, Jetson 시리즈는 Raspberry Pi보다 가격이 비싸다는 단점이 있어 가용 비용 또한 고려할 필요가 있다.

본 연구는 2가지 하드웨어 모두에서 실사용이 가능한 수준으로 모듈을 구현하고 성능을 비교 분석하여, 실제 산업 환경에서의 하드웨어 선정 가이드를 제공한다.

Ⅲ. 음성인식 모듈 설계 및 구현

본 연구에서 개발한 음성인식 모듈은 실시간 음성 입력을 받아 텍스트로 변환하기까지 일련의 처리 단계를 파이프라인 구조로 설계하였다. 전체 시스템은 네트워크 연결 없이 장치 내에서 독립적으로 동작한다. 시스템의 전체 구조는 그림 1과 같다.

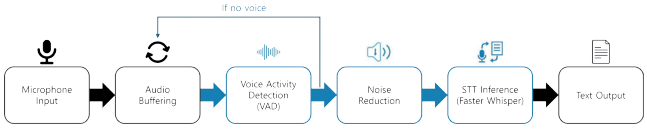

그림 1. 음성인식 모듈 처리 파이프라인

제안 시스템에서 핵심 구성 요소의 역할은 다음과 같다.

- 음성 활동 감지 (Voice Activity Detection): 입력된 오디오 신호에서 사람의 음성이 포함된 구간만을 검출한다. VAD는 시스템의 효율성을 극대화하는 핵심 요소로 침묵이나 배경 소음 구간에 대해 STT 추론을 수행하지 않도록 하여 전체적인 연산 부하를 크게 줄이고 실시간성을 확보한다.
- 잡음 제거 (Noise Reduction): 정비 현장과 같은 소음 환경에서 인식 정확도를 높이기 위해 배경 잡음을 제거한다. 이 단계는 STT 모델에 입력되는 오디오의 신호 대 잡음비(Signal-to-Noise Ratio, SNR)를 개선하여 모델의 성능을 안정적으로 유지하는 데 기여한다.
- 음성 텍스트 변환 (STT Inference): 모듈의 핵심부로, VAD를 통해 검출되고 잡음이 제거된 음성 데이터를 텍스트로 변환한다. 본 연구에서는 성능과 자원 사용량의 균형을 고려하여 base 모델을 사용하였으며, CTranslate2 엔진을 통해 추론을 수행한다. 모든 과정이 오프라인으로 처리되므로 네트워크 연결이 불안정하거나 불가능한 산업 현장에서도 안정적인 사용이 가능하다 [3].

Ⅳ. 실험 및 결과 분석

1. 실험 환경

제안된 음성인식 모듈의 성능을 객관적으로 비교 분석하기 위해, 사양이 다른 세 종류의 하드웨어 플랫폼에서 실험을 진행하였다. 각 플랫폼의 주요 사양은 표 1과 같다.

플랫폼	CPU	GPU/AI 가속기	RAM
노트북	AMD Ryzen 7 5700U	내장 그래픽	16GB DDR4
Raspberry Pi5	4-core ARM Cortex-A76	-	8GB LPDDR4X
Jetson Orin Nano	6-core ARM Cortex-A78AE	1,024 CUDA + 32 Tensor Cores	8GB LPDDR5

표 1 하드웨어 플랫폼 사양

소프트웨어 환경은 Ubuntu 22.04 운영체제와 Python 3.10을 기반으로 구성하였으며, Faster Whisper는 base 모델과 16비트 부동소수점(float16) 연산 타입을 사용하였다. 실험은 30자 이내의 간단한 문장 4개를 각 플랫폼에 입력 후 문장별 추론 시간을 결과 창에 출력하여 진행하였다.

2. 실험 결과

플랫폼	최솟값(s)	평균값(s)	최댓값(s)
노트북	0.0028	0.0040	0.0056
Jetson Orin Nano	0.0103	0.0193	0.0424
Raspberry Pi5	0.0816	0.0868	0.0929

표 2 하드웨어 플랫폼별 STT 모듈 추론 성능 비교

실험 결과, 모든 플랫폼에서 0.1초 미만의 추론 시간을 기록했다. 또한 UI의 갱신 속도, I/O 버퍼, 쓰레드/큐 동기화 지연 등 CPU/RAM/OS 레벨의 병목 요소까지 고려할 경우, 추론 결과가 화면에 출력되는 평균 시간은 노트북, Jetson Orin Nano, Raspberry Pi5에서 각각 0.72, 2.16, 2.82초로, 모든 플랫폼에서 실사용이 가능한 수준임을 보였다.

또한, 실험 결과를 통해 실제 모듈이 사용될 산업 환경에서는 별도의 UI 개발 여부, 외부 출력 장치 유무, 가용 비용 등을 함께 고려하여 적합한 하드웨어 플랫폼을 선택할 수 있을 것으로 판단한다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 장비 정비 교육용 AI 디지털 튜터를 위한 엣지 디바이스용 음성인식 모듈을 구현하고, 서로 다른 아키텍처를 가진 하드웨어 플랫폼에서 성능을 비교 분석하였다. 실험을 통해 Faster Whisper을 엣지 디바이스에서 실시간으로 구동하는 경우 실사용이 가능한 수준임을 보였다. 이는 복잡한 산업 현장에서 활용 가능한 핸즈프리 AI 디지털 튜터 시스템 개발을 위한 실질적인 청사진을 제공한다.

향후 연구는 현재 float16 정밀도로 동작하는 모델을 GPTQ, AWQ와 같은 후훈련 양자화(Post-Training Quantization, PTQ) 기법을 적용하여 8비트 정수(INT8) 모델로 변환하여 속도 및 메모리 점유율 측면에서 성능을 향상시키고, 정비 매뉴얼, 기술 문서 등의 텍스트 데이터를 기반으로 LoRA(Low-Rank Adaptation)와 같은 파라미터 효율적 파인튜닝(PEFT) 기법을 적용하여 모델을 특정 도메인에 적응시킬 것이다. 또한, LLM과 TTS를 활용하여 도메인 특화 합성 음성 데이터를 생성하고 이를 학습에 활용하는 방안도 탐색할 것이다. 이를 통해 목표 도메인에서의 단어 오류율(Word Error Rate, WER)을 획기적으로 낮춰 AI 디지털 튜터의 신뢰성과 실용성을 극대화하고자 한다.

ACKNOWLEDGMENT

이 논문은 중소벤처기업부와 중소기업기술정보진흥원의 지원을 받아 수행된 연구임. (과제번호: S3433427)

참 고 문 헌

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Proceedings of the 40th International Conference on Machine Learning, pp. 28492-28518, 2023.

[2] Santosh Gondi, Vineel Pratap, “Performance Evaluation of Offline Speech Recognition on Edge Devices,” Electronics, vol. 10, no. 21, pp. 2697, 2021.

[3] Jyotika Singh, “pyAudioProcessing: Audio Processing, Feature Extraction, and Machine Learning Modeling,” Proceedings of Python in Science Conference, pp. 152-158, 2022.