공학전문석사학위 연구보고서

음향 기반의 조립 결함 감지 시스템

Acoustic-based assembly defect detection system

2022년 2월

서울대학교 공학전문대학원
응용공학과 응용공학전공
이 한 수

음향 기반의 조립 결함 감지 시스템

Acoustic-based assembly defect detection system

지도 교수 성 우 제

이 프로젝트 리포트를 공학전문석사 학위
연구보고서로 제출함
2022년 2월

서울대학교 공학전문대학원
응용공학과 응용공학전공
이 한 수

이 한 수의 공학전문석사 학위 연구보고서를
인준함
2022년 2월

위원장 오 병 수 (인)
위 원 성 우 제 (인)
위 원 김 성 우 (인)

초 록

글로벌 제조업체들은 제품의 품질을 확보하기 위한 많은 노력에도 불구하고, 다양한 원인에 의한 생산 불량은 지속해서 발생한다. 생산 불량 제품이 소비자에게 전달될 경우 이것은 직접적인 비용 이외에도 브랜드 이미지를 일순간에 실추 시켜 기업경영에 타격을 줄 수 있는 중요한 문제이다.

최근 딥러닝 기술의 발전으로 제조 현장에서 이상 탐지(anomaly detection)기반의 많은 연구가 이루어지고 있다. 그러나 최근 연구된 관련 연구를 살펴보면 주로 비전(vision)검사를 통한 제품의 표면 결함을 다루거나 진동 데이터를 이용해 생산 설비의 상태 검사에 관한 연구가 대부분이다. 그러나 이러한 방식은 소리를 다루는 오디오 제품에 적용하기에는 적합하지 않다. 또한 하루에 수백, 수천 개를 생산하는 제조업에서 제품의 품질검사에 센서를 부착하기 위해서는 많은 시간과 비용이 소요된다.

본 연구는 단일 음향 센서로 측정한 스피커 출력데이터와 합성 곱 신경망을 활용하여 원시 오디오 신호에서 직접 표현을 학습하는 소리 분류에 대한 종단 간 접근 방법을 제시한다. 조립 결함의 분류 작업과 관련된 다양한 필터를 학습하기 위해 7개의 컨볼루션 레이어가 사용된다. 데이터 셋은 여러 대의 스피커에서 출력데이터를 통해 수집되었으며, 일부 스피커 출력에서 학습한 지식이 다른 스피커 결함 여부를 판단 할 수 있음을 보였고, 평균 정확도 99%를 달성하는 것으로 나타났다. 제안하는 조립결함 감지 기법은 기존의 2D 표현을 입력으로 사용하는 대부분의 방식보다 높은 성능을 보인다. 또한, 다른 아키텍처에 비해 적은 수의 매개변수를 가지고 있어, 실시간 제품 품질 검사에 효율적이다.

본 연구를 통해 제조 현장에서 TV, 차량용 AVN과 같이 스피커가 탑재된 제품에 대한 조립 공정 불량률을 감소시켜줄 것으로 기대한다.

주요어 : 딥러닝, 이상탐지, 불량 검출, 합성곱 신경망

학번 : 2020-21071

목 차

그 림 목 차

표 목 차

제 1 장

서론

1.1 연구 배경

제조업체들은 글로벌 시장에서 경쟁력을 키워나가기 위해 인건비, 공급 비용, 간접비, 재료비 등 비용 절감을 위한 많은 노력을 하였다. 그러나 이러한 방식은 가격 경쟁력은 확보하였으나 반대로 품질에 대한 문제점들을 노출하였다. 일례로 도요타 리콜 사태를 들 수 있다.

당시 도요타는 글로벌 경쟁에서 생존하고자 전사적인 원가절감에 사활을 걸었다. 하지만 그에 대한 반대급부로 설계부터 조립까지 이르는 모든 과정에서 품질관리를 소홀히 하게 되었고, 그 결과로 2009년 도요타 자동차는 가속 페달 문제를 시작으로 결함 부위와 대상 차종이 점차 확대됨에 따라 창사 최악의 리콜 사태를 맞이하게 되었다. 2010년 2월 전 세계에서 리콜 및 수리 대상이 도요타의 2009년 일본판매 대수에 필적하게 되었다 [1]. 직접적인 리콜 및 수리 비용 이외에도 기업이 오랜 세월 동안 쌓아온 브랜드 이미지를 일순간에 무너뜨리고 기업을 존폐에 위기에 빠지게 만든 대표적인 사례이다.

오늘날 이와 같은 기업들의 행보는 변화하고 있다. 미래 지향적인 기업은 품질에 높은 가치를 부여하고 성장과 수익성을 대체할 수 있는 경로를 모색하였다. 새로운 프로세스와 기술의 개발로 기업은 품질과 수익 창출에 중점을 두고 자동화되고 연결된 공장을 운영한다. 그러나 제조업의 생산 과정에서 품질을 확보하기 위한 위와 같은 노력에도 불구하고,

다양한 원인(부품결함 및 누락, 조립 불량 등)으로 인하여 불량제품은 지속해서 발생한다. 대부분의 불량의 원인인 휴먼 에러는 제조 공정의 자동화에 따라 발생 빈도가 현저하게 줄어들었으나, 기계로 대체할 수 없는 공정 설비는 여전히 존재한다.

필연적으로 발생하는 불량을 검출할 목적으로 컴퓨터 비전을 활용한 검출기가 제조 현장에 적용되어 불량 제품이 시장으로 유출되는 것을 최소화하고 있다. 하지만 생산되는 제품의 구조 및 크기에 따라 카메라 혹은 렌즈의 시야각이 나오지 않거나, 작업자에 의해 가려지는 등 이미지의 편차가 발생하게 되고 이는 검출률을 떨어뜨리게 원인이 된다. 또한 제조 라인에서는 생산모델이 변경되면서 길면 몇달이지만 빠르면 하루 이틀 단위로도 라인을 조정하게 되는데 그때마다, 이미지를 촬영하는 카메라의 위치를 조정함으로써 검사 환경 변화되는 것은 검출기 신뢰성 문제를 야기한다. 현재의 비전 검사기로는 다양한 제품 구조에 활용하기 어렵고, 특히 오디오 제품에 적용할 경우 제품의 특성상 비전 검사만으로는 제품의 가장 중요한 부분인 음향에 대한 부분을 커버할 수 없다. 따라서 오디오 제품군의 경우 음향을 기반으로 하는 별도의 품질 검사 방법이 필요하다.

스피커는 진동을 통해서 전기적 신호를 소리로 바꾸는 장치로 제품이 동작하는 동안 스피커에 의해서 제품 전체가 구조적으로 진동하게 된다. 조립과정에서 흡음재가 누락되거나 스크류가 느슨하게 조여져서 부품의 고정력이 약한 경우 스피커의 진동에 의해 부품이 떨리게 되고, 이것은 음질을 떨어뜨리는 요소가 된다.

과거의 오디오 제품군의 경우 제품의 동작과 관련된 시스템 및 앰프를 구성하는 회로 부품이 스피커와 분리되어 있어 스피커 진동에 의해 회로 부품들이 영향을 받지 않았다. 그러나 최근 오디오 제품의 소형화 추세에 의해 회로 부품이 챔버(chamber) 주변에 위치함에 따라 스피커 진동에

의해 주변부품이 떨리게 되어 소음 발생 가능성이 매우 높아지게 되었다. 하지만 조립과정에서 발생하는 미세 불량은 비전검사기로는 검출이 어려워 전문 시험원의 청취 검사에 의존하고 있다. 그러나 사람에 의존한 검사 방법은 고도로 훈련된 노동력이 필요할 뿐만 아니라, 일관되지 않은 검사 결과가 발생할 가능성이 높아 효율적이지 못하다. 이런 문제를 해결하기 위해 음향 제품군에 맞는 적합한 검사 모델 설계가 필요하다.

표 1: 구문 및 약어

약어	구문
ASC	Acoustic scene classification
CNN	Convolutional neural network
DNN	Deep neural network
FC	Fully connected network
FFT	Fast fourier transform
MFCC	Mel-frequency cepstral coefficient
MSE	Mean squared error
PCA	Principal component analysis
STFT	Short-time fourier transform
SVM	Support vector machine

1.2 접근

이상 탐지(Anomlay detection)는 이상치라고 하는 예상 동작과 일치하지 않는 비정상적인 패턴을 식별하는 데 사용되는 기술이다 [2]. 이상 탐지 기술은 라벨 여부에 따라 지도학습, 준 지도 학습, 비지도 학습으로 나누어지며 [3] 금융 [4], 의료 [5], 소셜 네트워크 서비스 [6], 산업 [7, 8, 9] 등 다양한 분야에 적용되고 있다.

산업 분야의 이상 탐지 에 대한 글로벌한 방법론은 진동 기반의 이상 탐지 기법으로, 진동 데이터를 통해 모니터링 대상의 상태를 파악한다. 이러한 방법은 가속도 센서에서 취득한 데이터를 처리함으로써 베어링과 같은 회전 기계에 대한 상태를 비롯하여 손상 위치, 손상 정도를 파악하는 것을 목적으로 한다. 이러한 진동 기반의 기계학습 알고리즘은 두 가지 과정을 거치는데 특징추출(feature extraction)과 분류(classification) 과정이다.

고전적인 특징 추출 과정은 수작업으로 특징을 추출하여 분류하였으나, 이러한 방식은 정상과 비정상을 구분하는 최적의 특징을 추출하기 어렵고, 연산 비용이 들어 실시간 검사 시스템에 효과적이지 않다. 이러한 문제에 대한 해결책으로 종단 간 학습(end-to-end learning) 방법이 제시 되었다 [9, 10]. 종단 간 학습은 입력에서 출력까지 하나의 파이프라인으로 데이터 입력에서 출력까지 사람의 개입 없이 판별 특성 표현을 학습한다.

진동기반의 이상 탐지 방법은 공정 설비의 상태 검사에 많은 연구가 진행되었으나, 이와 같은 방법은 제조업의 제품 품질 검사 공정에 적용하기에는 비효율적이다. 진동데이터는 한 개 이상의 가속도 센서에 의해 획득 되는데, 하루에도 수백에서 수천 개를 생산하는 제품에 센서를 부착하는 것은 큰 비용을 발생시키고 생산 효율을 떨어뜨린다.

이 연구는 조립 결함에 의해 발생하는 오디오 불량을 검출하려는 방법으로 음향 데이터를 기반으로 하는 방법론을 제안한다. 음향 데이터는 단일 음향 센서에 의해 획득되며, 음향 센서는 제품에 부착하는 방식이 아닌 제품 외부에 배치함으로써 라인에서 추가적인 공정이 발생하지 않도록 한다. 센서에 의해 획득 된 원시 신호에 대한 데이터 기반 식별을 목적으로 한다. 제안된 방법은 원시 오디오 신호를 종단 간 학습 방법으로 처리하여 유형별 불량을 분류 및 감지하기 위해 딥러닝에 크게 의존한다. 불량 검출 범위는 조립에 의해 발생하는 불량 중 소음을 발생 시켜 음질에 영향을 주는 것을 대상으로 하며, 음질이 아닌 부팅 불량 같은 시스템 문제, 혹은 무선성능 저하와 같은 기능성 불량은 본 연구에서 다루지 않는다.

1.3 연구보고서 구성

2장에서 합성 곱 신경망 아키텍처와 데이터 부족 문제를 해결하기 위한 방법으로 전이 학습, 데이터 증강에 대해 소개한다. 그 뒤 시작 감지 기법과 음향 장면 분류 연구에 대해 개관한다. 3장에서는 조립 결함에 대해 정의하고, 종단 간 1D 합성 곱 신경망 방법론을 제안한다. 또한 아키텍처 구조 및 데이터 증강 기법 등 실험에 적용하고자 하는 방법론에 대한 내용을 다룬다. 4장에서는 앞서 기술된 문제를 해결하기 위한 실험 방법 및 평가를 진행하며 마지막 장에서는 결론과 향후 연구 방안에 대해 정리한다.

제 2 장

관련 연구

본 장에서는 본 연구 보고서의 바탕이 되는 배경지식과 관련된 연구들에 관해 소개한다. 1절에서 본 연구의 기본 구조인 합성 곱 신경망에 관해 설명한다. 이어서 2절과 3절에서는 산업 분야에서 데이터 부족 문제를 해결하려는 방안으로 전이 학습과 데이터 증강에 대해 차례로 소개한다. 4절에서 음향 데이터의 이벤트 시점을 감지할 수 있는 시작 감지 기법에 관해 설명한 후 마지막 절에서 음향 장면 분류와 관련된 연구 흐름을 소개한다.

2.1 합성곱 신경망

합성 곱 신경망(CNN)은 데이터 처리하기 위한 특수한 종류의 신경망으로 시각 시스템의 구조에서 영감을 받았다 [11]. 합성 곱 신경망은 일반적으로 입력층(input layer), 제한된 수의 완전히 연결된 은닉층(hidden layer) 그리고 출력층(output layer)으로 구성되어있다. 이런 CNN 구조는 본질적으로 다층 퍼셉트론의 확장으로 볼 수 있지만, 컨볼루션 신경망은 컨볼루션과 풀링 작업에서 차이점을 보인다. 일반적인 신경망은 입력층의 각 뉴런은 은닉층의 뉴런에 연결되는 반면에 컨볼루션 신경망에서는 입력 영역 뉴런의 작은 부분만이 은닉층의 뉴런에 연결된다. 이것을 지역수용 필드(local receptive fields)라고 부른다. 지역수용 필드는 입력층에서 은닉층으로 시각적인 특징을 추출한다. 이렇게 함으로써 은닉층에 대해

동일한 가중치 및 바이어스 값을 갖게 된다. 이것은 은닉층의 뉴런이 전체 입력 이미지의 다른 부분에서도 동일한 특징을 감지할 수 있게 된다. 입력 층과 커널과의 컨볼루션 연산에 대한 출력을 특징맵(feature map)이라 하며, 일반적으로 컨볼루션 층은 여러 개의 특징맵으로 구성된다. 즉, 다른 특징맵은 다른 일련의 가중치와 바이어스를 가지고 있기 때문에 다른 특징(feature)을 감지한다.

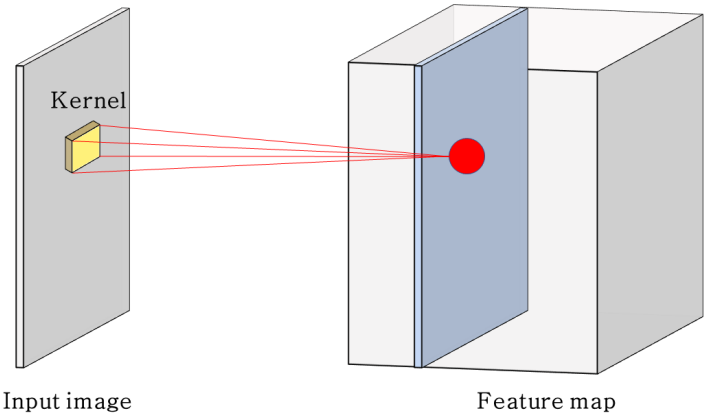


그림 1: 합성 곱 신경망 학습구조

이런 다수의 특징맵들은 정보를 유지하며 차원을 축소할 수 있는데, 이 과정을 풀링(pooling) 또는 서브 샘플링(sub-sampling)이라고 한다. 대표적인 기법으로 평균 풀링(average pooling)과 맥스 풀링(max-pooling) 방법이 있다. Average pooling은 필터의 값에서 평균값을 구하는 방법이며, 스트라이드(stride)에 따라 순차적으로 연산한다. Max-pooling 방식은 그림 2와 같이 필터에서 가장 큰 값을 추출하여 불필요한 정보를 간추리게 된다. 풀링은 입력을 약간 변환하더라도 대부분의 풀링된 출력값이 변경되지 않음으로써 입력에 대한 변환에 불변성을 유지하는데 도움을 준다

[10]. 이를 통해 데이터의 정보는 유지하면서 파라미터 수는 감소시키기 때문에 오버피팅을 억제한다. 또한 줄어든 파라미터 수에 비례하여 연산이 감소하기 때문에 하드웨어 리소스를 절약할 수 있다.

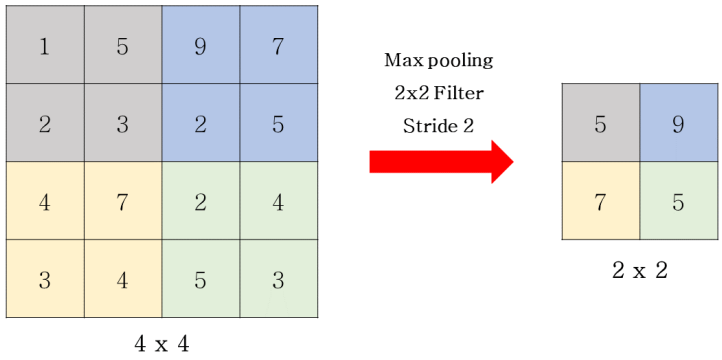


그림 2: Max polling(2x2 filter, stride 2)

위와 같은 방식으로 특징 맵이 추출되면 이 특징 맵에 활성화 함수를 적용하여 비선형의 값으로 바꿔주는 과정을 거친다. 대표적인 활성화 함수로는 시그모이드(sigmoid) 함수와 렐루(ReLU) 함수 가 있다. 컨볼루션 층은 네트워크에서 역전파에 의해 업데이트 되면서 학습이 진행되는데 시그모이드 함수를 활성화 함수로 사용할 경우 그라디언트 값이 $0 < y' \leq 0.25$로 신경망의 계층이 깊어지면 역전파 되는 에러값이 급격히 작아지는 그라디언트 소실(Gradient Vanishing) 문제가 발생한다. 그렇기 때문에 심층 망에서 이를 해결하기 위해 렐루함수를 적용한다.

$$f(x) = max(0, x) \tag{2.1}$$

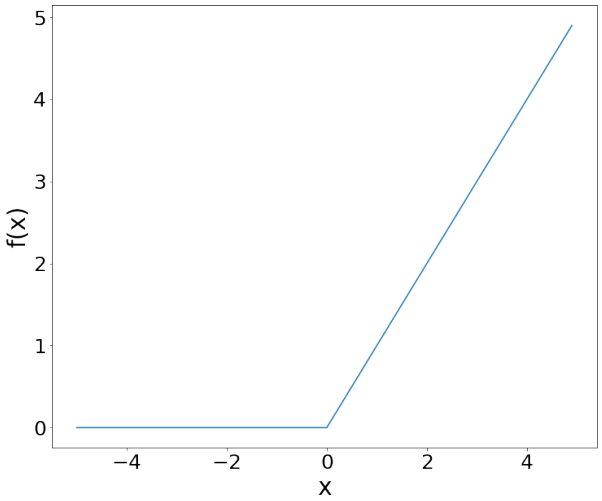


그림 3: ReLU 함수

식 2.1과 같이 렐루함수는 아주 간단한 구조로 되어 있어 빠른 계산이 가능하기 때문에 효율적이다 [12]. 활성화 효과는 있으면서 그림 3과 같이 그라디언트가 1로 유지되기 때문에 역전파 과정에서 소실 문제를 발생하지 않는다. 또한 렐루함수의 희소 활성화 구조(Sparse activation structure) [13]에 의해 입력이 고른 분포를 가진다면, 확률적으로 출력의 절반을 0으로 만들어 잡음에 영향을 덜 받고 오버피팅에 좋은 성능을 가진다. 단점으로는 0 이하일 때 그라디언트 값을 0 을 갖게 되어 정보가 사라지게 된다. 이러한 이유로 Leaky Rectified Linear Units (LeakyReLU) [14], Exponential linear unit (ElU) [15]와 같이 입력이 0 이하에서 기울기가 0이 아닌 대안이 제안되었다.

2.2 전이 학습

데이터가 희소한 상태에서 필요한 훈련 없이도 다른 표본으로로부터 학습된 지식을 통해서 성능을 높일 수 있다. 이러한 전이 학습(Transfer learning)은 관련 소스와 대상 도메인 간의 지식을 전달하는 것을 목표로 한다 [16]. 컴퓨터 비전 분야에서 지식을 전달하는 많은 시도가 많이 연구되었고 [17, 18], 최근에 오디오 관련된 전이학습이 연구되고 있다 [19].

컴퓨터 비전 분야의 경우 Pascal VOC [20], ImageNet [21]등 방대한 양의 이미지를 보유하고 있는 데이터 셋이 있어 전이학습에 적용하기 수월하지만 오디오의 경우 이런 대규모 데이터 셋이 부족한 단점이 있다. 이러한 대규모 데이터 셋의 부족으로 인해 Soundnet [22]은 오디오 이벤트 인식을 위해 시각적 모델에서 지식을 이전할 것을 제안했다.

교사-학생(Teacher-student)모델 [23]을 기반으로 시각 데이터에 대해 훈련된 CNN 모델을 사용하여 간단한 모델로 지식을 압축하여 학생 네트워크에 전송한다.

본 연구에서는 레이블 되지 않은 대량에 이미지 데이터셋에 의해 잘 훈련된 Soundnet [22]의 특징 층을 일부 차용하고, 최종 클래스 구분을 위한 분류기를 추가한다. 조립결함 데이터셋을 사용하여 종단 간 학습을 진행함으로써 세부조정(Fine tuning) 한다.

생산 초기에 방대한 데이터셋을 획득하기 어려운 산업 분야의 특성상 전이 학습은 학습된 지식을 활용하여 성능을 향상할 수 있는 좋은 방법이며, 조립 불량 공정에서 데이터에 기반한 식별의 일반화 가능성을 확인 할 수 있다.

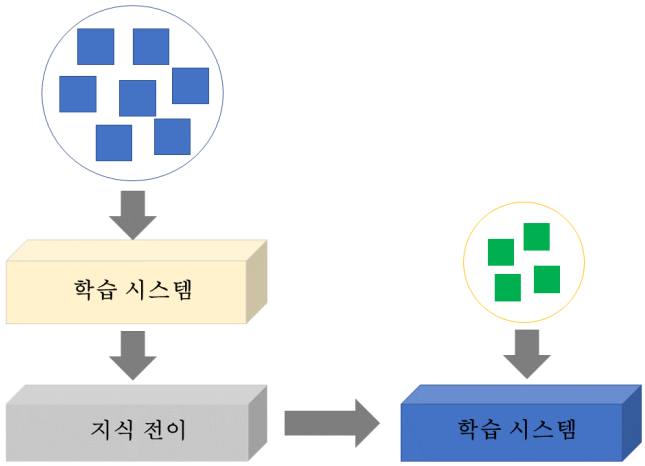


그림 4: 전이 학습 도식도

2.3 데이터 증강

모델 깊이가 깊은 심층 신경망일수록 입력에서 출력까지 비선형 함수를 학습하기 위해서는 대량의 훈련데이터가 필수적이다. 그러나 산업현장에서 발생하는 공정데이터는 생산 초기에 데이터가 충분하지 않고, 특히 정상 표본보다 비정상 표본은 발생 빈도가 현저하게 적기 때문에 상대적으로 더욱 희소하다.

정상데이터와 비정상 데이터의 불균형 및 데이터 부족 문제를 해결하기 위한 좋은 해결책은 데이터 증강(Data Augmentation)기법을 사용하는 것이다. 데이터 증강기법은 주석이 달린 훈련 샘플 모음에 하나 이상의 변형을 적용하여 추가적인 훈련 데이터를 생성하는 것이다 [12, 24]. 데이터 증강의 핵심 개념은 레이블이 지정된 데이터에 적용된 변형이 원 데이터의 레이블 의미를 변경하지 않는다는 것이다 [25]. 이미지로 예를 들면 자전거의 사진 이미지를 회전하거나, 미러링, 혹은 크기를 일부 변환

하는 것은 기존의 이미지와 다른 새로운 데이터를 생성하였으나, 기존에 라벨링 된 자전거의 의미는 그대로 유지가 된다. 추가로 생성된 데이터를 학습함으로써 분류 성능을 향상할 수 있다. 이러한 데이터 증강 기법이 오디오 분야에도 제안이 되었으며 모델의 분류 성능을 높이는 데 기여하였다 [25, 26].

2.4 시작 감지(Onset detection)

오디오의 시작(Audio onset)은 오디오 신호의 진폭의 인벨로프(envelope) 가 일시적으로 증가하는 순간으로, 다양한 오디오 이벤트의 시작 시점을 의미한다 [27]. 시작 감지는 자체로 유용성을 갖지는 않으나 상위 수준의 오디오 처리를 위한 기본 작업으로, 음악 자동주석 [28], 음악분석 (템포, 비트추적) [29], 음악 합성 [30] 및 분할 [31] 등 오디오 관련 활동 및 응용프로그램 분야에 적용되고 있다.

조립결함 데이터 셋은 공정에서의 편의성을 위하여 연속적인 데이터를 산정하였다. 연속 데이터에서 출력 레벨의 RMS 값 기반으로 제품의 출력을 감지하여 일차적으로 전처리를 하고, 시작 감지를 통해 이벤트의 시작 지점을 확인하여 3S의 오디오 구간을 추출하였다.

2.4.1 시간 영역 시작 감지

오디오 신호를 시계열에서 관찰할 때 앞에서 정의하였듯이 시작의 발생은 신호 진폭의 인벨로프가 증가하는 현상을 보인다. 따라서 초기에는 인벨로프를 정제 및 평활화하여 감지를 하였다 [32]. 이러한 방법은 원시 신호의 감소로 인해 피크 선택에 의한 감지에는 안정적이지 않았기 때문에 제곱을 통해 에너지에 대한 변형 식으로 발전하였다. 그 뒤 에너지

의 미분을 이용해 급격한 상승 구간에 대한 피크를 발견하였으며, Klapuri [33]는 음량의 변화가 전체 음량에 대해 상대적으로 인지된다는 결과를 도출하였다.

2.4.2 주파수 영역 시작 감지

신호의 스펙트럼을 이용한 시작 감지를 하는 다양한 기술들이 발전되었다. 주파수 도메인에서 에너지의 증가는 광대역의 이벤트로 나타나는 경향이 있다. 과도현상은 고주파일수록 두드러지는 특성을 보이는데 Masri [30]는 의해 주파수에 비례하여 가중치를 부여하는 HFC 함수를 제안하였다. 주파수 영역도 시간 영역의 시작 감지와 마찬가지로 순간 변화량을 고려한다. 따라서 인접한 STFT 프레임 사이의 거리로서 감지 기능을 공식화할 수 있다 [27].

2.5 음향 장면 분류

음향 장면 분류(ASC)는 주로 수동으로 제작된 오디오 특징(MFCC, Mel-filter bank, 스펙트로그램 등)에 서포트 벡터 머신 및 다수결 투표와 같은 분류기를 기반으로 소리를 구분한다 [34, 35]. 최근 CNN을 적용한 방법들이 제안되었으며, 환경 소리 분류 [25, 36, 37], 소음 분류 [38, 39]와 같은 여러 응용 분야에 활용되었다. 본연구는 로그-멜 스펙트로그램을 입력으로 하는 대신 원시파형을 사용하여 전처리를 최소화하였다. 원시파형을 사용하는 것이 2차원의 멜 스펙트로 그램과 유사한 기능을 하는 것이 [19, 40] 에 의해 증명되었다.

제 3 장

문제 정의 방법론

본 장에서는 산업 현장에서 발생하는 조립 결함 문제를 정의하고, 해당 문제의 해결방법을 제안한다. 오디오 제품의 생산공정에서 조립 시 발생 가능한 불량을 진단할 수 있는 시스템을 구축하여, 현재의 시험원들이 검사하는 공정을 대체하여 자동화하는 것을 목적으로 한다. 오디오 스피커 출력을 통해 조립 불량 여부 및 위치를 진단하는 최적의 모델을 도출하여 불량 검출 성능을 향상할 수 있는 방법을 제시한다. 본 장은 다음과 같이 구성된다. 1절에서는 본 연구에서 해결하고자 하는 문제를 상세히 설명하고 목적 함수를 정의한다. 2절에서는 해결 방안에 대해 간략히 개관한다.

3.1 조립 결함의 정의

제품의 생산에서 발생하는 주요 조립 결함의 원인 중 소음을 유발시키는 대표적인 원인으로 조립 공차에 의한 마찰음, 흡음재와 같은 부자재의 누락 등을 들 수 있다. 이러한 조립 결함은 생산라인에서 시험원이 테스트할 때 직접 검출이 어렵다. 또한, 결함을 발견했더라도 소음원의 위치를 판별하기가 쉽지 않기 때문에 제품을 수리하는 데 많은 시간이 소요하게 된다.

오디오 스피커는 진동을 통해 전기적 신호를 소리로 변환하는 장치이기 때문에 조립 결함이 발생한 오디오 제품의 경우 제품 동작 시 고정되

지 않은 부위에서 기계적 진동에 의한 마찰음이 발생한다. 이러한 진동은 특정 주파수 재생 시 스팩트럼에서 피크를 발생시킨다 [7].

그림 5와 같이 가청 주파수 대역의 주파수가 가변하는 스윕 신호를 출력한 파형을 특정 윈도우에서 FFT로 변환해보면 그림 6과 같은 결과를 얻을 수 있다. 정상 제품인 그림(a)과 달리 결함이 있는 제품 그림(b)의 경우 원 재생 주파수인 300Hz 이외에 다른 주파수에서 피크를 발생시킨다. 이러한 노이즈는 청취 시 오디오 품질을 저하하는 요인이 된다. 이에 본 연구에서는 생산라인에서 발생하는 조립 불량 중에 소음을 발생시키는 불량을 결함(Fault)으로 정의한다.

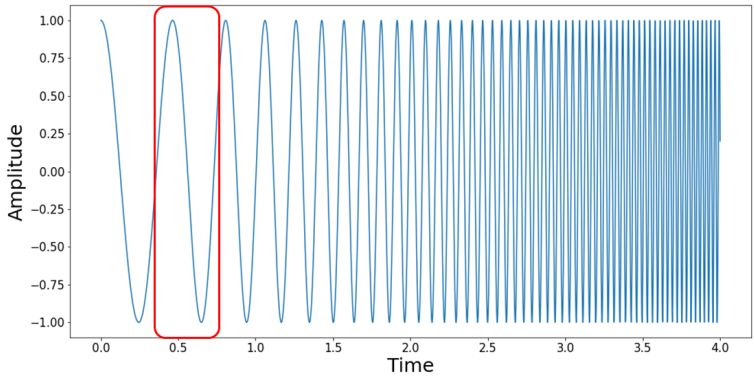
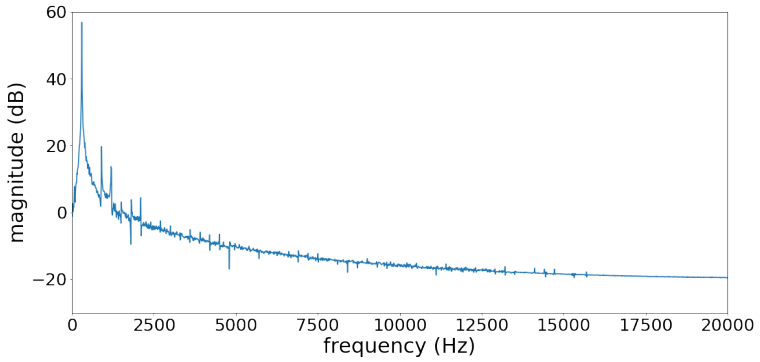
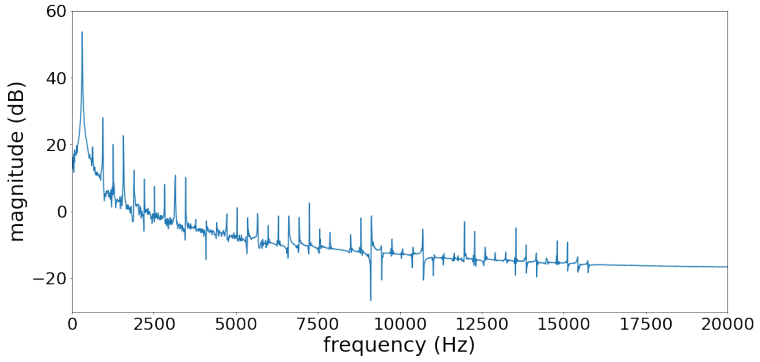


그림 5: 주파수 스윕 신호



(a)



(b)

그림 6: 300hz 재생 신호의 FFT 변환. (a) 정상시료 (b) 비정상시료

3.2 제안된 종단간 아키텍쳐

종단 간 학습(End-to-end learning)은 입력에서 출력까지의 파이프라인 네트워크 없이 한 번에 처리하는 아키텍처를 지칭한다 [41]. 이러한 종단 간 학습 방법은 처리하고자 하는 문제에 사전 지식의 필요성을 크게 줄이고, 데이터를 분석하는 엔지니어링을 개입을 최소화시킨다. 모델 파라미터만 일부 조정함으로써 학습이 가능하다. 이러한 학습 방법은 데이터 분석에서 널리 사용되고. 특히 CNN 구조를 이용하여 이미지 인식 [12, 42]에서 높은 성능을 보이고 있다. CNN은 이러한 이미지 분류 이외에도 음성인식 [43], 음악 태깅 [44], 음악 장르 분류 [19], 환경 소리 분류 [36, 45], 결함 진단 [8, 7, 9] 및 기타 여러 응용 분야에서 사용되었다.

그림 7은 음향 기반의 CNN 아키텍처를 표현하고 있는 것으로 이미지 인식에서 사용되는 구조와 유사하다. (a)에서 나타내고 있는 구조는 멜-스펙트로그램을 입력으로 하는 일반적인 CNN 모델로 오디오 분류에 널리 사용된다. 이러한 멜-스펙트로그램은 오디오 파형으로부터 STFT(short-time Fourier Transform) 을 통해 크기를 로그스케일로 변환 후 그 뒤 선형-멜 매핑을 및 크기 압축의 단계를 거치며 얻어질 수 있다[38]. 멜-스펙트로그램을 이용한 방법이 음향분류에 높은 성능을 보이지만, 창 크기, 홉 크기, 멜 필터 뱅크 크기 같은 고정 파라미터를 엔지니어가 수작업을 통해 선택하는 되는 단점이 있다. 종단 간 학습은 이러한 단점을 보완하기 위해 개발되었고, 도식(b)처럼 학습되어진 필터를 통해 STFT와 Mel-filter bank를 맵핑 시키는 과정을 대체할 수 있다. 이 연구는 종단 간 학습을 활용하여 빠르고 정확한 공정 불량 감지 시스템을 제안한다.

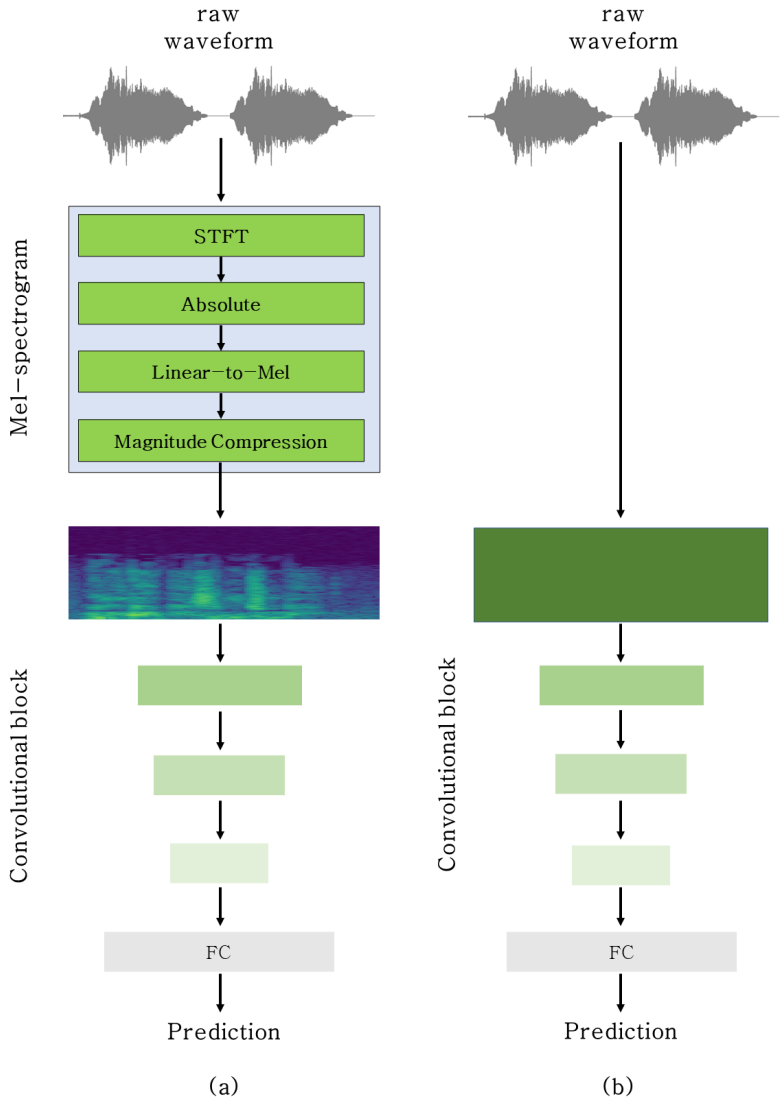


그림 7: 음향 기반의 CNN 아키텍쳐. (a) 멜 스펙트로그램 기반 학습 (b) 종단 간 학습

3.2.1 1D CNN 토폴로지

1D CNN은 일반적인 신경망과 유사하지만, 수작업으로 추출된 특징을 입력데이터로 사용하는 대신에 미가공 데이터를 사용한다는 점에서 차이가 있다. 이 연구에서 1D CNN을 이용하여 녹음된 미가공 오디오 데이터로부터 특징추출과 학습을 통합하는 과정을 거친다.

제안된 1D CNN은 커널과 특징맵이 1D array를 사용한다는 점에서 2D CNN과 구조적인 차이를 보인다. 입력벡터는 입력의 적절한 표현을 학습하기 위해 여러 훈련 가능한 컨벌루션 레이어를 통해 처리된다. 이때 입력 레이어의 작은 영역만이 히든레이어의 뉴런에 연결된다. 입력은 오디오 파형인 1D 배열이며 X로 표시한다. 네트워크는 식 3.1에 의해 주어진 계층적 특징 추출에 따라 예측 O에 입력을 매핑하기 위해 파라미터 집합 θ를 학습하도록 설계되었다. 여기에서 k는 네트워크의 히든레이어의 개수를 의미한다.

$$O = F(X|\theta) = f_k(...f_2(f_1(X|\theta_1)|\theta_2)|\theta_k) \tag{3.1}$$

컨벌루션 레이어에서 n번째 레이어의 순전파식은 다음과 같이 표현될 수 있다.

$$O_n = f_n(X_n|\theta_n) = h(W * X_n + b), \theta_n = [W, b] \tag{3.2}$$

식 3-2에서 *는 컨벌루션 연산을 의미하고, X_n는 n-1 레이어의 출력값으로 N개의 특징맵을 갖고 있는 이차원의 매트릭스다. W는 N개의 1차원 커널, b는 바이어스, h(·) 는 활성함수(activation function)을 의미한다. 그리고 각 컨벌루션 레이어 사이에 서브 샘플링 레이어를 적용하여

네트워크를 따라 이동할 때 특징 맵을 다운 샘플링하였다. 서브 샘플링 레이어에 따라 입력 배열의 크기가 정해진다. 마지막 컨벌루션 레이어의 출력은 FC layer에 의해 1차원 배열의 형태로 평탄화 된다.

다중 클래스의 분류 문제에서 출력 레이어의 노드 개수는 클래스의 개수이다. 각 노드의 값은 소프트 맥스 함수를 통해 정규화되어 각 요소를 확률값으로 변환한다. 주어진 $z \in R^n$ 에서 소프트 맥수 함수는 다음과 같이 정의된다.

$$\sigma(Z)_i = \frac{exp(z_i)}{\sum_{j=1}^{c} exp(z_j)} \tag{3.3}$$

C개의 카테고리의 조립 불량 데이터의 예측 확률은 $\hat{y}_c = \sigma(Z)_c$ 가 된다. 네트워크의 매개변수(parameter)는 역전파(back-propagation) 과정을 거치면서 손실 함수를 최소화하는 방향으로 조절된다[2]. 분류 문제에서 MSE(mean square error) loss보다 더 빨리 수렴하는 cross-entropy loss function(L)이 사용되었다. 이 손실을함으로써 각 데이터의 C 클래스에 대한 확률을 출력하도록 네트워크를 훈련한다.

$$Loss = -\sum_{i}^{c} y_i log(\hat{y}_i) + \frac{1}{2}\lambda||\theta||^2 \tag{3.4}$$

여기에서 y_i는 정답(ground truth), $\hat{y}_i$는 각 클래스 i 의 CNN 스코어가 된다. 다중 클래스 분류의 레이블은 one-hot encoded 이므로 positive class 만 손실 텀을 가지고 있어, 벡터에는 하나의 요소만을 가지고 있다.

3.2.2 네트워크 아키텍쳐

그림 8에 상세한 네트워크 아키텍쳐가 설명되어 있다. 총 7-레이어의 네트워크 아키텍처로 구성되어 있으며, 7개의 컨벌루션 레이어와 3개의 맥스풀링 레이어로 구성 되어 있다. 본 연구에서 사용된 CNN 구조는 레이블이 지정되지 않은 비디오에서 사운드 표현을 학습하기 위한 CNN 구조인 Soundnet [22]의 일부를 사용하였다. SoundNet은 서포트 벡터 머신 분류기와 두개의 동시 CNN을 사용하여 오디오 및 비디오의 동기화를 통해 학습한다. 반면에 제안된 1D CNN 아키텍쳐는 최종 컨벌루션 레이어인 con8을 대체하여 FC 레이어에 입력함으로써 원시의 오디오 파형에서 직접 표현을 학습하고 분류하였다.

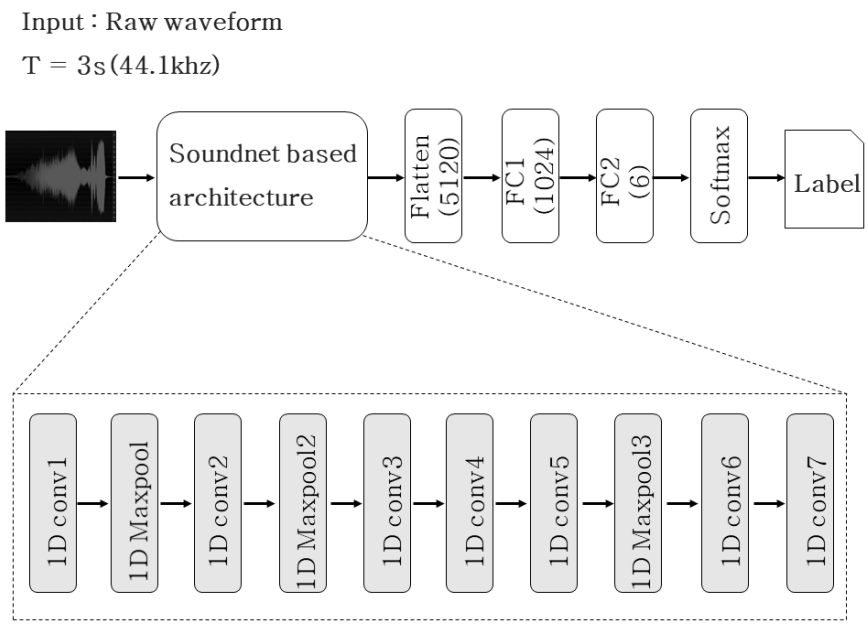


그림 8: 조립 결함 감지 분류 아키텍쳐

조립결함감지 네트워크는 비선형성의 특징을 가진 1차원 컨벌루션을 사용하고 각 레이어 사이에 렐루 활성 함수를 사용하여 사운드를 처리한다. 컨벌루션 네트워크는 학습해야 하는 파라미터 수를 줄임으로써 효율성을 높이는 장점이 있다. 또한 네트워크의 레이어를 쌓음으로써 일련의 하위 수준의 감지기를 통해 상위 수준의 표현을 학습할 수 있다. 또한 길이가 다른 오디오는 1차원 컨벌루션 연산에 의해 출력의 길이가 달라지는데 본 연구에서는 T= 3.0s 로 고정하였다. 이를 위해서 입력에 따라 시작감지를을 하게 되고, 감지 지점에서 길이가 부족한 경우 뒤에서부터 0으로 채워지게 된다. 진폭은 [-1, 1] 사이의 값으로 정규화하였다. 컨벌루션 레이어에 의해 각 특징들이 표현되고 출력 차원 크기는 5120으로 평탄화된다. FC 레이어를 통해 최종적으로 6개의 클래스 개수로 출력된다. 각 레이어의 상세 차원수는 표 5에서 자세히 확인 할 수 있다.

표 2: 네트워크 파라미터

Layer	conv1	pool1	conv2	pool2	conv3
Dimension	66151	8268	4135	516	259
Filters	16	16	32	32	64
Filter size	64	8	32	8	16
Stride	2	8	2	8	2
Padding	32	0	16	0	8
Layer	conv4	conv5	pool5	conv6	conv7
Dimension	130	66	16	9	5
Filters	128	256	256	512	1024
Filter size	8	4	4	4	4
Stride	2	2	4	2	2
Padding	4	2	0	2	2

오디오 신호의 샘플링 속도는 입력 샘플의 크기에 직접적인 영향을 미치고 모델의 계산 비용에도 영향을 준다. 나이키스트 이론에 의거 샘플링 주파수는 신호의 최대 주파수의 2배가 되어야 한다. 사람의 가청주파수 대역은 [20Hz, 20kHz]로 이 대역폭에서 표본화 정리를 만족시키는 최소 샘플링 레이트는 40kHz이므로, 일반적으로 음원에서 사용되는 샘플링 레이트인 44.1kHz로 설정하였다. 따라서 입력 차원은 44,100 * 3(s)로 132,300이 된다.

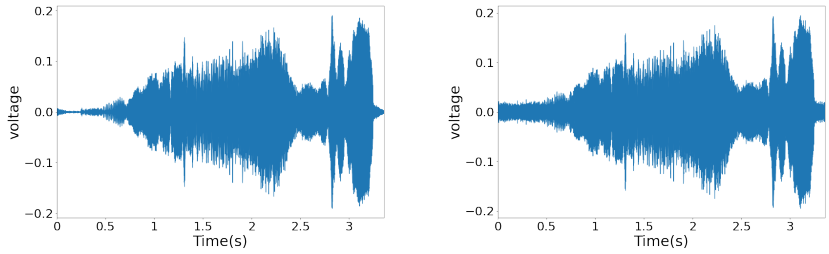
3.3 조립결함 오디오 데이터 증강

우리의 모델의 경우 2가지 오디오 데이터 변형을 통해 1:1의 비율로 데이터를 생성한다. 각 변형은 입력으로 변환하기 전에 원시 오디오 상태에 적용한다. 비교 지표로 사용하는 2D 합성 곱 신경망을 위한 데이터도 동일하게 원시 오디오 상태에 데이터 증강을 적용한 뒤 로그 멜 스펙트로그램 형태로 변형한다. 변형 및 증가는 데이터는 아래에서 확인 할 수 있다.

1) 화이트 노이즈 추가 : 화이트 노이즈를 원 소스에 추가하여 잡음에 대해서 강인한 모델을 학습할 수 있다. 생성된 데이터 z는

$$z = x + wy \tag{3.5}$$

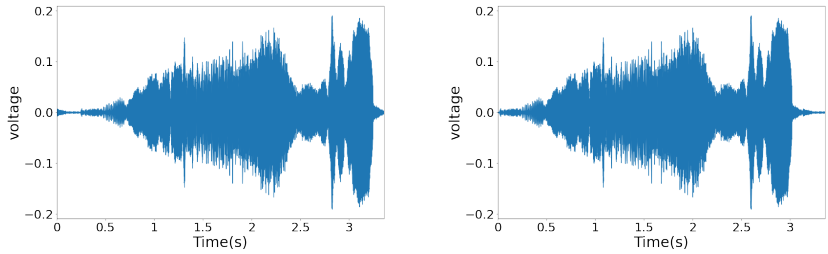
여기에서 x는 오리지널 샘플의 오디오 시그널, y는 정규분포를 갖는 샘플, w는 노이즈 비율이다. $w = 0.005$ 를 적용하여 잡음데이터를 생성하였고, 그림 9과 같이 원본데이터에 비해 그래프가 두꺼워진 것을 볼 수 있다.



(a) (b)

그림 9: 조립결함 데이터 셋 데이터 증강 예시1. (a) 오리지널 신호 (b) 화이트 노이즈 추가

2) 시간 시프팅 : 오디오 데이터를 밀어주거나 당기는 기법. 무작위로 오디오를 왼쪽 또는 오른쪽으로 이동한다. 오디오 이벤트 되는 시점을 변형하지만 이벤트 자체는 바뀌지 않음. 그림 10과 같이 0.3s 왼쪽으로 당겨진 것을 볼 수 있다.



(a) (b)

그림 10: 조립결함 데이터 셋 데이터 증강 예시2. (a) 오리지널 신호 (b) 시간 시프팅

제 4 장

실험 및 평가

본 장에서는 CNN 기반의 조립 결함 감지 시스템을 구현하고 해당 시스템을 글로벌 방법론인 원시 오디오 데이터를 멜 스펙트로 그램으로 변환한 2D CNN과 함께 성능 비교 평가를 실행한다. 본 장은 다음과 같이 구성한다. 먼저 조립공정 불량 분석을 위한 오디오 데이터를 간략하게 설명하고 해당 연구의 실험을 수행한다. 첫 번째로 제안한 종단 간 학습 1D CNN을 적용하여 데이터의 유형에 따라 정확도를 확인하고 불량을 검출한다. 두 번째는 글로벌 방법인 2D CNN을 적용하여 제안한 1D CNN과 결과를 비교확인 한다.

4.1 조립 결함 데이터셋

조립 결함 데이터 셋은 생산공정에서 제조된 무선스피커를 이용하여 수집되었다. 무선스피커에 가청주파수 대역의 주파수 스윕 음원을 재생하여 단일 센서에 의해 출력 신호가 수집되어 진다. 음원의 $fs = 44.1k$로 하며 블루 투스를 통해 재생한다. 소비자가 선택가능한 이퀄라이저중 출력 레벨이 가장 높은 이퀄라이저를 선정하여 제품의 최대 볼륨으로 측정한다. 수집된 데이터셋은 정상을 포함한 다섯 가지의 결함과 함께 총 6가지의 클래스로 나뉘어 진다. 결함의 원인 및 결함 위치는 표 3에 표기되어 있다.

생산은 일반적으로 여러 개의 하위 어셈블리(Sub-assembly) 형태로

표 3: 다섯가지 결함 구분 및 원인

No	구분	조립 결함 원인
0	N/A(정상)	N/A
1	Bottom case	미끄림 방지 고무 미부착
2	Bottom case	느슨한 스크류
3	Top case	흡음제 미부착
4	Chamber	느슨한 스크류
5	Top case	느슨한 스크류

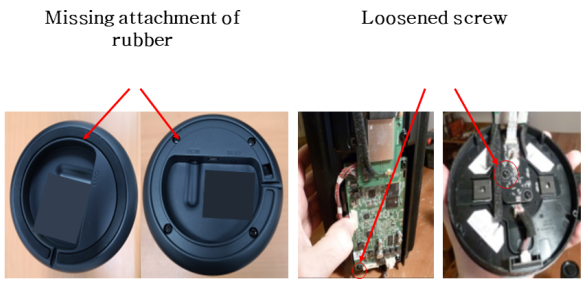


그림 11: 조립 결함 예시

1차 조립을 한 뒤 총조 라인에서 각 하위 어셈블리 들을 조립하여 제품을 완성한다. 제품의 불량이 발생했을 때 결함이 발생한 원인을 알 수 있게 되면 결함이 발생한 위치 즉, 하위 어셈블리를 교체 함으로써 생산량을 증대시킬 수 있다. 이 제품의 하위 어셈블리는 3종(하단 케이스/상단 케이스/챔버)으로 구성된다. 각 결함은 고무 미부착, 흡음재 미부착, 그리고 각 서브 아세이 체결 스크류의 느슨함으로 분류하였다. 이와 같은 원인으로 인하여 각 하위 어셈블리의 구성 부품이 오디오 재생 시 스피커와 같이

진동하게 되어 노이즈를 발생시키게 된다. 결함 조건의 상세 예시는 그림 11과 같다. (a)는 케이스 하단의 고무 미부착 (b)는 챔버의 느슨한 스크류 (c)는 케이스 상단의 느슨한 스크류의 예시다.

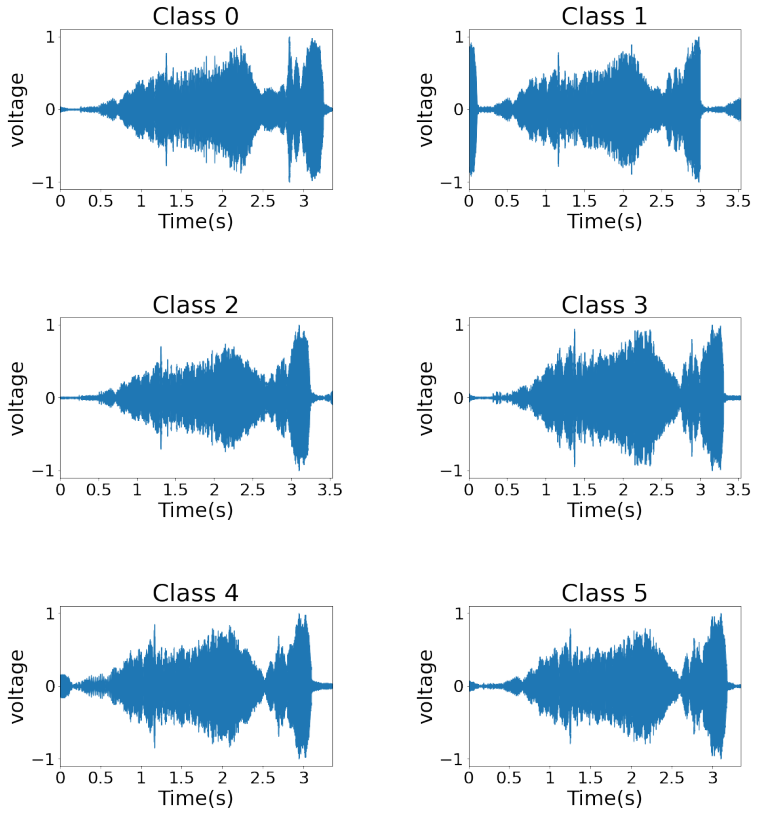


그림 12: 조립결함 데이터셋

조립 결함 데이터 셋은 6대의 오디오 제품에서 수집되었다. 각 세트당 1,250개의 데이터가 수집되어 총 7,500개의 오디오 데이터로 구성된다. 그림 12은 각 클래스별 데이터 셋을 시각화한 것으로 진폭은 0과 1사이 값으로 정규화되었다.

제안된 1D CNN의 성능 검증을 위한 비교 지표로 글로벌한 방법론인 멜 스펙트로그램 기반의 CNN을 활용하기 위해 전 처리된 데이터가 그림 13에 나타나 있다. 원시파형의 조립결함 데이터셋에서 STFT변환 후 로그 스케일의 멜 스펙트로 그램으로 변환한 형태로 원시 파형과 마찬가지로 다른 클래스의 데이터와 유사한 형태를 보이지만, Class 4의 멜 스펙트로그램 이미지를 살펴보면 1s 구간에 다른 클래스의 이미자와 대비해 높은 db 스케일의 주파수 성분을 육안으로도 포착 할 수 있다.

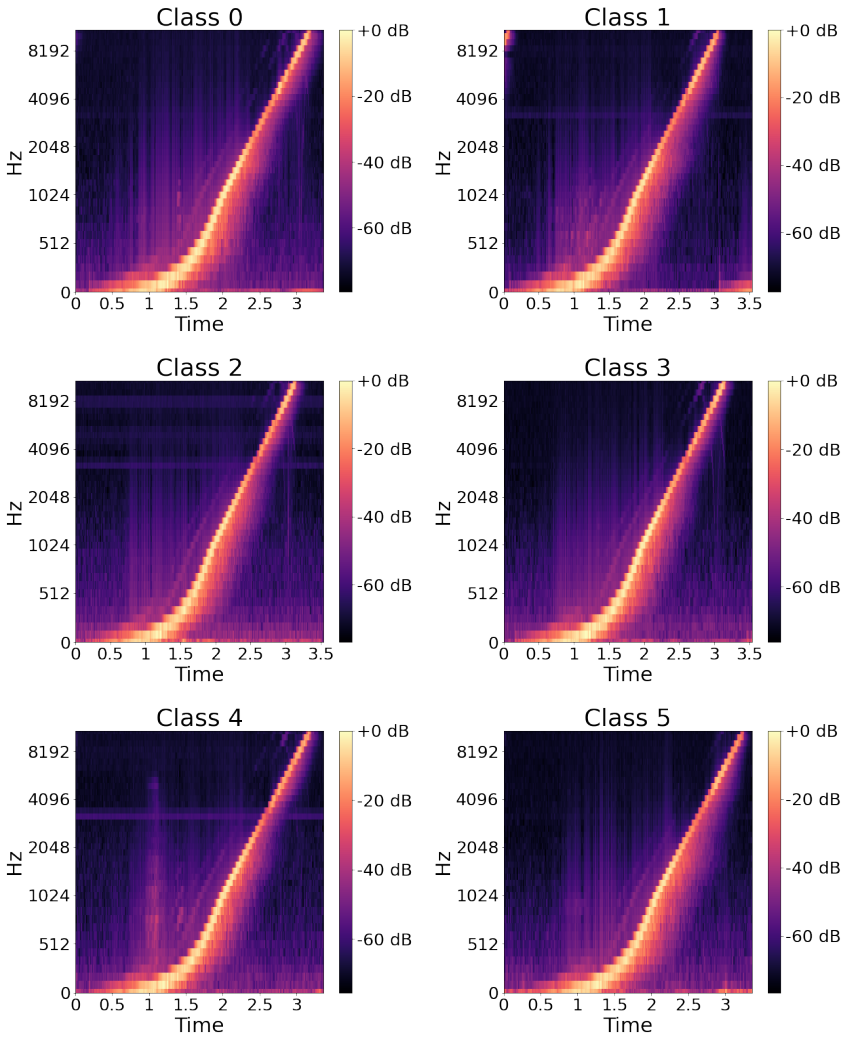


그림 13: 멜스펙트로그램으로 전처리된 데이터셋

4.2 실험 환경

조립 결함 데이터 셋은 서울대학교 34동 연구실험실에서 수집되었다. 그림 15는 오디오 소스 및 수신기 배열에 대한 테스트 환경을 보여

준다. 사각형은 무선스피커 즉 오디오 소스의 위치를 나타내는 것으로 Z 축을 따라 1.25m 높이에 위치해 있다. 이는 제품의 크기를 고려하여 생산 공정에서 작업이 용이한 생산라인의 높이와 유사하게 조성하였다. 실험에서 사용된 무선스피커의 높이는 210mm로 소형의 제품이지만 크기가 큰 제품에 적용한다면 실제 생산 환경과 유사하도록 제품의 높이 배치는 낮아져야 한다.

그림 14: 리시버 130F20

표 4: 130F20 스펙

성능	
mic diameter	1 / 4 inch
Frequency response($\pm 3dB$)	10 - 20000 Hz
Frequency response($\pm 4db$)	10 - 20000 Hz
Sensitivity	45mV / Pa
Dynamic range(3% distortion)	$> 122dB$

오디오 소스에서 x축으로 0.4m 떨어져 있는 위치에 리시버를 배치하였고, 높이는 1m에 위치 시켜 리시버와 오디오 소스의 z축을 일치시켰다. 본 실험에 사용된 리시버는 PCB PIEZOTRONICS사의 130F20 ARRAY MICROPHONE 제품으로 그림 14에 표기되어 있다. [10Hz,20kHz]의 주파수 응답을 가져 가청주파수 대역을 커버하며, 다이내믹 레인지가 넓어 실험에 사용된 오디오 출력에 대해 왜곡 없이 녹음이 가능하다. 상세 스펙은 표4에 표기되어 있다. 리시버의 위치는 그림 15의 단색 원에서 확인 할 수 있다.

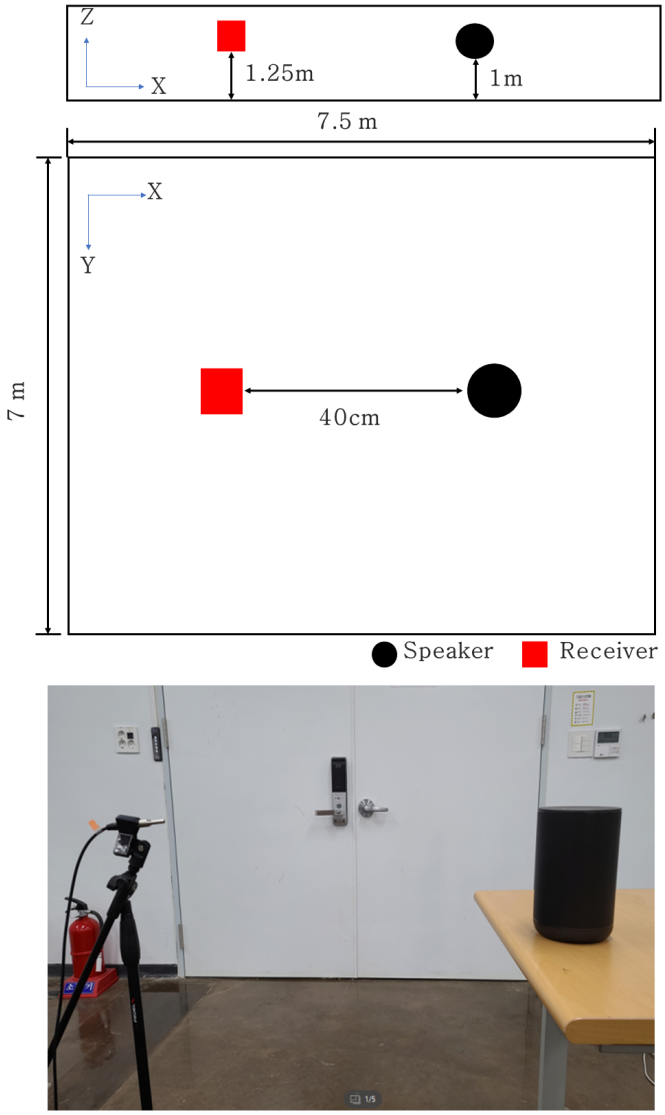


그림 15: 오디오 소스와 리시버 배치 환경

실험에서 사용된 무선스피커의 스피커 유닛은 지향성 스피커로 스피커의 방향에 따라 해당 방향 축을 벗어날수록 음압의 크기가 다른 지향

특징을 나타낸다 [46]. 고주파 대역일수록 그 경향성은 더욱 크다. 이러한 특징은 그림 4.3의 실험에서 측정된 스피커 방향별 스피커 출력 데이터에서 확인할 수 있다. (a) 정면 방향의 데이터와 달리 (b)(c)는 3s 부근인 고주파 대역에서 음압이 떨어지는 것을 확인할 수 있다. 이와 같은 스피커 방향에 따른 지향성 차이를 일반화하기 위해 그림 16와 같이 리시버 정면을 기준으로 좌/우 45도씩 스피커 방향을 변경하여 데이터를 수집하였다. 데이터는 각도별로 1:1:1로 구성하였다.

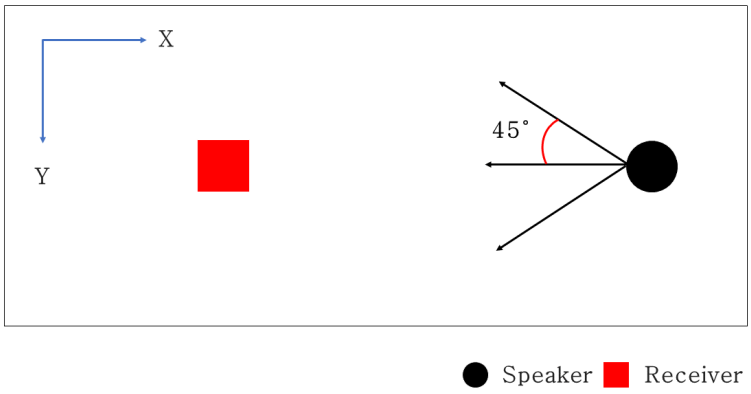
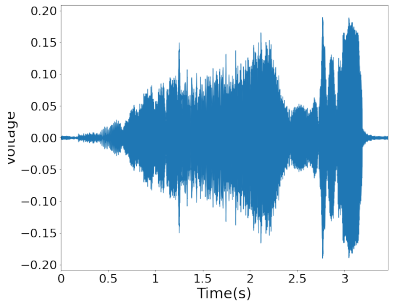
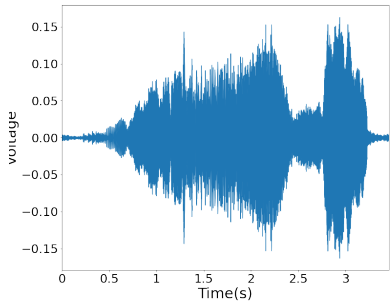


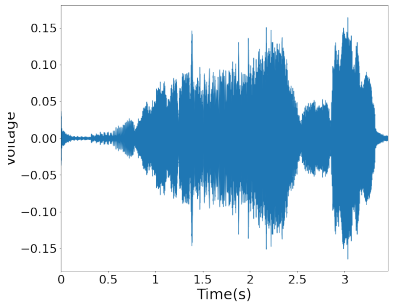
그림 16: 스피커 방향별 실험 조건



(a) 0°



(b) +45°



(c) −45°

그림 17: 스피커 방향별 출력 파형

4.3 학습 방법

3절에서 설명한 바와 같이 *Loss*는 크로스 엔트로피 로스에 L_2-regularization을 결합함으로써 페널티 텀을 주어 loss를 최소화하는 방식으로 학습한다. λ와 learning rate는 무작위 탐색 방법을 사용하여 최적의 하이퍼 파라미터를 찾는다. 무작위 탐색 방법은 그리드 탐색 방법에 비해 하이퍼 파라미터 최적화에 더 효율성을 보인다 [47]. 각각의 하이퍼 파라미터는 $[10^{-4}, 10^{2}]$ 범위의 로그 공간에서 연속 균등분포를 따르는 각기 다른 50개의 난수 쌍 lr, λ 을 생성한다. 생성된 값은 50번의 학습 반복을

통해 유효 정확도를 측정한다. 생성 된 50개의 난수 쌍 중에 유효 정확도가 가장 높은 값을 갖는 값이 하이퍼 파라미터로 선택된다.

Optimizer로 Adam을 사용하여 훈련은 300번 반복 후에 종료된다. 우리의 네트워크에서 사용된 가중치는 레이블이 지정되지 않은 비디오 [22]에서 사전 훈련된 가중치로 초기화되었다. 각 신경망 레이어 중간에 배치 정규화를 적용하여 gradient의 scale이나 초기값에 대한 dependancy를 최소화 하였다 [48].

4.1장에서 설명한 것처럼 데이터셋은 총 6개의 오디오 제품에서 수집됐다. 본 연구는 학습된 모델을 통해 생산라인에서 다른 제품의 결함을 감지하는 것을 목적으로 하고 있기 때문에 표5와 같이 훈련 셋, 유효 셋 테스트 셋을 제품 별로 구분한다. 즉 1 4번 제품의 오디오 출력 데이터는 훈련 셋, 5번 제품의 오디오 출력 데이터는 유효 셋 6번 제품의 데이터는 테스트 셋으로 구성하여 조립결함 시스템의 제품별 일반화성능 평가한다.

표 5: 데이터셋 구성

데이터셋	Train	Validation	Test	Total
수량(EA)	10000	2500	2500	15000
구성비(%)	(66.6)	(16.6)	(16.6)	(100)
오디오 제품번호	1-4	5	6	1-6

4.4 평가

조립 결함 분류 모델의 성능 평가를 위하여 성능을 시각화 할 수 있는 혼동 행렬(Confusion matrix)을 사용한다. 이진 분류(Binary Classification) 문제에서 혼동 행렬은 표 6와 같다. TP, TN은 각각 True positive, True negative로 실제값을 맞게 예측한 부분이며, FP, FN은 False positive, False negative로 실제값과 다르게 예측한 부분을 의미한다.

표 6: 혼동 행렬(Confusion Matrix)

Division		Prediction	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

정밀도(precision)는 모델의 예측값이 얼마나 정확하게 예측됐는가를 나타내며, 재현도(recall)는 실제값 중에서 모델이 검출한 실제값의 비율을 나타내는 지표로 각각 식4.1과 식4.2로 표현 된다.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4.2}$$

정확도(accuracy)는 모델이 바르게 분류된 부분의 비율로, 혼동행렬에서 대각선 부분이다.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.3}$$

F1 score는 정밀도와 재현율의 조화 평균으로 식4.4 와 같이 표현 할 수 있다. 이 점수는 FP와 FN을 모두 고려한다. 직관적으로 이해하기는 쉽지 않지만, F1 score는 일반적으로 정확도보다 더 유용하다. 특히 클래스간 분포가 고르지 않은 불균형의 경우에 더욱 더 유용하다.

$$\text{F1-Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4.4}$$

조립 결함 감지 모델은 다중 클래스에 대한 평가 지표로 6개의 클래스에 대한 혼동 행렬 표를 적용하며, scikit-learn을 통해 수행한다. 조립결함 검출 데이터셋은 표 2와 같은 6개의 클래스로 분류되어 레이블링 되었다: 0, 1, 2, 3, 4, 5. 50번 반복 동안 Adam을 사용하여 loss를 최소화하는 최적의 파라미터를 찾는다.

4.4.1 네트워크별 성능 평가

제안 된 네트워크의 성능 검증을 위해 기존의 글로벌한 방법인 스펙트로그램을 입력으로 하는 2D 컨벌루션 방법과 비교 평가를 한다. 제안된 종단 간 학습 방법에 대한 내용은 4.3절에 설명해 놓았고, 비교평가를 위한 2D 컨벌루션에 대한 실험은 동일한 조립 결함 데이터 셋에서 스펙트로그램으로의 전처리 과정을 추가한다. 조립 결함 데이터 셋을 STFT을 사용하여 주파수 도메인으로 변환한다. 변환된 값의 파워스펙토그램을 Mel-filterbank에 통과 시켜 Mel-scale 에 대한 주파수 크기로 변환한다. Alexnet과 VGG19 네트워크를 사용하여 학습하며 32개 샘플의 배치 크기가 사용되었다.

조립 결함 데이터 셋에서 데이터 증강 과정을 거치지 않고 조립결함을 분류한 결과가 표 7에 나타나 있다. 각 네트워크 모두 사전 학습된 매개변수를 이용하여 전이 학습을 진행하였다. 조건 모두 반복회수 300회를 학습한 결과로 평균 정확도 soundnet 42.2%, Alexnet 52.5%, VGG19 56.0%의 분류 성능을 보였다. 세가지 네트워크 모두 품질검사에 적용하기엔 현저히 낮은 정확도를 보였다. 제안된 1D 네트워크의 경우 2D 네트워크에 비해 열세의 성능을 보였다. Training accuracy는 몇번의 반복만으로도 90% 이상으로 충분히 높은 정확도를 보였으나, Test accuracy가 42.2%로 심각한 오버 피팅(overfitting)이 발생 되었다. 전처리 과정을 통해 특징을 추출하고 차원크기를 줄인 2D CNN(입력 차원수 224*224=50,176)에 비해 1D CNN의 경우 더욱 도드라지는 현상을 보였다.(입력 차원수 44,100*3=132,300).

각 네트워크의 분류 결과의 혼동행렬표가 그림 18에 나타나있다. (a)의 경우 class 0과 2를 제외하고 평균 이상의 정확도를 보인다. 상대적으로

표 7: 데이터 증강 적용하지 않은 네트워크 별 조립결함 분류 성능평가

unit : percent(%)

	Dimension	Accuracy
Soundnet	1D	42.2
Alexnet	2D	52.2
VGG19	2D	56.0

정확성이 떨어지는 class 0과 2 사이에서 혼동하는 경향을 보인다. (b)와 (c)의 경우 class 1에서 98%이상의 높은 정확도를 보였고 class 2와 class 5에서 상대적으로 정확도가 낮았다. class 2와 class 5의 경우 예측에 대한 특정한 패턴을 찾을 수 없었다.

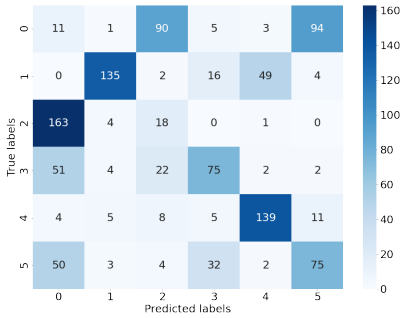
오버피팅을 피하고 분류 성능을 높이기 위해 3.3절에서 언급한 바와 같이 데이터 증강(Data augmentation) 중 화이트 노이즈 추가와 시프팅 기법을 적용하여 데이트를 증량 하였다. 그 외 학습방법은 앞선 실험과 동일하게 진행하였다.

표 8는 데이터 증량 적용 후에 모델의 조립결함분류 성능을 보여준다. 세가지 아키텍처 모두 정확도 94% 이상으로 높은 성능을 보여 앞선 실험 대비 현저한 분류 성능 차이를 보였다. 제안된 SoundNet 기반의 1D CNN 모델은 VGG19와 더불어 가장 높은 정확도인 99.9%를 달성하였고, 2D CNN 대비 동등 이상의 성능을 보였다.

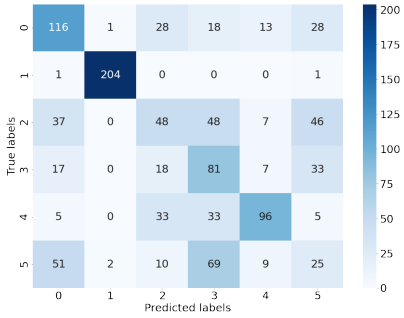
표 8: 데이터 증강을 적용한 네트워크 별 조립결함 분류 성능평가

unit : percent(%)

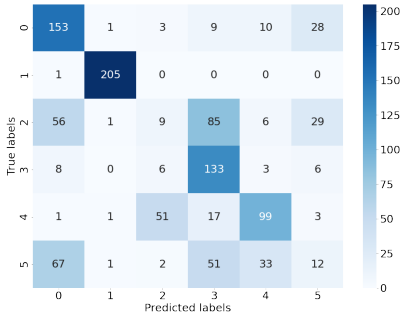
	Dimension	Accuracy
Soundnet	1D	99.9
Alexnet	2D	94.6
VGG19	2D	99.9



(a) Soundnet

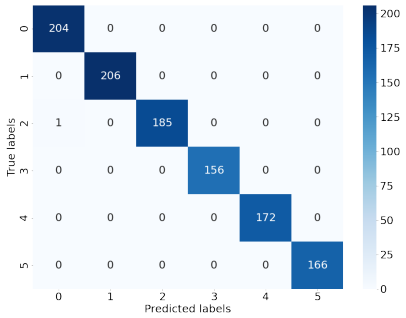


(b) Alexnet

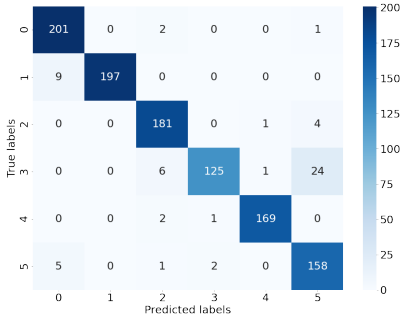


(c) VGG19

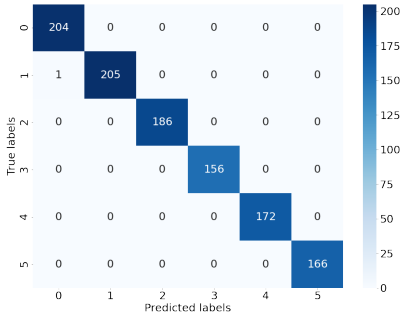
그림 18: 데이터 증량을 적용을 하지 않은 조립결함 감지 결과 혼동행렬표



(a) Soundnet



(b) Alexnet



(c) VGG19

그림 19: 데이터 증량 적용한 조립결함 감지 결과 혼동행렬표

표 9: 조립 결함 분류 결과 Precision, recall, and F1 score (데이터증강 미적용시)

		0	1	2	3	4	5	Average
Soundnet	F1 score	0.0455	0.7542	0.1091	0.5190	0.7554	0.4261	0.4349
	Precision	0.0394	0.8882	0.1250	0.5639	0.7092	0.4032	0.4548
	Recall	0.0539	0.6553	0.0968	0.4808	0.8081	0.4518	0.4245
Alexnet	F1 score	0.5383	0.9879	0.2972	0.4000	0.6316	0.1645	0.5032
	Precision	0.5110	0.9855	0.3504	0.3253	0.7273	0.1812	0.5134
	Recall	0.5350	0.9809	0.1268	0.4508	0.6556	0.1538	0.4838
VGG19	F1 score	0.6245	0.9880	0.0700	0.5898	0.6130	0.0984	0.4973
	Precision	0.5350	0.9809	0.1268	0.4508	0.6556	0.1538	0.4838
	Recall	0.7500	0.9951	0.0484	0.8526	0.5756	0.0723	0.5490

표 10: 조립 결함 분류 결과 Precision, recall, and F1 score (데이터증강 적용시)

		0	1	2	3	4	5	Average
Soundnet	F1 score	0.9976	1.0000	0.9973	1.0000	1.0000	1.0000	0.9991
	Precision	0.9951	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992
	Recall	1.0000	1.0000	0.9946	1.0000	1.0000	1.0000	0.9991
Alexnet	F1 score	0.9349	1.0000	0.9427	0.9766	0.9883	0.8449	0.9479
	Precision	0.9594	0.9777	0.9577	0.8803	0.9854	0.8952	0.9459
	Recall	0.9853	0.9563	0.9731	0.8013	0.9826	0.9518	0.9417
VGG19	F1 score	0.9975	0.9975	1.0000	1.0000	1.0000	1.0000	0.4973
	Precision	0.9951	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992
	Recall	1.0000	0.9951	1.0000	1.0000	1.0000	1.0000	0.9992

4.4.2 스피커 지향성 분석

오디오를 구성하는 여러 요소(스피커, 앰프, 케이블 등) 중에서 한계가 명확하고 발전이 더딘 부분이 스피커이다. 소리와 마찬가지로 형태가 없이 공기가 진동해 소리가 발생한 소스로부터 모든 방향으로 동일한 소리를 재생하는 것이 이상적인 스피커 모델이나, 진동판을 통해 진동 시켜 소리를 내는 방식으로는 진동판의 특성이나, 그것을 둘러싸는 인클로저 등 많은 부분이 필요하고 그로 인해 왜곡을 만들어 낸다. 중 · 저음역대의 소리는 방향과 상관없이 잘 퍼져 나가는 성질이 있으나, 고역으로 갈수록 지향성의 영향을 크게 받는다 [46]].

이러한 스피커의 지향성 특징으로 인해 우리의 모델을 공정에 적용할 때 작업자가 제품의 방향을 마이크와 일치 시켜 놓지 않으면 해당 분류 모델에 성능이 떨어질 소지가 있다. 이런 스피커 방향성에 대한 일반화 성능을 갖기 위해 방향별 스피커 출력 데이터를 훈련 셋에 포함하여 학습하였다. 본 절에서는 스피커 방향별 데이터를 훈련 셋에 제약을 주면서 모델의 성능 차이를 확인한다.

학습은 총 세 가지 경우로 진행을 한다. 첫 번째 학습에서는 0도 방향의 데이터만 훈련 셋으로 구성하고 테스트 셋으로 +45도, -45도 방향의 데이터를 구성한다. 두 번째 학습은, 0도 +45도 방향의 데이터를 훈련셋에 -45도방향의 데이터를 테스트셋으로 구성한다. 마지막으로 세 번째 학습은 0도, +45도, -45도의 일부 데이터를 훈련셋으로 포함하고, 훈련셋에 포함되지 않은 데이터를 테스트셋으로 구성한다. 마지막 시험은 앞선 4.4.1 절의 실험 결과를 사용한다.

스피커 지향성에 따른 학습 결과는 표 11과 표 12를 통하여 확인 할 수 있다. 표 11은 Soundnet을 통하여 특징추출 없이 end-to-end 로 학습한

결과이고, 표 4.8은 mel-spectrogram으로 특징 추출 후 Alexnet으로 학습한 결과이다. 두 가지 결과 모두 스피커 방향이 일부만 학습데이터에 포함된 경우보다, 여러 방향을 모두 학습시켰을 때 더 높은 정확성을 보이는 것을 확인할 수 있다. 위 실험 결과를 보아 스피커 지향성의 일반화 성능을 위해서는 학습데이터에 여러 방향의 스피커 출력 데이터가 반드시 필요한 것을 알 수 있다.

그 외 추가적인 특징으로는 방향 별 데이터가 일부만 학습데이터에 포함된 경우 2D CNN을 통해 학습한 결과가 84.5%로 1D CNN을 통해 학습한 결과 62.0% 대비 더 높은 분류 성능을 보였다. 그러나 지향성에 따른 데이터가 고르게 입력에 포함될 경우 1D CNN이 99.9%로 2D CNN의 정확도 94.5% 보다 더 높은 성능을 보임을 알 수 있었다.

표 11: 스피커 지향성 데이터 포함 여부에 따른 1D CNN 학습 결과

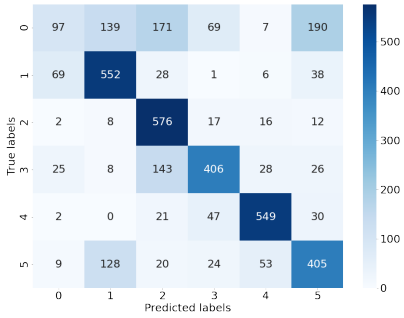
unit : percent(%)

	0	1	2	3	4	5	Average
0°	47.6	64.8	60.1	72.0	83.8	57.8	62.0
0°/45°	72.9	85.0	76.4	79.7	87.9	85.3	79.4
0°/45°/-45°	99.8	99.4	100.0	99.5	100.0	100.0	99.9

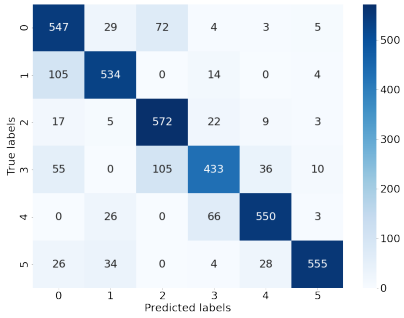
표 12: 스피커 지향성 데이터 포함 여부에 따른 2D CNN 학습 결과

unit : percent(%)

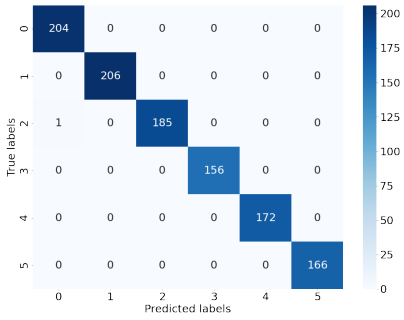
	0	1	2	3	4	5	Average
0°	85.6	98.2	83.9	70.7	95.1	82.2	84.5
0°/45°	89.4	99.7	96.1	82.1	98.6	77.8	88.5
0°/45°/-45°	93.4	100.0	94.2	97.6	98.8	84.4	94.4



(a) 0°

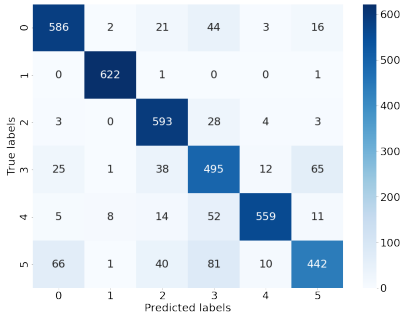


(b) 0°/45°

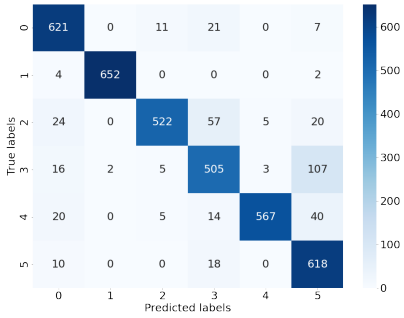


(c) 0°/45°/-45°

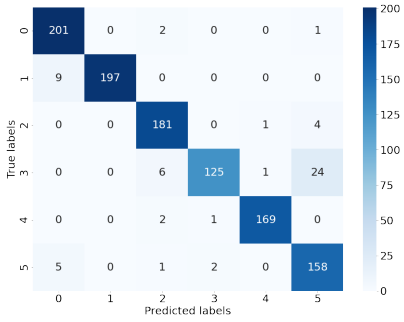
그림 20: 스피커 지향성 데이터 포함 여부에 따른 1D CNN 학습 결과 혼동 행렬표



(a) 0°



(b) 0°/45°



(c) 0°/45°/-45°

그림 21: 스피커 지향성 데이터 포함 여부에 따른 2D CNN 학습 결과 혼동 행렬표

제 5 장

결론

본 연구 보고서는 산업 제조 공정에서 발행하는 오디오의 조립 불량을 검출하기 위한 종단 간 1D 합성 곱 신경망을 제안하였다. 산업 현장의 데이터 부족 문제를 해결하고 학습 성능을 극대화하는 방안으로 데이터 증량(Data Augmentation) 및 전이 학습을 수행하였다. 데이터 증량 기법으로는 화이트 노이즈 추가 및 시간 축 이동을 하여 1:1의 비율로 데이터를 증량하였고, 많은 양의 라벨링 되지 않은 비디오에 의해 사전 학습된 매개변수를 사용하였다.

제안된 시스템은 양산 중인 오디오 제품의 오디오 출력 데이터를 이용하여 실험적으로 검증되었다. 총 15,000개의 오디오 샘플 데이터 셋에 대해 평가되었으며, 비교 평가를 위해 종단 간 학습 방법 이외에도 로그 스케일의 멜-스펙트로그램으로 변환하여 입력으로 사용하는 2D CNN을 같이 시험하였다. 제안된 1D 아키텍처인 SoundNet은 데이터 셋에 포함된 6개의 결함 유형에 대해 99.9%로 가장 높은 분류 성능을 보여준다.

제안된 종단 간 1D 아키텍처는 음향 분류를 위한 대부분의 다른 CNN 아키텍처보다 매개변수가 적다. 또한, 오디오 분류를 위한 별도의 신호 처리가 필요하지 않기 때문에 공정의 실시간 검증에 매우 적합하고, 임베디드 시스템에서 사용하기 좋다. 또한, 제안된 접근법은 훈련 데이터와 테스트 데이터를 서로 다른 제품에 사용함으로써 일반화 가능성을 확인하였다.

글로벌하게 사용되는 방법인 진동데이터 기반의 불량 감지 기법은 산업용의 설비같은 대형 장비에 사용하기 좋으나, 제조업의 제품 품질 검증에 적용하기 위해서는 센서를 부착해야 하는 등의 추가적인 비용이 많이 발생하는 문제가 발생한다. 제안된 방법은 음향 기반의 데이터를 사용함으로써 이 문제를 해결하기 위한 효과적인 솔루션을 제공한다.

이 실험은 오디오 조립공정에서 진행되는 시험원에 의한 청취시험 조건과 유사한 조건으로 구성되었다. 주변 소음을 고려하여 실험이 진행되었으나, 실제 공정에서는 주변 소음이 아주 높기 때문에 우리의 실험과 같은 청취룸이 없다면 분류 성능이 저하 될 수 있다. 그러나 고성능의 지향성 마이크를 이용한다면 이러한 주변 소음은 극복이 가능하다.

본 연구를 통해 제조 현장에서 스피커 출력이 있는 제품(오디오, TV 등)에 대한 조립 공정 불량률을 감소시켜줄 것으로 기대한다. 또한 최근 디지털 마이크가 탑재된 하이엔드 오디오 제품군의 경우 별도의 시험공간을 구성하지 않아도 제품의 자체의 마이크를 통해 라인에서 검사가 가능할 것으로 판단한다. 이를 이용하여 소비자에게 판매된 뒤에서 자체 임베디드를 이용하거나 클라우드 서버를 이용하여 지속적인 자가 테스트로 결함 유무를 체크 할 수 있을 것을 기대한다.

본 연구는 오디오 신호의 1D 표현을 사용하여 높은 분류 성능을 보였으나, 일부 스피커 지향성에 대한 데이터에 제약을 줄 경우, 일부 클래스에서 2D 표현(로그 스케일의 멜 스펙트로그램)에 비해 성능이 저하될 수 있다. 그렇지만 1D와 2D 필터 사이에 상보성이 존재할 수 있음으로, 앙상블 등 분류 성능을 확보하는 방안은 향후 과제로 남긴다. 또한, 이러한 스피커 지향성에 관한 문제는 마이크가 탑재된 제품 즉, 음성인식 스피커나 휴대전화 등의 제품에 실장 된 마이크를 통해 녹음 함으로써 On device 형태로 불량을 검출하게 되면, 스피커 방향성에 대한 요소를 제거 할 수

있다. 차후 이것에 대한 검증이 필요하다.

다만 본 제안된 모델은 다른 제품군에 적용할 경우 미세조정의 과정을 거쳐야 하므로 초기 공정에서 데이터의 레이블 작업은 추가적인 비용을 발생시킨다. 이러한 문제를 해결하기 위한 비지도 학습 방법은 향후 과제로 남긴다.

참고 문헌

[1] 김양희, "도요타 리콜사태의 발생 원인과 교훈," *[KIEP] 오늘의 세계경제*, vol. 2010, no. 2, pp. 0–0, 2010.

[2] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*, pp. 4393–4402, PMLR, 2018.

[3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[4] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in *International conference on neural information processing*, pp. 483–490, Springer, 2016.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] K. Anand, J. Kumar, and K. Anand, "Anomaly detection in online social network: A survey," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 456–459, IEEE, 2017.

[7] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.

[8] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-d convolutional neural networks," *IEEE*

Transactions on Industrial Electronics, vol. 63, no. 11, pp. 7067–7075, 2016.

[9] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman, "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks," *Journal of Sound and Vibration*, vol. 388, pp. 154–170, 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, JMLR Workshop and Conference Proceedings, 2011.

[14] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[15] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[17] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *2011 international conference on computer vision*, pp. 2252–2259, IEEE, 2011.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

[19] J. Lee, J. Park, K. L. Kim, and J. Nam, "Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, p. 150, 2018.

[20] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, pp. 892–900, 2016.

[23] L. J. Ba and R. Caruana, "Do deep nets really need to be deep?," *arXiv preprint arXiv:1312.6184*, 2013.

[24] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[25] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[26] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation.," in *ISMIR*, vol. 2015, pp. 248–254, 2015.

[27] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on speech and audio processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[28] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 37–40, IEEE, 2011.

[29] M. F. McKinney, D. Moelants, M. E. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *Journal of New Music Research*, vol. 36, no. 1, pp. 1–16, 2007.

[30] P. Masri and A. Bateman, "Imroved modelling of attack transients in music analysis-resynthesis.," in *ICMC*, Citeseer, 1996.

[31] P. Brossier, J. P. Bello, and M. D. Plumbley, "Real-time temporal segmentation of note objects in music signals," in *ICMC*, Citeseer, 2004.

[32] W. A. Schloss, *On the Automatic Transcription of Percussive Music–From Acoustic Signal to High-level Analysis*. PhD thesis, Stanford University, 1985.

[33] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 6, pp. 3089–3092, IEEE, 1999.

[34] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 171–175, IEEE, 2015.

[35] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[36] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2015.

[37] H. Zhou, Y. Song, and H. Shu, "Using deep convolutional neural network to classify urban sounds," in *TENCON 2017-2017 IEEE Region 10 Conference*, pp. 3089–3092, IEEE, 2017.

[38] H. Choi, S. Lee, H. Yang, and W. Seong, "Classification of noise between floors in a building using pre-trained deep convolutional neural networks," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 535–539, IEEE, 2018.

[39] H. Choi, H. Yang, S. Lee, and W. Seong, "Classification of inter-floor noise type/position via convolutional neural network-based supervised learning," *Applied Sciences*, vol. 9, no. 18, p. 3735, 2019.

[40] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[41] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Advances in neural information processing systems*, pp. 739–746, Citeseer, 2006.

[42] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616, 2009.

[43] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks,"

Advances in neural information processing systems, vol. 22, pp. 1096–1104, 2009.

[44] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 366–370, IEEE, 2018.

[45] H. Choi, H. Yang, S. Lee, and W. Seong, "Type/position classification of inter-floor noise in residential buildings with a single microphone via supervised learning," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 86–90, IEEE, 2021.

[46] J.-H. Kim, J.-T. Kim, J.-O. Kim, and J.-K. Min, "Acoustic characteristics of a loudspeaker obtained by vibration and acoustic analysis," *Transactions of the Korean Society of Mechanical Engineers A*, vol. 21, no. 10, pp. 1742–1756, 1997.

[47] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

Abstract

Acoustic-based assembly defect detection system

Hansu Lee
Graduate School of Practical Engineering
Seoul National University

Despite many efforts by global manufacturers to secure product quality, production defects due to various causes continue to occur. If defective products are delivered to consumers, this is an important problem that can damage corporate management by instantly destroying brand image in addition to direct cost.

Recently, with the development of deep learning technology, many studies based on anomaly detection are being conducted in manufacturing sites. However, if you look at the related studies that have been recently studied, most of them deal with the surface defects of products through vision inspection or the state inspection of production facilities using vibration data. However, these methods are not suitable for application to audio products that deal with sound. In addition, it takes a lot of time and money to attach a sensor to apply it to quality inspection of products in a manufacturing industry that produces hundreds or thousands of units per day.

In this paper, we present an end-to-end approach to sound classification that learns representations directly from raw audio signals using speaker output data measured by a single acoustic sensor and a synthetic product neural network. Seven convolutional layers are used to learn various filters related to the classification task of assembly defects. The dataset was collected through output data from multiple speakers, and it was shown that the knowledge learned from the output of some speakers can determine whether other speakers are defective, with 99% accuracy. The proposed assembly defect detection method shows higher performance than most methods that use the existing 2D representation as input. In addition, it has fewer parameters compared to other architectures, making it efficient for real-time product quality inspection.

Through this study, it is expected that the assembly process defect rate will be reduced for product groups with speakers installed inside the product, such as TVs and AVNs for vehicles.

Keywords : Deep learning, Anomaly detection, Fault detect, CNN

Student Number : 2020-21071