KIPS Trans. Softw. and Data Eng.
Vol.8, No.5 pp.193~204 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.5.193

Enhanced Sound Signal Based Sound-Event Classification

Yongju Choi† · Jonguk Lee†† · Daihee Park††† · Yongwha Chung††††

ABSTRACT

The explosion of data due to the improvement of sensor technology and computing performance has become the basis for analyzing the situation in the industrial fields, and various attempts to detect events based on such data are increasing recently. In particular, sound signals collected from sensors are used as important information to classify events in various application fields as an advantage of efficiently collecting field information at a relatively low cost. However, the performance of sound-event classification in the field cannot be guaranteed if noise can not be removed. That is, in order to implement a system that can be practically applied, robust performance should be guaranteed even in various noise conditions. In this study, we propose a system that can classify the sound event after generating the enhanced sound signal based on the deep learning algorithm. Especially, to remove noise from the sound signal itself, the enhanced sound data against the noise is generated using SEGAN applied to the GAN with a VAE technique. Then, an end-to-end based sound-event classification system is designed to classify the sound events using the enhanced sound signal as input data of CNN structure without a data conversion process. The performance of the proposed method was verified experimentally using sound data obtained from the industrial field, and the f1 score of 99.29% (railway industry) and 97.80% (livestock industry) was confirmed.

Keywords : Noise Robustness, Sound Signal Generation, End-to-End Architecture, Deep Learning

향상된 음향 신호 기반의 음향 이벤트 분류

최 용 주† · 이 종 욱†† · 박 대 희††† · 정 용 화††††

요 약

센서 기술과 컴퓨팅 성능의 향상으로 인한 데이터의 폭증은 산업 현장의 상황을 분석하기 위한 토대가 되었으며, 이와 같은 데이터를 기반으로 현장에서 발생하는 다양한 이벤트를 탐지 및 분류하려는 시도들이 최근 증가하고 있다. 특히 음향 센서는 상대적으로 저가의 가격으로 현장 정보를 왜곡 없이 음향 신호를 수집할 수 있다는 큰 장점을 기반으로 다양한 분야에 설치되고 있다. 그러나 소리 취득 시 발생하는 잡음을 효과적으로 제어하지 못한다면 산업 현장의 이벤트를 안정적으로 분류할 수 없으며, 분류하지 못한 이벤트가 이상 상황이라면 이로 인한 피해는 막대해질 수 있다. 본 연구에서는 잡음 상황에서도 강인한 시스템을 보장하기 위하여, 딥러닝 알고리즘을 기반으로 잡음의 영향을 개선 시킨 음향 신호를 생성한 후, 해당 음향 이벤트를 분류할 수 있는 시스템을 제안한다. 특히, GAN을 기반으로 VAE 기술을 적용한 SEGAN을 활용하여 아날로그 음향 신호 자체에서 잡음이 제거된 신호를 생성하였으며, 향상된 음향 신호를 데이터 변환과정 없이 CNN 구조의 입력 데이터로 활용한 후 음향 이벤트에 대한 식별까지도 가능하도록 end-to-end 기반의 음향 이벤트 분류 시스템을 설계하였다. 산업 현장에서 취득한 음향 데이터를 활용하여 제안하는 시스템의 성능을 실험적으로 검증한바, 99.29%(철도산업)와 97.80%(축산업)의 안정적인 분류 성능을 확인하였다.

키워드 : 잡음 견고성, 음향 신호 생성, End-to-End 구조, 딥러닝

1. 서 론

정보 통신 기술(information communication technology) 및 융합 기술(convergence technology)의 발전은 4차 산업혁명의 실마리가 되었으며, 다양한 산업 분야로 영향력이 커지고 있다[1]. 이와 같은 기술의 발전을 근간으로 원격으로 산업 현장을 모니터링하기 위해 설치한 수많은 센서로부터 산업 현장과 관련된 대용량의 데이터가 발생하고 있으며, 이를 토대로 산업 현장의 상황을 분석하고자 하는 시도들이 최근 급격하게 증가하고 있다[2, 3]. 이처럼 산업 현장을 관찰하기 위해 사용하는 다양한 센서 중 음향 센서는 다른 센서에 비하여 상대적으로 저가의 비용으로 설치할 수 있으며, 현장 상황을 왜곡 없이 정보를 수집할 수 있다는 장점이 있다[4]. 이

※ 본 연구는 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018R1D1A3B07044938).
† 준 회 원 : CJ대한통운 정보전략팀 연구원
†† 정 회 원 : 고려대학교 컴퓨터융합소프트웨어학과 초빙교수
††† 정 회 원 : 고려대학교 컴퓨터융합소프트웨어학과 교수
†††† 종신회원 : 고려대학교 컴퓨터융합소프트웨어학과 교수
Manuscript Received : December 20, 2018
Accepted : February 3, 2019
* Corresponding Author : Jonguk Lee(eastwest9@korea.ac.kr)

와 같은 장점을 기반으로 산업 현장에서 발생하는 이벤트를 음향 센서를 기반으로 수집한 후, 수집된 음향 정보를 분석하는 학술적인 시도들이 다양하게 보고되고 있으며, 현장의 이벤트를 소리 기반으로 원격 분석이 가능함을 보고하고 있다[5-9].

그러나 소리 기반 연구의 가장 큰 문제점은 현장에서 발생하는 음향에는 중요한 이벤트뿐만이 아니라 다양한 잡음에도 노출이 되며, 해당 잡음으로 인하여 시스템의 성능이 저하된다는 부분이다[10-12]. 이와 같은 잡음에 대한 문제점을 해결하기 위하여, 기존 소리를 이용한 연구들은 다음과 같은 전형적인 구조에서 잡음을 제어하는 연구를 진행하였다. 해당 구조는 아날로그 소리 신호 취득 모듈, 소리 특징 추출 모듈, 추출된 특징을 기반으로 이벤트를 식별하는 모듈로 구성된 3단 파이프라인(pipeline) 구조이며, 잡음 제거를 위한 과정은 다음과 같다. 제한된 시간과 저성능의 컴퓨팅 자원으로 인하여 아날로그 원시 신호(raw waveform) 그 자체를 분석하기 어려운 관계로[13], 해당 원시 신호를 대체하여 표현할 수 있는 시간 영역(time domain) 특징 또는 주파수 영역(frequency domain) 특징과 같이 연구자들이 정의한 특징(hand-crafted feature)을 기반으로 소리 정보를 추출한 후 추출된 특징 정보에서 잡음을 제어하는 연구를 진행하였다. 이처럼 시간 또는 주파수 영역의 정보로 변환된 특징 정보를 기반으로 잡음을 제어하는 방법은 원시 신호 그 자체에서 잡음을 제어하는 방법이 아닌 관계로 원시 신호 정보가 일부 손실되며, 이로 인한 소리 기반의 이벤트 분류 시스템 자체의 성능 저하가 발생한다. 또한, 해당 모듈에 대한 전문적인 지식 및 현장 상황에 맞는 세세한 튜닝(tuning) 작업과 같은 어려움도 존재한다[13, 14].

최근 향상된 컴퓨팅 성능과 딥러닝 기술을 기반으로 이미지 인식 분야에서 급격한 성능 개선이 이루어졌으며, 음향 분야에서도 이를 받아들여 다양하게 응용되고 있다. 특히, 본 연구의 관심사인 잡음 제거에도 그 응용된 사례가 최근 발표되었다. 해당 방안은 딥러닝 알고리즘 중 하나인 원본에 가까운 진짜 같은 데이터를 생성하는 GAN(Generative Adversarial Network)을 기반으로 VAE(Variational Auto-Encoder) 기술이 구현된 SEGAN(Speech Enhancement GAN)을 활용한 방안으로서[15], 원본 음성 신호 그 자체에서 잡음 제어가 가능함을 보고하였다. 또한, 기존의 모듈화 된 단계를 거쳐 음향 이벤트를 분류하는 전통적인 3단 파이프라인 구조에서 벗어나 원시 데이터를 시스템의 입력으로 직접 사용한 후 특징 생성 및 이벤트 분류까지 한 번에 수행이 가능한 end-to-end 방식을 음향 분석에 적용하여 성능이 개선된 사례들도 보고되고 있다[16-19].

본 논문에서는 위와 같은 최근 학계의 연구 성과들을 참조하여 실제 산업 현장에서 응용이 가능한 음향 기반의 이벤트 분류 시스템을 설계하고자 한다. 제안된 시스템은 먼저, 원시 음향 신호에서 잡음의 영향력을 제거한 후, 잡음이 제거되어 향상된 원시 신호를 CNN(Convolutional Neural Network)의 입력으로 사용하여 음향 식별에 효과적인 소리 특징을 딥러닝을 기반으로 생성한다. 생성된 특징은 MLP(Multi-Layer Perceptron)에 적용되어 해당 음향 이벤트에 대한 분류까지도 가능한 end-to-end 기반의 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 주제인 음향 이벤트를 이용한 최신 연구들과 그 한계점에 대해 살펴보며, 3장에서는 본 연구에서 제안하는 시스템에 대하여 상세히 다룬다. 4장에서는 본 논문에서 제안하는 시스템의 실험 결과 및 성능 분석을, 마지막으로 5장에서는 본 연구의 결론 및 향후 연구를 언급한다.

2. 관련 연구

본 연구의 관심 대상인 산업 현장에서 발생하는 음향을 기반으로 이벤트를 분류하는 학술적인 노력 중 철도산업과 축산업 분야를 기준으로 살펴보면 다음과 같다.

먼저 철도산업 분야에서의 연구를 살펴보면, 열차의 진로 방향을 제어하는 철도 부품 중 하나인 선로전환기의 결함은 열차의 탈선 및 충·추돌을 발생시킬 수 심각한 문제이며, 실제 2018년 12월 8일에 발생한 KTX 열차의 탈선 문제도 선로전환기의 이상으로 발생하였다. 따라서, 선로전환기의 이상 여부를 조기에 탐지하는 것은 매우 중요한 문제이다. Lee 등[5]은 선로전환기의 전환 시 발생하는 소리 신호에서 MFCC (Mel-Frequency Cepstrum Coefficient) 특징 정보를 추출한 후 SVM(Support Vector Machine) 분류기를 이용하여 선로전환기의 이상 상황을 탐지하는 연구를 수행하였으며, Choi 등[6]은 선로전환기의 노후화에 따른 장비 교체시기를 예측하기 위하여, 선로전환기의 노후화로 인한 피로도를 시간과 주파수 대역의 소리 특징을 조합하여 SVM 분류기에 적용 후 선로전환기의 피로도를 탐지하는 연구를 발표하였다.

다음은 축산업 분야의 돼지 음향 데이터를 활용한 연구이다. Guarino 등[7]은 소리 주파수 대역에서 필터링 기법과 진폭 변조 등의 기법을 적용하여 소리 특징 벡터를 생성한 후, DTW(Dynamic Time Warping) 알고리즘을 적용하여 돼지의 정상 소리와 호흡기 질병의 DTW 값의 차이를 기반으로 호흡기 질병을 탐지하는 연구를 보고하였으며, Chung 등[8]은 돼지 호흡기 질병으로 인한 기침 여부를 MFCC 특징 정보와 단일 클래스 탐지기인 SVDD(Support Vector Data Description)를 이용하여 호흡기 질병을 탐지하고 SRC(Sparse Representation Classifier)를 기반으로 질병의 종류를 분류하는 이중 구조를 제안하였다. 또한, Lee 등[9]은 돼지의 호흡기 질병을 효과적으로 탐지하기 위하여, 시간 영역과 주파수 영역의 다양한 소리 특징 중 호흡기 질병 탐지에 유효한 특징들만을 선택 및 조합하는 방법에 관한 연구를 발표하였다.

이처럼 산업 현장에서 발생하는 음향 데이터를 활용하여 특정 이벤트를 탐지 및 분류 가능성을 확인하는 학술적인 성과가 보고되었으나, 앞서 언급된 연구들은 실제 소리 취득환경에서 발생하는 잡음의 영향력을 최소화하는 부분에 대한 고민은 상대적으로 미흡한 편이다. 최근 이와 같은 잡음에 대한 영향력을 제어하여 실제 산업 현장에서 발생하는 잡음 환

경에서 시스템의 인식 성능을 높이기 위한 시도들이 보고되고 있다.

Sharan 등[20]은 잡음 환경에서도 강인한 성능을 보장하기 위하여, 음향 신호를 스펙트로그램(spectrogram)으로 변환한 후 해당 스펙트로그램을 작은 영역으로 나눈 모듈화(modulation) 특징을 기반으로 잡음 환경에서 음향 이벤트를 분류하는 연구를 수행하여 일정 부분 이벤트 식별 성능이 개선됨을 확인하였다. 그러나, 신호의 백색 잡음(white noise) 에너지가 강하게 포함된 경우에는 식별 성능이 현저하게 낮아진 결과를 보였다. 또한, Choi 등[12]은 이미지의 잡음을 개선하는 DNS (Dominant Neighborhood Structure) 알고리즘을 소리 분야로 확장하여 소리 잡음을 제어하는 방법을 제안하였다. 해당 방법론은 소리 신호를 선형 변환하여 2차원 이미지로 변환한 후 이미지로 변환된 음향 신호를 DNS 기법에 적용하여 잡음을 제어하였으며, Lee 등[21]은 음향 신호를 스펙트로그램으로 변환한 후 CNN의 커널 기법을 이용하여 잡음 환경에 강인한 특징이 생성됨을 확인하는 결과를 보고하였다. 그러나 위에서 소개한 연구들의 일부분 성공적인 결과에도 불구하고 잡음의 원인을 원본 신호에서 직접 제거한 것이 아닌, 음향 신호를 특정한 도메인의 특징 정보로 변환한 후 변환된 특징 정보에서 잡음을 제거하는 간접적인 접근방법이다.

최근 딥러닝 학습 시 부족한 데이터 문제를 해결하기 위하여, 취득한 진짜 데이터와 매우 유사한 데이터를 자동으로 생성하는 GAN 알고리즘[22]이 발표되었으며, 다양한 분야로 확산 및 응용되고 있다. 특히, 음성 신호에서 잡음을 제거하기 위해 제안된 SEGAN은 GAN 작동 원리를 차용 및 확장한 방법으로써, 학습 수행 시 입력 데이터는 잡음이 포함된 음향 신호를 사용하고, 출력 결과는 잡음이 제거된 음향 신호로 설정한다. 입력된 데이터가 GAN의 학습 과정을 반복 수행하는 과정에서 잡음이 없는 음향 신호가 생성되면 학습이 종료하도록 설정하였다. 이때 인코딩(encoding)과 디코딩(decoding)이 반복되며 잡음이 제거되는 과정은 원본 음향 신호에서 잡음을 제거하는 데 쓰였던 VAE와 유사한 기능을 수행하게 된다. 특히, SEGAN은 VAE에 비해 잡음을 복원하는 과정에서 blur 현상과 과적합(over fitting) 경향이 상대적으로 적게 나타나 더욱 주목받고 있다[23].

본 연구에서는 이처럼 원시 신호 자체에서 잡음을 제거할 수 있는 SEGAN의 장점을 산업 분양에서 발생하는 음향에 확장 적용하고자 한다. 또한, 원시 음향 신호 자체에서 잡음을 제거하게 되면 전통적인 모듈 기반의 파이프라인 구조가 아닌 원본 신호를 시스템의 입력 및 특징 추출 그리고 분류까지 한 번에 진행하는 end-to-end 구조를 활용할 수 있다는 부분이다. End-to-end 구조를 음향에 도입한 사례를 살펴보면, Dieleman 등[16]은 음악 분류를 위해 원본 오디오 신호뿐만이 아니라 스펙트로그램을 동시에 CNN의 입력값으로 사용하여 음악을 분류하는 end-to-end 구조를 제안하였으며, Collobert 등[17]은 사람의 음향 신호를 입력받아 해당 음성에 맞는 문자로 변환시키기 위하여, 원시 아날로그 신호를 입력으로 받는 CNN 기반의 음향 모델과 음성 신호를 문자로 변환해주기 위한 그래프 디코딩(graph decoding)을 결합한 end-to-end 기반의 자동 음성 인식 시스템을 제안하였으며, 해당 시스템은 기존의 HMM(Hidden Markov Model)과 GMM (Gaussian Mixture Model)을 활용한 방법보다 수행시간이 빠르고 정확한 성능을 보였다. Zhang 등[18]은 기존 음성 인식 시스템의 경우 음성 인코더 네트워크 구성 시 네트워크의 계층(layer) 구조를 얕게 쌓아 사용하였지만, 높은 인식 및 일반화 성능을 확보하기 위해 매우 깊게 계층을 쌓은 CNN 구조의 end-to-end 음성 인식 시스템을 제안하였으며, Kim 등[19]은 가변 길이 오디오에 대한 처리가 가능한 CTC (Connectionist Temporal Classification) 알고리즘과 주목(attention) 기법 기반의 인코더-디코더를 결합한 end-to-end 방식으로 우수한 성능을 보이는 음성 인식 모델을 제안하였으며, 긴 길이의 오디오에서도 실시간 처리가 가능함을 확인하였다.

이와 같은 학계의 연구 성과들을 참조하여, 본 논문에서는 원시 신호 자체에서 SEGAN을 활용하여 잡음을 제거한 후, 잡음이 개선된 음향 신호를 CNN 구조의 입력으로 사용하여 특징을 자동 생성한다. 생성된 특징을 MLP에 적용하여 잡음 상황에도 강인하게 음향 이벤트를 분류할 수 있는 end-to-end 구조를 제안한다.

3. 향상된 음향 신호 기반의 음향 이벤트 분류

3.1 제안하는 시스템

본 논문에서 제안하는 향상된 음향 신호 기반의 음향 이벤트 분류 시스템의 전체 구조는 Fig. 1과 같다. 본 구조의 특징은 안정적인 음향 이벤트 분류 성능을 확보하기 위해서, 음성 신호 자체에서 잡음을 개선하기 위해 제안된 SEGAN을 활용하여 산업 현장에서 발생한 음향 신호 원본에서 잡음을 개선하는 것이다. 또한, 음향에서 잡음을 제거하는 것에 그치는 것이 아닌 해당 음향 이벤트를 탐지 및 분류하기 위하여, CNN과 MLP를 활용하여 효과적인 특징 생성 및 이벤트를 분류하도록 설계하였다. 이를 위하여, 본 시스템은 크게 3단계의 end-to-end 구조로 구성되며, 해당 구조는 다음과 같다. 음향 신호 자체에서 잡음이 개선된 음향 신호를 생성하는 생성자(generator), 생성자에서 생성된 음향 신호가 실제 원본 신호와 같은 신호로 생성되었는지 판별하기 위한 판별자(discriminator), 그리고 분류기(classifier)는 CNN 기반의 특징 자동 생성 및 CNN 마지막 계층의 MLP 분류기를 활용하여 음향 이벤트를 분류한다.

3.2 제안하는 시스템의 학습 과정

1) 판별자 학습 과정

a) 잡음이 없는 신호, 즉 클린(clean) 신호와 잡음이 포함된 신호를 판별자에 입력으로 사용한 후 판별자가 클린 신호에 대해서 참(ture)이라고 판단할 때까지 학습을 진행한다. 현재 판별자는 클린 신호를 명확하게 참이라

고 판별하는 부분에 초점을 맞췄다.

b) 잡음 신호를 생성자에 입력한다. 현재는 생성자에서 잡음 제거에 대한 학습이 진행되기 전인 관계로, 잡음 신호가 입력되면 잡음이 제거되지 않은 신호가 결과로 출력되며, 해당 출력 신호는 판별자에 입력되어 거짓(false)으로 판별될 때까지 판별자의 학습을 수행한다. 현 과정이 완료되면 판별자는 잡음이 있는 신호와 클린 신호를 효과적으로 구분할 수 있다.

c) 앞선 과정에 의해 잡음이 포함된 신호와 클린 신호를 안정적으로 판단하게 되는 수준까지 오면 판별자의 학습을 정지한다.

2) 생성자 학습 과정

a) 판별자의 학습 완료 후, 잡음이 포함된 신호가 입력되면 잡음이 제거된 신호가 생성되도록 생성자의 학습을 진행한다.

b) 생성자의 학습은 입력된 잡음이 포함된 신호를 생성자에 입력하여 생성된 신호가 판별자에서 잡음이 없는 클린 신호로 판별할 때까지 진행된다.

앞서 설명한 모델 학습을 위해 사용이 되는 목적 함수(object function)를 수학적으로 표현하면 Equation (1)과 같다.

$$\min_{G}\max_{D} V(D,G) = E_{x,x_c\sim p_{data}(x,x_c)}[\log D(x,x_c)] + \\ + E_{z\sim p_z(z),x_c\sim p_{data}(x_c)}[\log(1-D(G(z,x_c),x_c))] \quad (1)$$

위 식에서 우변의 첫 번째 항은 판별자가 클린 신호를 가려낼 때의 엔트로피(entropy)이며, 두 번째 항은 생성자에서 생성해낸 잡음이 제거된 신호를 판별기가 진짜 잡음이 제거되었는지(true) 아닌지(false)를 가려낼 때의 엔트로피이다. 판별자는 클린 신호를 정확하게 분류하기 위해 엔트로피를 최대화하는 방향으로 학습하고, 생성자는 판별자가 클린 신호(true)와 생성자에서 만들어 낸 잡음이 제거된 신호(false)의 차를 인식하지 못하는 상태, 즉 엔트로피가 최소화되도록 학습을 수행한다.

3) 분류기 학습 과정

a) 클린 신호만을 활용하여 이벤트 분류에 최적의 특징을 CNN 학습 과정에 의해 생성하고, CNN의 마지막 계층인 MLP에서 이벤트를 효과적으로 분류하도록 학습을 수행한다.

b) 분류기의 학습 완료 후, 시스템에 대한 검증은 잡음이 포함된 신호가 제안된 시스템에 입력되면 생성자에 의해 잡음이 제거되고, 향상된 음향 신호는 분류기에 입력이 되어 음향 이벤트를 분류하게 된다.

3.3 제안하는 시스템의 상세 구조 및 작동 흐름

1) 생성자

생성자의 계층 구조는 11개의 1차원 convolutional layer와 11개의 1차원 de-convolutional layer로 구성되어있다. 모든 계층에서 커널의 크기는 1×32로 고정, 커널의 개수는 convolutional 계층에선 {16, 32, 32, 64, 64, 128, 128, 256, 256, 512, 1024} de-convolutional 계층에선 {512, 256, 256, 128, 128, 64, 64, 32, 32, 16, 1}로 설정하였으며, convolutional 계

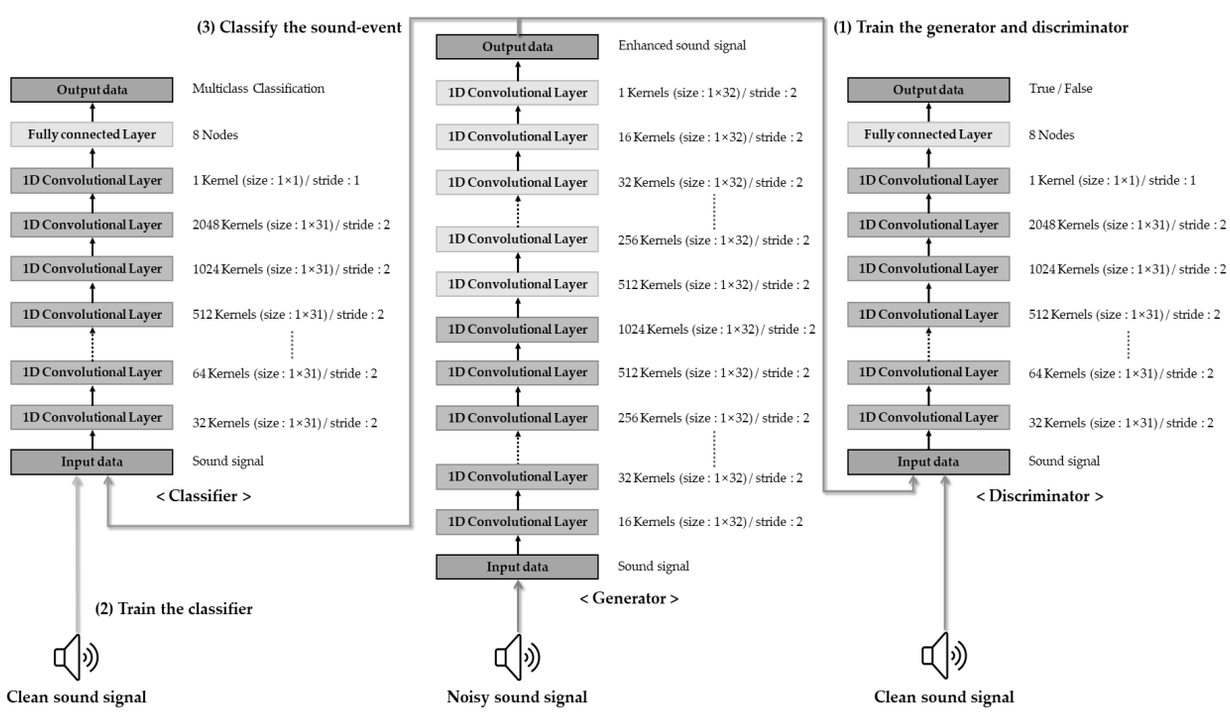


Fig. 1. Overall Structure of the Proposed Method

층과 de-convolutional 계층은 서로 대칭의 skip connection 구조를 이룬다. 생성자의 계층 구조 중 convolutional 연산에서는 인코더의 기능을 수행하며, de-convolutional 연산에서는 디코더의 기능을 수행한다. 이때, 인코더와 디코더의 연산은 기존의 잡음을 제거해주는 VAE의 잡음 제거(de-noising)와 같은 기능을 수행하게 되어, 소리 신호에 포함된 잡음을 제거하게 된다. 그리고 skip connection은 계층을 건너뛰어 정보를 전달하는 선을 이어주는 기법으로, 깊은 계층 구조에서도 vanishing gradient 문제를 완화해주고 학습 시간을 단축해주는 장점이 있다.

2) 판별자

판별자의 계층 구조는 12개의 convolutional 계층과 1개의 fully connected 계층으로 구성되며 모든 계층에서 커널의 크기는 1×31로 고정, 커널의 개수는 {32, 64, 64, 128, 128, 256, 256, 512, 512, 1024, 2048, 1}로 설정하였다. 마지막 계층의 활성함수로는 시그모이드(sigmoid) 함수를 사용하여 생성자에서 생성해낸 신호를 참과 거짓으로 구분할 수 있도록 설계하였다.

3) 분류기

마지막으로, Fig. 1의 왼쪽 모델인 분류기는 잡음 상황에서의 음향 이벤트를 분류해주는 역할을 하며, 판별자와 동일한 구조로 구성하였다. 판별자와의 차이점으로는 맨 마지막 계층의 활성 함수를 시그모이드 함수 대신 소프트맥스(softmax) 함수를 사용하여 분류 기능을 수행할 수 있도록 설계하였다.

4) 작동 흐름

전체 시스템의 작동 흐름은 먼저, 잡음이 포함된 음향 신호가 입력되면 생성자는 잡음이 제거된 향상된 신호가 생성되도록, 판별자는 생성자에서 출력된 신호를 잘 판별하도록 학습하게 된다(Fig. 1의 1번 과정). 다음으로, 클린 신호만을 이용하여 음향 신호를 잘 식별하도록 분류기를 학습한다(Fig. 1의 2번 과정). 마지막으로, 모든 모델의 학습이 완료되면, 생성자는 입력된 잡음이 포함된 음향 신호에서 잡음을 제거하여 향상된 음향 신호를 생성하고, 생성된 신호는 분류기에 입력되어 음향 신호를 명확하게 분류하게 된다(Fig. 1의 3번 과정).

4. 실험 및 결과 분석

본 논문에서 제안하는 시스템을 검증하기 위하여, 본 장에서는 철도산업과 축산업 분야의 음향 데이터를 사용하였다. 4.1절에서는 철도산업 분야의 음향 데이터를, 4.2절에서는 축산업 분야의 음향 데이터를 이용한 검증 실험을 수행하였으며, 4.3절에서는 본 논문에서 제안하는 시스템과 다른 방법론 간의 정량적·정성적 분석에 대한 고찰을 다루었다.

4.1 철도산업 분야

1) 실험 데이터

실험에 사용한 데이터는 2016년 1월 1일 대전광역시 유성구에 있는 ㈜세화 연구소에서 다음과 같이 실험 데이터를 확보하였다. 선로전환기가 작동 시 발생하는 음향 신호를 선로전환기에서 약 1m 떨어진 정중앙에서 마이크(SHURE SM137)를 이용하여 수집하였으며, 기상 상황은 약한 바람이 부는 0~6℃의 환경이었다. 본 실험에서는 Asada 등[24]이 정리한 선로전환기의 어골도(Fishbone diagram)에 기초하여 수집할 음향 클래스를 선정한 후, 해당 이벤트에 맞는 환경을 설정하여 데이터를 수집하였다. 수집한 음향 이벤트는 잡음이 섞이지 않은 총 588개이며, 정상(normal) 이벤트 150개, 선로에 자갈이 낀(gravel) 이벤트 142개, 선로에 얼음이 낀(ice-covered) 이벤트 141개, 그리고 선로전환기에 나사가 풀린(unscrewed) 이벤트 155개이다. 수집한 음향 이벤트들은 소리 파형을 보고 수작업으로 편집하였으며, 4.5~5.7초의 길이를 가졌다(44,100Hz, mono 타입).

또한, 데이터 수집 시 발생했던 새 소리(bird chirping), 헬리콥터 소리(helicopter), 바람 소리(wind), 빗소리(rain)를 선로전환기의 운행 시 발생할 수 있는 환경 잡음으로 설정하였으며, 인위적인 백색 잡음(SNR 비율: 18, 15, 12, 9, 6, 3, 0dB)을 잡음이 없는 신호에 합성하여 잡음 데이터를 준비하였다. 수집한 환경 잡음들에 대한 기초 통계표는 Table 1과 같으며, SNR 값은 수치가 작을수록 잡음의 세기가 강해짐을 의미하다.

2) 실험 및 결과 분석

본 논문에서 제안하는 향상된 음향 기반의 선로 전환 시 발생하는 소리 이벤트 분류 시스템을 검증하기 위해, Ubuntu 16.04, TensorFlow 0.12.1 컴퓨터 환경에서 제안된 모델(Fig. 1 참조)의 학습을 수행하였다.

a) 생성자·판별자 학습 및 결과 분석

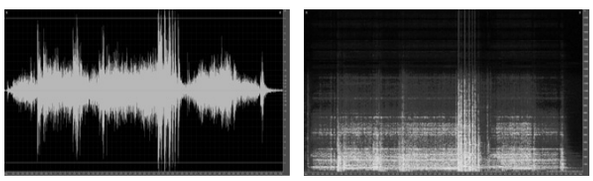
선로 전환기 이벤트 분류 시스템에 적용하기 위한 잡음이 제거된 음향을 생성하기 위하여, 먼저, 판별자와 생성자의 학습에 사용된 신호는 다음과 같다. 앞서 설명한 잡음이 섞이지 않은 데이터(588개), 백색 잡음(SNR 18, 12, 6, 0)과 환경 잡

Table 1. Basic Statistics of Environmental Noise on Railway Point Machine Sound Data

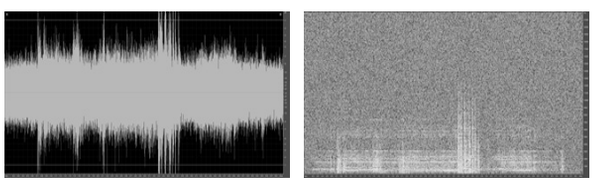
	Bird chirping	Helicopter	Wind	Rain
SNR (dB)	38.1146	14.5317	11.332	8.4212
Mean intensity (dB)	-1.5×10^{-5}	4.2×10^{-6}	-1.9×10^{-5}	-1.3×10^{-5}
Max intensity (dB)	0.0097	0.2429	0.2849	0.2560
Min intensity (dB)	-0.0103	-0.2724	-0.2559	-0.2863

음(bird chirping, helicopter, wind, rain)을 이용한, 총 8가지 잡음 신호를 학습 데이터로 선정하였다(588개 × 8가지 잡음 상황 = 4,704). 모델 학습 시, 판별자의 학습률은 0.0001, 생성자의 학습률은 0.001, batch size는 100, Leaky ReLU(alpha = 0.2), dropout 비율은 50%, Xavier 함수를 이용하여 모든 노드를 초기화하였으며, 전체 학습 횟수는 50,000회로 설정하였다.

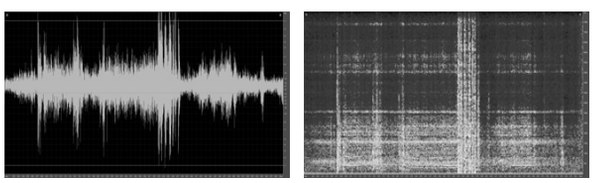
Fig. 2는 잡음이 없는 원본 음향 신호, 백색 잡음이 매우 강하게 합성된(SNR 0) 음향 신호, SNR 0가 합성된 신호를 생성자에 입력했을 때 생성된 신호에 대한 예시이다. 파형(왼쪽)과 스펙트로그램(오른쪽)을 모두 확인한 결과, 생성자에 의해 SNR 0이 합성된 신호 (c)는 잡음이 효과적으로 제거되어 잡음이 없는 원본과 유사해졌음을 시각적으로 확인할 수 있다.



(a) Raw waveform (noise-free)



(b) Noise-synthesized signal (SNR 0)



(c) Noise enhanced signal (SNR 0)

Fig. 2. Sample Waveform and Spectrogram of Railway Point Machine Sound Data

또한, 잡음이 제거된 정도를 정량적인 수치로 확인하기 위하여, PESQ(Perceptual Evaluation of Speech Quality), SSNR(Segmental SNR) 그리고 코사인 유사도를 사용하였다. PESQ는 국제 전기 통신 연합(ITU)의 P.862.2에서 권장하는 광대역 버전을 사용하여 소리 품질을 평가하는 객관적인 지표로 -0.5~4.5의 값을 가지며[25], SSNR(Segmental SNR)는 프레임 단위 신호 대 잡음 비로 0~∞의 값을 갖는다[26]. 마지막으로 코사인 유사도는 두 신호 사이의 코사인 거리를 계산하여 역수를 취한 값으로 0~1 사이의 값을 갖는다[27]. 세 지표 모두 값이 클수록 비교 대상과 유사성이 높음을 의미하며, SSNR과 코사인 유사도에 대한 수식은 다음과 같다.

$$SSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\sum_{i=Nm}^{Nm+N-1}\left(\frac{\sum_{i=1}^{N}x^2(i)}{\sum_{i=1}^{N}(x(i)-y(i))^2}\right) \quad (2)$$

$$Cosine\ Similarity = \frac{\sum_{i=1}^{n}A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

잡음 상황별 원본 음향 신호 대비, 1) 잡음 신호(Noisy)와 2) 잡음이 제거되어 향상된 음향 신호(Enhanced)에 대해 3개의 유사도 측정 지표를 사용하여 각각 측정한 결과는 Table 2와 같다. 3개의 지표 모두 향상된 음향 신호가 모든 잡음 상황에서 원본 음향 신호와의 유사도가 더 높은 것을 정량적으로 확인할 수 있었다.

b) 분류기 학습 및 결과 분석

선로 전환 시 발생하는 음향 이벤트의 클래스를 분류하기 위하여, 잡음이 없는 깨끗한 신호 588개를 기반으로 Fig. 1의 분류기를 학습하였고, 학습을 위한 파라미터는 이전 절에서

Table 2. Results of Similarity Measurement Between Noisy Signal and Enhanced Signal on Railway Sound Data

Noise Conditions	PESQ		SSNR		Cosine Similarity	
	Noisy	Enhanced	Noisy	Enhanced	Noisy	Enhanced
SNR 18	2.7846	3.1730	9.0236	10.9763	0.9576	0.9569
SNR 15	2.5745	3.0900	8.7471	10.3493	0.9552	0.9549
SNR 12	2.3140	2.9896	8.1101	9.6238	0.9406	0.9513
SNR 9	2.0221	2.8761	7.2477	8.8156	0.9114	0.9465
SNR 6	2.1151	2.7318	6.5046	7.7625	0.8940	0.9384
SNR 3	1.9723	2.5130	5.9568	6.4660	0.8717	0.9243
SNR 0	1.9223	2.2094	3.7190	4.9367	0.8459	0.8979
Bird Chirping	2.8458	3.5623	9.0416	9.7840	0.9297	0.9312
Helicopter	2.1769	2.7263	6.1316	7.1095	0.8905	0.9194
Wind	1.9684	2.0310	2.9315	4.6440	0.8519	0.8828
Rain	1.8829	3.2802	6.7130	9.5065	0.8657	0.9299
Average	2.2344	2.8348	6.7388	8.1795	0.9013	0.9303

설명한 판별자의 학습을 위해 사용한 설정값을 동일하게 사용하였다.

제안하는 시스템의 분류기 성능 테스트는 학습에 사용되지 않은 백색 잡음 데이터(SNR 18, 15, 12, 9, 6, 3, 0)와 실제 환경 잡음 데이터(bird chirping, helicopter, wind, rain)가 합성된 데이터(588개 × 11가지 잡음 상황 = 6,468)를 생성자에 입력한 후, 잡음이 제거된 음향 신호를 분류기에 전달하여 분류 성능을 확인하였다.

음향 이벤트 분류 성능을 정량적으로 평가하기 위하여 정밀도(precision), 재현율(recall), f1 score를 성능 지표로 이용하였다. Precision은 특정 음향 이벤트를 검출 후 검출된 결과에 실제 해당 이벤트가 속해 있는 비율을 의미하며, recall은 특정 음향 이벤트를 분류 시스템이 성공적으로 검출한 비율을 의미한다[28, 29]. F1 score는 precision과 recall 간의 트레이드 오프(trade off)를 고려하여 precision과 recall의 조화평균으로 계산[30]되며, 정밀도, 재현율, f1 score 모두 1에 가까울수록 좋은 성능을 의미한다. 각각의 수식은 다음과 같다. 수식에서 tp, fn, fp는 각각 true positive, false negative, false positive를 의미한다.

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

$$Recall = \frac{tp}{tp + fn} \tag{5}$$

$$F1\,score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

비교 실험으로 Choi 등[12]이 수행한 연구를 첨부하였으며, 모든 잡음 상황에서 제안하는 시스템이 우수한 성능을 보였고 특히 상대적으로 강한 잡음인 SNR 0이 합성된 상태와 비가 오는(rain) 상황에서도 안정적인 분류 성능을 확인할 수 있었다(Fig. 3 참조).

4.2 축산업 분야

1) 실험 데이터

충청남도와 경상남도에 위치한 돈사에서 25~30kg의 총 36마리의 돼지(Yorkshire, Landrace, Duroc)를 대상으로 개체로부터 약 1m의 거리에서 디지털 캠코더(JVC GR-DVL520A, Yokohama, Japan)를 사용하여 실험에 사용할 음향 신호를 수집하였다[8]. 이벤트의 레이블링을 위해 호흡기 질병으로 의심되는 돼지의 혈액을 채취한 후, 바이러스 검사와 혈청 검사를 통해 PMWS(Postweaning Multisystemic Wasting Syndrome), PRRS(Porcine Reproductive and Respiratory Syndrome), 그리고 MH(Mycoplasma Hyopneumonia)에 감염된 개체(22마리) 및 질병에 걸리지 않은 돼지를 확인한 후 음향 데이터를 수집하였다. 수집된 음향 이벤트는 총 710개이며, 돼지가 내는 일반적인(normal) 이벤트 350개와 호흡기 질병의 PMWS 이벤트 150개, PRRS 이벤트 140개, MH 이벤트 70개다. 수집한 음향 이벤트들은 소리 파형을 보고 수작업으로 편집하였으며, 0.13~2.66초의 길이를 가졌다(44,100Hz, mono 타입).

또한, 데이터 수집 시 발생했던 1~2마리가 움직일 때의 발걸음(weak footstep) 소리, 여러 마리가 움직일 때의 발걸음(strong footstep) 소리, 돼지의 안정을 위해 틀어주는 라디오(radio operation) 소리, 돈사 관리인이 사료를 주거나 배설물을 청소할 때 돈사의 문을 여닫는(door opening) 소리를 돈사 운영 시 발생할 수 있는 환경 잡음으로 설정하였으며, 인위적인 백색 잡음(SNR 비율: 18, 15, 12, 9, 6, 3, 0dB)을 잡음이 없는 신호에 합성하여 잡음 데이터를 준비하였다. 수집한 환경 잡음들에 대한 기초 통계표는 Table 3과 같다.

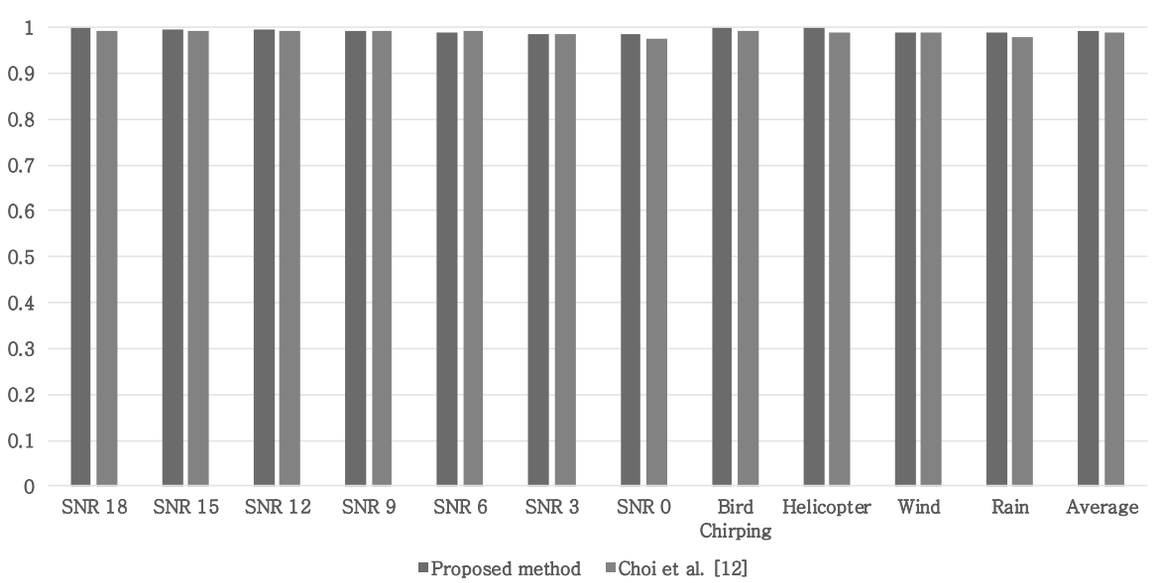


Fig. 3. F1 Score of the Proposed Method on Railway Sound Data Under Various Noise Conditions

Table 3. Basic Statistics of Environmental Noise on Porcine Sound Data

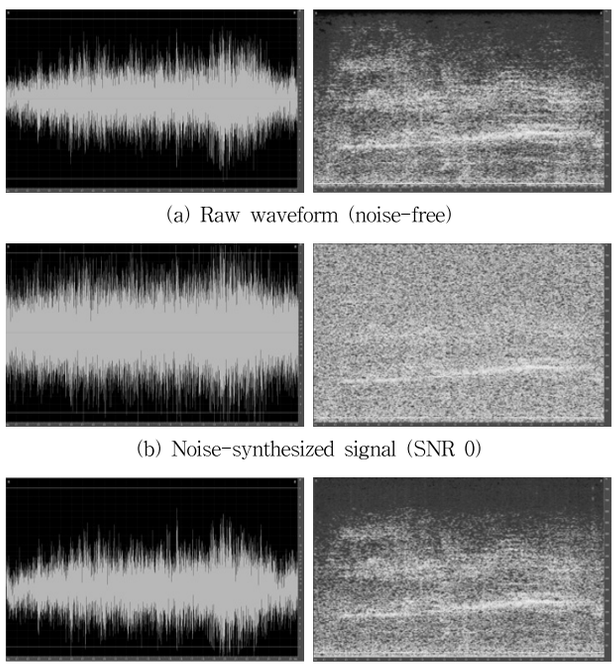
	Weak footstep	Radio Operation	Strong footstep	Door Opening
SNR (dB)	9.1172	8.7971	7.4681	4.6820
Mean intensity (dB)	2.9×10^{-5}	-9.5×10^{-6}	-1.1×10^{-5}	-3.7×10^{-5}
Max intensity (dB)	0.4594	0.3682	0.9198	0.8978
Min intensity (dB)	-0.5862	-0.3615	-0.9794	-0.8593

2) 실험 및 결과 분석

a) 생성자·판별자 학습 및 결과 분석

돼지의 호흡기 질병 분류 시스템의 잡음 제거 과정은, 선로 전환 시 취득한 소리의 잡음 제거 과정과 일관성을 유지하기 위해, 모델 학습 시 필요한 속성들의 값을 4.1절의 (a)에서 설명한 내용과 동일하게 설정하여 실험하였다. 판별자 및 생성자의 학습을 위해 사용된 데이터는 잡음이 섞이지 않은 데이터(710개)와 백색 잡음(SNR 18, 12, 6, 0)과 환경 잡음(weak footstep, radio operation, strong footstep, door opening)을 이용한, 총 8가지 잡음 신호를 학습 데이터로 선정하였으며(710개 × 8가지 잡음 상황 = 5,680). Fig. 3은 잡음이 없는 깨끗한 원본 음향 신호, SNR 0의 잡음이 합성된 음향 신호, 생성자를 통해 잡음이 제어되어 향상된 음향 신호에 대한 파형(왼쪽)과 스펙트로그램(오른쪽)에 대한 예시이다. Fig. 2와 유사하게, Fig. 4에서도 제안하는 시스템에 의해 SNR 0이 합성되었던 신호의 잡음이 제거되어 원본 신호에 유사하게 신호가 변경된 것을 시각적으로 확인할 수 있다.

또한, 잡음이 섞인 신호와 원본과의 유사도와 시스템에 의해 잡음이 제거된 신호와 원본과의 유사도를 3개의 지표를 활용하여 비교한 내용을 Table 4에 기술하였다. 제안한 시스템에 의해 생성된 향상된 음향 신호가 원본 음향 신호와의 유사도가 더 높은 것을 정량적으로 확인할 수 있다.



(a) Raw waveform (noise-free)

(b) Noise-synthesized signal (SNR 0)

(c) Noise enhanced signal (SNR 0)

Fig. 4. Sample Waveform and Spectrogram of Porcine Sound Data

b) 분류기 학습 및 결과 분석

돼지가 내는 음향 이벤트의 클래스를 분류하기 위하여, 잡음이 없는 돼지 음향 신호 4개 클래스 710개를 기반으로 Fig.

Table 4. Results of Similarity Measurement Between Noisy Signal and Enhanced Signal on Porcine Sound Data

Noise Conditions	PESQ		SSNR		Cosine Similarity	
	Noisy	Enhanced	Noisy	Enhanced	Noisy	Enhanced
SNR 18	2.9561	2.9976	11.2147	14.3308	0.9542	0.9679
SNR 15	2.7266	2.7554	9.0179	12.3902	0.9493	0.9457
SNR 12	2.6050	2.6716	8.3915	11.1479	0.9433	0.9409
SNR 9	2.3262	2.5446	6.6020	9.5986	0.9144	0.9322
SNR 6	2.1906	2.2841	5.7431	7.8063	0.8957	0.9147
SNR 3	1.7752	1.9841	3.3496	5.7769	0.8187	0.8803
SNR 0	1.4642	1.6588	1.3324	3.9195	0.7107	0.8266
Weak footstep	1.8162	2.2533	7.4384	11.214	0.8927	0.9378
Radio operation	1.7227	2.3889	7.2869	12.096	0.8968	0.9401
Strong footstep	1.4915	1.9140	6.3268	5.3288	0.8805	0.8709
Door opening	1.3452	2.0364	4.8875	9.4716	0.8121	0.9146
Average	2.0381	2.3172	6.5083	9.3710	0.8789	0.9156

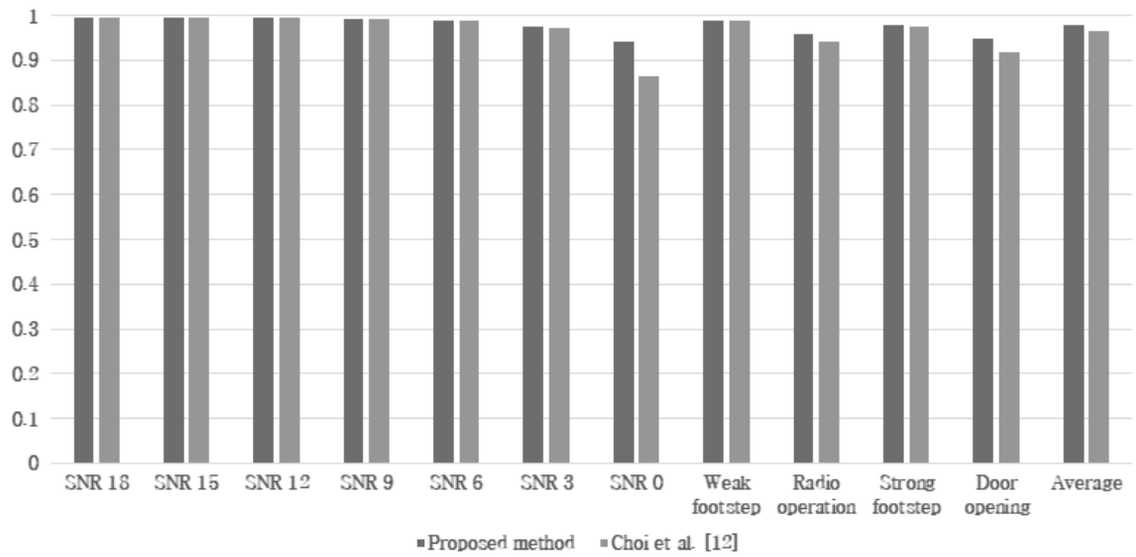


Fig. 5. F1 Score of the Proposed Method on Porcine Sound Data Under Various Noise Conditions

1의 분류기를 학습하였으며, 학습을 위한 파라미터는 이전 4.1절에서 설명한 학습 파라미터 설정값과 동일하게 사용하였다.

학습이 완료된 후 분류기의 성능을 확인하기 위하여, 학습에 사용되지 않은 백색 잡음 데이터(SNR 18, 15, 12, 9, 6, 3, 0)와 실제 환경 잡음 데이터(weak footstep, radio operation, strong footstep, door opening)가 합성된 데이터(710개 × 11가지 잡음 상황 = 7,810)를 생성기에 입력으로 사용한 후, 잡음이 제어된 돼지 음향 신호를 분류기에 전달하여 이벤트 분류 성능을 검증하였다.

잡음 상황에서의 돼지 음향 이벤트 분류 성능은 Choi 등[12]이 수행한 연구와 비교한 후 Fig. 5에 정리하였다. 다양한 잡음 상황에서 본 연구에서 제안한 시스템의 분류 성능이 우수함을 확인할 수 있다. 특히, 상대적으로 강한 잡음인 SNR 0과 door opening 상황에서도 안정적인 분류 성능을 보인다.

4.3 고찰

본 논문에서 제안한 시스템과 잡음 환경을 고려한 기존 연구들에 대한 정량적·정성적 분석을 수행하였으며, 해당 결과를 Table 5에 정리하였다. 기존의 잡음 환경을 고려한 이벤트 분류에 관한 Choi 등[12]과 Sharan 등[20]의 방법론은 잡음의 원인을 원시 음향 신호에서 직접 제거한 것이 아닌, 신호를 가공한 후 변환된 도메인에서 잡음을 제거하는 간접적인 방식이다.

본 논문에서는 이러한 한계를 보완하고 시스템을 발전시

Table 5. Quantitative and Qualitative Comparison Analysis Between the Proposed Method and Other Methods

		Proposed method	Choi et al.[12]	Sharan et al.[20]
Data generation and augmentation		○	×	×
Noise filtering		○	×	×
Data preprocessing		×	Linear transformation, Normalization	Spectrogram
Feature extraction		CNN (End-to-end)	DNS	Modulation and MFCC
Classifier			CNN	SVM
Data input type		Raw waveform	Texture image	Feature vector
Execution time	Railway	3.1696	4.3141	5.6012
	Livestock	0.8374	1.4881	2.9685
F1 score	Railway	0.9929	0.9880	0.5464
	Livestock	0.9780	0.9657	0.8230
Standard deviation	Railway	0.0052	0.0063	0.3081
	Livestock	0.0195	0.0416	0.0700

키기 위하여, SEGAN 모델을 활용하여 음향 신호 자체에서 잡음을 제거하였으며, 잡음이 제거되어 향상된 음향 신호를 분류 알고리즘과 연계하여 음향 이벤트의 분류까지도 가능하도록 end-to-end 구조로 구현하였다. 이처럼 향상된 음향 신호를 사용하는 본 시스템은 소리 이벤트 분류를 위한 전통적인 파이프라인 구조의 개별 모듈에 의한 데이터 전처리 및 데이터 변환으로 인한 정보의 손실과 해당 모듈의 특성에 따른 성능의 저하가 없는 관계로, 전통적인 구조와 비교하여 더욱 안정적인 이벤트 분류 성능을 확보하였다.

또한, 음향 이벤트 하나당 이벤트 분류를 위한 알고리즘 수행시간을 확인한 결과 철도산업의 음향은 평균 3.1696초의 수행시간을, 축산업 음향 응용에선 평균 0.8374초의 수행시간을 기록하였으며, 다른 방법론보다 상대적으로 처리속도가 빠름을 확인할 수 있다. 이때, 철도산업의 음향 신호가 평균 5.53초의 길이인 관계로 이벤트 분류에 대한 의사 실시간(pseudo real-time) 성능을 보장하지만, 축산업의 원시 음향은 철도산업 음향과 비교하면 상대적으로 짧은 0.46초의 기침 소리 신호이지만, 본 논문에서 제안하는 잡음 제거 알고리즘의 계산량이 많은 관계로 짧은 신호에서는 상대적으로 제한된 시간 내에서 잡음이 개선된 소리를 생성하는 소요 시간이 상당하다. 즉, 원시 신호의 길이가 길수록 제안하는 시스템의 처리 효율성이 상대적으로 큰 것을 실험적으로 확인할 수 있었다. 따라서 짧은 원시 음향 신호에 대해서도 효율적인 알고리즘 수행시간을 위한 연구가 추가로 필요하다.

5. 결론 및 향후 연구

본 논문에서는 실제 산업 현장에서 발생하는 잡음 환경에서도 안정적인 성능을 확보할 수 있도록 아날로그 음향 신호 자체에서 잡음을 제거한 후 이를 이용하여 음향 이벤트를 분류하는 시스템을 제안하였다. 이를 위하여, 음향 신호 자체에서 잡음을 제거하는 SEGAN 알고리즘을 활용하였으며, 원본 신호 자체에서 잡음이 제거된 음향 데이터를 생성하였다. 생성된 신호는 데이터의 변환과정 없이 CNN의 입력으로 곧바로 사용되어 소리 특징을 생성한 후, CNN의 마지막 계층인 MLP에 적용되어 이벤트를 분류하는 end-to-end 방식으로 설계하였다.

철도산업과 축산 분야의 현장에서 취득한 음향 데이터를 활용하여 제안된 시스템의 성능을 실험적으로 검증한바, 99.29%(철도산업)와 97.80%(축산업)의 f1 score를 각각 기록하였다. 또한, 제안된 시스템은 Choi 등[12]의 연구보다 음향 이벤트 하나당 수행시간 측면에서 철도산업 분야에선 1.36배, 축산업 분야에선 1.78배의 향상된 성능을 확인하였다.

제안된 시스템은 아날로그 음향 신호에 대한 직접적인 잡음 제거 및 특징 생성, 그리고 다양한 잡음 환경에서 높은 수준의 정확도를 유지하면서 경제적인 비용(저가의 소리 센서)으로 실제 산업 현장에 독립적인 혹은 기존 방법들의 보완책으로 사용될 수 있을 것으로 기대된다. 향후 연구 과제로는 본 연구에서 제안된 시스템을 실세계에서 구현 및 운용하기 위해서 알고리즘 수행시간을 더욱 단축하고, 시스템의 적용 범위를 확장하기 위하여 다양한 음향 공개 데이터베이스(open databases)를 이용한 추가 검증 작업을 진행할 예정이다.

References

[1] Y. Choi, J. Lee, D. Park, and Y. Chung, “Noise-Robust Porcine Respiratory Diseases Classification Using Texture Analysis and CNN,” *KIPS Transactions on Software and Data Engineering*, Vol.7, No.3, pp.91-98, 2018.

[2] Y. Kim, J. Sa, Y. Chung, D. Park, and S. Lee, “Resource-Efficient Pet Dog Sound Events Classification Using LSTM-FCN Based on Time-Series Data,” *Sensors*, Vol.8, No.18, pp.4019, 2018.

[3] J. Sa, Y. Choi, Y. Chung, H. Kim, D. Park, and S. Yoon, “Replacement Condition Detection of Railway Point Machines Using an Electric Current Sensor,” *Sensors*, Vol.17, pp.263, 2017.

[4] 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.

[5] J. Lee, H. Choi, D. Park, Y. Chung, H.Y. Kim, and S. Yoon, “Fault Detection and Diagnosis of Railway Point Machines by Sound Analysis,” *Sensors*, Vol.16, No.4, pp.549, 2016.

[6] Y. Choi, J. Lee, D. Park, J. Lee, Y. Chung, H.Y. Kim, and S. Yoon, “Stress Detection of Railway Point Machine Using Sound Analysis,” *KIPS Transactions on Software and Data Engineering*, Vol.5, No.9, pp.433-440, 2016.

[7] M. Guarino, P. Jans, A. Costa, J.M. Aerts, and D. Berckmans, “Field Test of Algorithm for Automatic Cough Detection in Pig Houses,” *Computers and Electronics in Agriculture*, Vol.62, No.1, pp.22-28, 2008.

[8] Y. Chung, S. Oh, J. Lee, D. Park, H. Chang, and S. Kim, “Automatic Detection and Recognition of Pig Wasting Diseases Using Sound Data in Audio Surveillance,” *Sensors*, Vol.13, No.10, pp.12929-12942, 2013.

[9] J. Lee, L. Jin, D. Park, Y. Chung, and H. Chang, “Acoustic Features for Pig Wasting Disease Detection,” *International Journal of Information Processing and Management*, Vol.6, No.1, pp.37-46, 2015.

[10] R. Zazo, T.N. Sainath, G. Simko, and C. Parada, “Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection,” *In Proceeding of Interspeech*, pp.3668-3672, 2016.

[11] H. Zhang, I. McLoughlin, Y. Song, “Robust Sound Event Recognition Using Convolutional Neural Networks,” *IEEE International Conference on Acoustics, Speech and Signal*

Processing, pp.559-563, 2015.

[12] Y. Choi, O. Atif, J. Lee, D. Park, and Y. Chung, "Noise-Robust Sound-Event Classification System with Texture Analysis," *Symmetry*, Vol.10, No.9, pp.402, 2018.

[13] Z. Zhang, J. Geiger, J. Pohjalainen, A.E.D. Mousa, W. Jin, and B. Schuller, "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments," *ACM Transactions on Intelligent Systems and Technology*, Vol.9, No.5, pp.49, 2018.

[14] Y. Choi, Y. Jung, Y. Kim, Y. Suh, and H. Kim, "An End-to-End Method for Korean Text-to-Speech Systems," *Phonetics and Speech Sciences*, Vol.10, No.1, pp.39-48, 2018.

[15] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," *In Proceedings of Interspeech*, pp.3642-3646, 2017.

[16] S. Dieleman and B. Schrauwen, "End-to-End Learning for Music Audio," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.6964-6968, 2014.

[17] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2Letter: An End-to-End ConvNet-Based Speech Recognition System," arXiv preprint arXiv:1609.03193, 2016.

[18] Y. Zhang, W. Chan, and N. Jaitly, "Very Deep Convolutional Networks for End-to-End Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4845-4849, 2017.

[19] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-Attention Based End-to-End Speech Recognition Using Multi-Task Learning," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4835-4839, 2017.

[20] R.V. Sharan and T.J. Moir, "Noise Robust Audio Surveillance Using Reduced Spectrogram Image Feature and One-against-all SVM," *Neurocomputing*, Vol.158, pp.90-99, 2015.

[21] J. Lee, Y. Choi, D. Park, and Y. Chung, "Sound Noise-Robust Porcine Wasting Diseases Detection and Classification System Using Convolutional Neural Network," *Journal of Korean Institute of Information Technology*, Vol.16, No.5, pp.1-13, 2018.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative Adversarial Nets," *In Advances in Neural Information Processing Systems*, pp.2672-2680, 2014.

[23] C. Zhang and Y. Peng, "Stacking VAE and GAN for Context-aware Text-to-Image Generation," *IEEE Fourth International Conference on Multimedia Big Data*, pp.1-5, 2018.

[24] T. Asada, C. Roberts, and T. Koseki, "An Algorithm for Improved Performance of Railway Condition Monitoring Equipment: Alternating-Current Point Machine Case Study," *Transportation Research Part C: Emerging Technologies*, Vol.30, pp.81-92, 2013.

[25] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol.2, pp.749-752, 2001.

[26] J. Hansen and B. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms," *International Conference on Spoken Language Processing*, Vol.7, pp.2819-2822, 1998.

[27] B. Shao, D. Wang, T. Li, and M. Ogihara, "Music Recommendation Based on Acoustic Features and User Access Patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.17, No.8, pp.1602-1611, 2009.

[28] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," 3rd ed., Morgan Kaufman, San Francisco, CA, USA, 2012.

[29] S. Theodoridis and K. Koutroumbas, "Pattern Recognition," 4th ed., Academic Press: Kidlington, Oxford, UK, 2009.

[30] D.M. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness Markedness and Correlation," *Journal of Machine Learning Technologies*, Vol.2, No.1, pp.37-63, 2011.

최 용 주

http://orcid.org/0000-0003-4661-6196
e-mail : aaa928@korea.ac.kr
2017년 고려대학교 컴퓨터정보학과(학사)
2019년 고려대학교 컴퓨터정보학과(석사)
2019년～현 재 CJ대한통운 정보전략팀 연구원
관심분야 : 인공지능, 융합IT, 기계학습, 딥러닝

이 종 욱

https://orcid.org/0000-0002-2077-4850
e-mail : eastwest9@korea.ac.kr
2002년 고려대학교 전산학과(학사)
2005년 고려대학교 전산학과(석사)
2014년 고려대학교 전산학과(박사)
2014년～현 재 고려대학교 컴퓨터융합소프트웨어학과 초빙교수
관심분야 : 딥러닝, 데이터마이닝, 융합 IT, 음향분석

박 대 희

https://orcid.org/0000-0003-4726-4508
e-mail : dhpark@korea.ac.kr
1982년 고려대학교 수학과(학사)
1984년 고려대학교 수학과(석사)
1989년 플로리다 주립대학 전산학과(석사)
1992년 플로리다 주립대학 전산학과(박사)
1993년~현 재 고려대학교 컴퓨터융합소프트웨어학과 교수
관심분야 : 빅데이터, 데이터마이닝, 인공지능, 융합 IT

정 용 화

https://orcid.org/0000-0001-6539-167X
e-mail : ychungy@korea.ac.kr
1984년 한양대학교 전자통신공학과(학사)
1986년 한양대학교 전자통신공학과(석사)
1997년 U. of Southern California(박사)
1986년~2003년 한국전자통신연구원 생체인식기술연구팀(팀장)
2003년~현 재 고려대학교 컴퓨터융합소프트웨어학과 교수
관심분야 : 병렬처리, 영상처리, 융합 IT