소음 환경을 고려한 지속 학습 시스템 기반 음성 핵심어 검출 시스템 구현

신채림, 김태구, 조용훈, 신기훈, 정혜선, 곽도균, 구동한, 나상진, 백윤주*
부산대학교
{cofla0429, tbg8577, kchoyh95, skh2929209, ahttjd55, kwakdg, ehdgksrn, sktkdwls1222}@pusan.ac.kr,
*yunju@pusan.ac.kr

Implement of a Keyword Spotting System Based on Continual Learning System Considering the Noise Environment

Shin Chae Rim, Kim Tae Gu, Cho Yong Hun, Shin Ki Hun, Jeong Hye Sun, Kwak Do Gyun,
Koo Dong Han, Na Sang Jin, Baek Yun Ju*
Pusan National Univ.

요 약

AI를 활용한 음성 인식 기술의 지속적인 발전으로 스마트폰, 스마트 스피커, 의료 분야 등 다양한 산업 분야에서 적극적으로 도입되고 있다. 하지만 산업 현장에서 발생하는 극심한 노이즈에 강건한 인식 성능을 보이는 실시간 음성 인식 기술은 여전히 미비한 상황이다. 긴급한 상황에 사용되는 음성 인식 기술의 경우 오인식을 개선하는게 매우 중요하다. 오인식 발생 시 작업자의 안전과 생산성에 큰 영향을 미칠 수 있으며, 음성 인식 시스템의 신뢰성이 떨어질 수 있다. 본 논문에서는 산업 현장에서 긴급 상황에서 장비 제어나 긴급 대응과 같은 용도로 활용할 수 있는 음성 인식 모델을 구현하고 성능을 평가한다. 학습 데이터에 배경 소음을 추가하고, 키워드와 비키워드를 구분할 수 있도록 음성학적으로 유사한 단어를 함께 학습하는 것으로 모델의 오인식률을 감소시킬 수 있다. 제안하는 시스템은 최대 90%의 오인식률 감소를 나타낸다.

Ⅰ. 서 론

최근 딥러닝 기반 음성인식 기술의 발전에 따라 대부분의 컴퓨터와 스마트 기기에 음성 어시스턴트가 활용되고 있다. 딥러닝 기반의 음성인식 기술인 자동 음성 인식(Automatic Speech Recognition, ASR) 모델은 대규모 데이터 세트를 요구하며, 모델의 크기가 매우 크기 때문에 높은 메모리 대역폭과 연산 능력을 필요로 한다. 때문에 일반적인 ASR 모델은 고성능 디바이스가 필요하다는 단점이 있다. 이러한 한계를 극복하기 위해 저성능 디바이스에서 활용할 수 있는 경량 딥러닝 기반 음성인식 기술에 대한 연구가 활발하다. 음성 핵심어 검출(Keyword Spotting, KWS) 시스템은 음성 신호 중 학습한 키워드만을 인식하는 기술로 모델의 크기와 복잡성이 낮아 성능이 제약적인 디바이스에서 활용할 수 있다.

산업 현장에서 간단한 키워드 인식을 통해 긴급한 상황을 식별하고 기계의 동작을 중지시키는 것은 음성 핵심어 검출 기술의 중요한 응용 분야 중 하나이다. 산업 현장의 소음에 대응하기 위해 다양한 소음 환경에 강건한 인식 성능을 보이는 음성 핵심어 검출 모델이 요구된다. 이를 위해 소음이 있는 환경에서의 발화를 사용하여 모델을 훈련하는 다중 조건 훈련에 대한 연구가 진행되고 있다[1]. 하지만 이는 좋은 성능을 달성하기 위해 큰 모델을 요구한다는 한계가 존재한다. 음성 핵심어 검출 시스템의 클래스 불균형 문제를 해결하여 모델의 잘못된 예측을 줄이고, 정확한 데이터를 학습하는 negative data mining 연구도 활발하다[2]. 하지만 일반적인 실생활 환경 대비 극심한 소음이 발생하는 산업 현장을 대상으로 강건한 인식 성능을 보이는 모델 개발에 대한 연구는 미비하며, 실제 음성인식 디바이스에 탑재해 실시간 인식을 구현한 사례 또한 미비하다.

본 논문에서는 주변 소음에 강건하고 정확도와 오인식률을 개선한 경량 딥러닝 기반 음성 핵심어 검출 시스템을 제안한다. 다양한 환경에서 수집한 테스트 데이터를 사용하여 개발된 모델의 오인식률을 측정하고, 오인식된 음성 파일을 기존 데이터 세트에 추가하여 모델을 지속적으로 재학습시킨다. 이를 통해 학습된 핵심어 검출 모델을 자원 제약적인 음성인식 디바이스에 탑재하여 극심한 소음이 발생하는 산업 현장의 긴급한 상황에서 동작할 수 있는 음성 핵심어 검출 시스템을 구현한다.

Ⅱ. 본론

1. 음성 데이터 세트 구성 및 전처리

긴급한 상황임을 전달하고, 키워드를 통해 장비를 제어할 수 있는 '긴급', '중지'라는 두 키워드를 선정하였다. 인식 정확도 향상을 위해 긴급한 상황(E)과 일상적인 상황(N)에서 발화로 데이터를 구분하여 수집하였다. 수집한 키워드에 대해 '긴급E', '긴급N', '중지E', '중지N'으로 라벨링하였으며, 총 참가자 100명으로부터 각 키워드에 대해 약 10회 반복 녹음하여 1초 단위의 wav 파일을 총 96,000개 수집하였다. 모든 데이터에는 실제 산업 현장에서 수집한 소음과 실생활에서 발생하는 소음 데이터를 합성하여 학습 및 평가에 활용하였다. 소음은 웃음소리, 울음소리, 백색잡음과 같은 기본적인 생활 소음과 공장 배경 소음, 기계음 등의 산업 현장 소음을 활용하였다. 정확도 향상을 위해 대용량의 음성 데이터 세트가 필요하지만, 자체적으로 데이터 세트를 구축하는 방식은 한계가 있다. 따라서 다양한 형태의 음성을 학습하기 위해 수집된 음성 데이터에 time shifting, changing pitch, changing speed, volume changing 증강기법을 사용한다[3]. 음성 신호의 시계열 정보를 보존하며 주파수 특성을 검출하기 위해 Short time fourier transform과 hann-window를 적용하여 Mel-Frequency Cepstral Coefficient(MFCC)를 추출하여 음성 딥러닝 모델의 입력으로 활용한다.

학습한 모델에 대한 오인식률을 평가하기 위해 AI-Hub 공개 데이터 세트에서 사전에 정의된 키워드가 포함되지 않은 3개의 데이터 세트를 선정하였다. 표 1에 나타낸 각 데이터 세트는 약 260,000개의 1초 단위의 wav 파일을 포함한다.

표 1. 수집한 테스트 데이터 세트 (AI-Hub)

데이터 세트	데이터 세트명
1	저음질 전화망 음성인식 데이터
2	방송 콘텐츠 대화체 음성인식 데이터
3	문학작품 낭송.낭독 음성 데이터

2. 음성 핵심어 검출 모델

자원 제약적인 환경에서 음성 핵심어 검출 모델을 탑재하기 위해 Depthwise Separable Convolution(DSC)를 적용하여 모델의 연산량과 파라미터 수를 줄인다. Convolution 연산마다 배치 정규화(batch normalization)와 Rectified Linear Unit(ReLU) 연산을 수행하며, Gloabal Average Pooling(GAP)과 Softmax 연산을 통해 키워드를 인식한다.

오인식 감소를 위한 지속 학습 시스템의 전체적인 구조는 그림 1과 같다. 사전에 정의된 키워드를 포함하는 음성 데이터와 함께 모델의 학습 정확도를 높이기 위해 음성학적으로 비슷하고, 자주 쓰이는 단어를 포함하는 네거티브 데이터 세트를 입력으로 사용한다. 오인식된 음성 파일을 테스트 데이터 세트에서 추출하여 네거티브 데이터 세트를 업데이트한다. 이를 통해 지속적으로 데이터 세트를 갱신하면서 모델의 지속적인 학습을 진행한다. 구현된 모델은 Raspberry pi 3에 탑재한다.

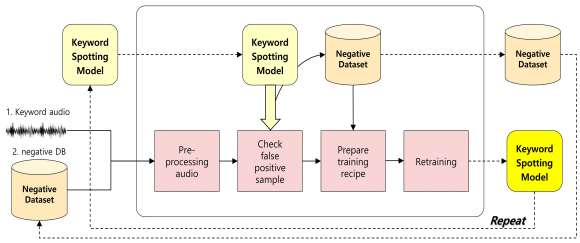

그림 1. 지속 학습 시스템 구조

Ⅲ. 실험

제안하는 시스템의 평가 지표는 모델의 인식 정확도, 파라미터 개수 및 용량, 추론 시간으로 선정하였다. 표 2는 평가 지표에 따른 음성 핵심어 검출 모델의 실험 결과를 나타낸다. 개발한 음성 핵심어 검출 모델의 정확도는 99.75%로 매우 정밀하게 소음 환경에서 키워드를 인식할 수 있다. 또한, DSC 모델을 적용함으로써 모델의 파라미터 수와 메모리 용량이 자원 제약적인 임베디드 보드에 탑재하기에 적합함을 확인할 수 있다. 음성 인식 디바이스에서 추론 시간을 줄이기 위해 음성 신호를 짧은 프레임으로 나누어 각 프레임에서 특징을 추출하고, 연속적인 음성 데이터를 모델의 입력으로 사용하기 위해 sliding window를 적용하여 추론 시간 0.46초를 달성하였다.

표 2. 음성 핵심어 검출 모델의 성능

정확도 (%)	파라미터 수	메모리 (KB)	추론시간 (s)
99.75	38,758	151.40	0.46

각 테스트 데이터 세트별로 재학습 빈도에 따른 오인식률의 변화를 비교 및 분석한다. 베이스 모델에서 1회와 5회 재학습 시 오인식률이 가장 낮게 나타났으며, 그림 2에서 확인할 수 있다. 이후 오인식률이 지속적으로 증가하는 경향을 보이다가 5회에서 다시 감소한다. 표 2는 베이스 모델의 재학습 빈도를 0으로 설정하여 1회와 5회 재학습 시 오인식률을 비교한 결과이다. 데이터 세트의 특징에 따라 오인식률이 감소 되는 비율 또한 달라진다. 배경 소음이 포함된 데이터 세트 1과 2에서 재학습 시 베이스 모델 대비 오인식률이 각각 91%, 82%로 크게 감소하는 반면, 소음이 없는 환경에서 수집된 데이터 세트 3은 상대적으로 오인식률이 높고, 재학습 시 감소율이 20%로 타 데이터 세트 대비 미미하다. 이는 소음이 많은 환경에서 수집된 테스트 데이터 세트를 재학습하는 과정에서 실제 환경의 음성 데이터의 특성을 더 정확히 파악하고, 키워드를 정확하게 인식할 수 있기 때문이다. 하지만 재학습의 횟수가 증가할수록 시간이 많이 소요되며, 과적합이 발생할 수 있다. 따라서 재학습 빈도는 1회가 가장 적절하며, 재학습 시 사용될 데이터 세트는 소음이 포함된 환경에서 수집된 것이 적합하다.

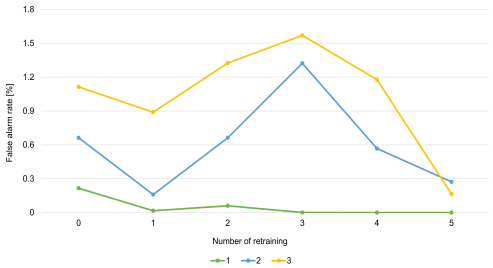

그림 2. 지속 학습 횟수에 따른 오인식률

표 3. 테스트 데이터 세트에 따른 오인식률 비교

데이터 세트	1			2			3		
재학습 빈도	0	1	5	0	1	5	0	1	5
오인식률 (%)	0.23	0.02	0	0.66	0.12	0.27	1.11	0.89	0.16

Ⅳ. 결론

본 논문은 극심한 소음이 발생하는 산업 현장의 긴급 상황에서 동작하는 낮은 오인식률의 음성 핵심어 검출 시스템을 구현하고, 이를 자원 제약적인 음성인식 디바이스에 탑재하였다. 경량화를 위해 DSC 모델을 이용하였으며, 구현한 모델은 99.75%의 높은 정확도를 나타내었다. 오인식률을 최소화하기 위해 다양한 음성 데이터를 활용하여 지속적인 네거티브 데이터 세트 업데이트 및 재학습을 진행하였다. 실험 결과, 재학습의 빈도는 1회가 적합함을 확인하였으며, 학습 데이터 세트와 유사한 데이터 세트로 재학습을 진행했을 때 오인식률이 최대 90% 감소하는 것을 확인하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2024-RS-2023-00260098)

이 연구는 2024년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(20015052)

참 고 문 헌

[1] Gu, Y., Du, Z., Zhang, H., and Zhang, X., "A monaural speech enhancement method for robust small-footprint keyword spotting," arXiv prepring arXiv:1906.08415, Jun. 2019

[2] J. Hou, Y. Shi, M. Ostendorf, M. -Y. Hwang and L. Xie, "Mining Effective Negative Training Samples for Keyword spotting," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 129-132, May. 2020.

[3] S. Daniel, C. william, "SpecAugment. A simple Data Augmentation Method for Automatic Speech Recognition," arXiv preprint arXiv:1904.08779, Apr. 2019