JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 2, pp. 339-346, Feb. 2025

Check for updates

음성 및 오디오 신호 분류를 위한 딥러닝 기반 서브밴드 분석

김 광 기*
나사렛대학교 IT인공지능학부 부교수

Deep Learning-Based Subband Analysis for Speech and Audio Signal Classification

Kwangki Kim*

Associate Professor, Department of IT Artificial Intelligence, Nazarene University, Cheonan 31172, Korea

[요 약]

본 논문은 통합 코딩 시스템에서 음성 신호와 오디오 신호를 분류하는 딥러닝 기반 서브밴드 분석 방법을 제안한다. 기존 G.718 분류기는 Modified Discrete Cosine Transform (MDCT) 에너지 비율과 같은 단일 매개변수에 의존하여 일정 수준의 분류 성능만을 지닌다는 한계를 가진다. 또한, 서브밴드 기반 이진 결정 트리는 고정된 결정 경계를 사용하여 신호의 특성을 분류하여 비선형적인 신호 특성이나 급격하게 변하는 신호를 모델링하는 데 어려움이 있다. 이러한 기존 분류기의 문제를 해결하기 위해 본 연구에서는 컨볼루션 신경망(CNN; convolutional neural network과 순환 신경망(RNN; recurrent neural network)을 활용한 딥러닝 기반 서브밴드 분석 기법을 적용한 새로운 분류기를 제안한다. 제안된 방법은 서브밴드 분석을 통해 다양한 주파수 대역에서 관련된 특징을 추출함으로써 딥러닝 모델이 음성 및 오디오 신호 간의 복잡한 관계와 패턴을 효과적으로 학습할 수 있도록 설계되었다. 실험 결과, 제안된 분류기는 G.718 및 서브밴드 기반 이진 결정 트리 모델에 비해 분류 정확도 측면에서 큰 성능 향상을 보였다.

[Abstract]

This study proposes a deep learning-based subband analysis method for classifying speech and audio signals in integrated coding systems. The existing G.718 classifier relies on a single parameter, namely the Modified Discrete Cosine Transform (MDCT) energy ratio, which constrains its classification performance. Furthermore, the subband-based binary decision tree employs fixed decision boundaries to classify signal characteristics, making it inadequate for modeling nonlinear or rapidly varying signal features. To overcome these limitations, this research introduces a novel classifier that incorporates deep learning-based subband analysis utilizing Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The proposed approach extracts salient features from multiple frequency bands through subband analysis, allowing the deep learning model to effectively capture the intricate relationships and patterns between speech and audio signals. Experimental results reveal that the proposed classifier achieves a substantial enhancement in classification accuracy compared to the G.718 classifier and subband-based binary decision tree models.

색인어 : 음성/오디오 분류기, 컨볼루션 신경망, 순환 신경망, 서브밴드 분석, 특징 추출

Keyword : Speech/Audio Classifier, Convolutional Neural Network, Recurrent Neural Network, Subband Analysis, Feature Extraction

http://dx.doi.org/10.9728/dcs.2025.26.2.339

Received 10 December 2024; **Revised** 14 January 2025
Accepted 24 January 2025

***Corresponding Author; Kwangki Kim**

Tel: +82-41-570-1434
E-mail: k2kim@kornu.ac.kr

Ⅰ. 서 론

최근 통합 음성 및 오디오 코딩 기술은 다양한 응용 분야에서 중요한 역할을 차지하고 있다. 특히, 음성 인식, 오디오 압축, 멀티미디어 콘텐츠 분석 등에서 이러한 기술들이 핵심적인 역할을 하면서 음성 및 오디오 신호의 정확하고 효율적인 분류는 그 어느 때보다 중요해졌다. 예를 들어, AI (Artificial Intelligence) 가상 비서 시스템에서는 스마트 기기와의 원활한 상호작용을 위해 사용자의 음성 명령을 정확하게 분류하고 해석하는 기술이 필요하다. 또한, 멀티미디어 콘텐츠 분석에서는 오디오 신호의 정확한 분류가 자동 태그 지정, 콘텐츠 검색, 파일 분할 등 다양한 작업을 지원하는 데 필수적이다. 이와 같은 응용 분야에서 음성 및 오디오 신호의 정확한 분류는 사용자 경험을 크게 향상시키며, 특히 오디오 압축, 검색, 음성 명령 인식 등과 같은 애플리케이션의 성능에 중대한 영향을 미친다.

G.718 코딩 시스템은 음성 및 오디오 신호 분류 기술 중의 하나로 오랫동안 널리 사용되어 왔다. G.718의 음성/오디오 분류기는 주로 Modified Discrete Cosine Transform (MDCT) 에너지 비율과 같은 간단한 매개변수에 의존하여 신호를 분류한다[1],[2]. 그러나 이러한 시스템은 신호의 스펙트럼 특성이나 복잡한 주파수 대역에 대한 세밀한 분석을 제공하지 못하고, 단일 매개변수만을 기반으로 분류를 수행하기 때문에 복잡한 스펙트럼 특성을 가진 신호에서 높은 오분류율을 보인다. 특히, G.718 시스템은 다양한 콘텐츠 유형에 대해 일관된 분류 정확도를 유지하는 데 어려움을 겪으며, 이로 인해 전체적인 오디오 품질이 저하될 수 있다. 즉, G.718 코딩 방식은 단순화된 특성으로 인해 고속으로 변화하는 신호나 복잡한 스펙트럼을 잘 처리하지 못하며, 결과적으로 멀티미디어 콘텐츠 분석에서의 정확도에 심각한 영향을 미친다.

이러한 G.718의 한계를 극복하기 위해 서브밴드 기반의 이진 결정 트리 분류기가 제안되었다[3]-[5]. 서브밴드 기반 방식은 입력 신호를 여러 개의 주파수 대역으로 나눈 뒤, 각 대역에서 중요한 특징을 추출하여 이를 바탕으로 신호를 분류하는 방법이다. 이러한 방식은 특정 주파수 대역에서 신호의 특성을 추출할 수 있어 단일 주파수 대역에서 분류할 때보다 더 높은 정확도를 제공할 수 있다. 그러나 서브밴드 기반의 결정 트리 분류기는 고정된 결정 경계를 사용하여 신호의 특성을 분류한다. 이 고정된 경계는 비선형적인 신호 특성이나 급격하게 변하는 신호를 모델링하는 데 한계가 있다. 특히, 스펙트럼 특성이 중첩되거나 빠르게 변화하는 신호에 대해서는 적절하게 대응하지 못하여 이는 결국 정확한 분류를 방해하게 된다.

이를 해결하기 위해 본 논문에서는 딥러닝 기반의 서브밴드 분석 방식을 제안한다. 딥러닝 모델은 특히 복잡한 패턴을 학습하고 고차원 데이터의 의미를 효과적으로 캡처하는 데 매우 강력하다. 본 연구에서는 컨볼루션 신경망(CNN; convolutional neural network) 과 순환 신경망(RNN; recurrent neural network) 을 결합하여 음성 및 오디오 신호의 공간적 및 시간적 특성을 모두 포착할 수 있는 모델을 제안한다. CNN은 스펙트로그램에서 주파수 대역 내의 공간적 특징을 추출하는 데 뛰어난 성능을 발휘하며, RNN은 신호의 시간적 의존성을 학습하여 동적 특성을 보다 잘 반영할 수 있게 한다[6],[7]. 이를 통해 제안된 모델은 기존의 G.718 방식이나 서브밴드 기반의 결정 트리 분류기보다 우수한 분류 정확도를 제공하며, 오디오 코딩 시스템에서 신호의 정확한 분류를 가능하게 한다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 음성 및 오디오 분류기를 살펴보고, 3장에서 제안하는 딥러닝 서브밴드 분석 기반 분류기를 설명한다. 4장에서는 실험을 통해 제안된 방법의 성능을 확인하며, 5장에서 결론을 제시한다.

Ⅱ. 기존의 음성/오디오 분류기

음성 및 오디오 신호 분류는 오랜 기간 다양한 응용 분야에서 중요한 연구 주제로 다루어져 왔다. 기존 연구들은 음성 신호와 음악 신호를 구분하기 위해 주로 특징 기반 접근법을 채택하였으며, 계산 효율성과 정확도를 균형 있게 유지하는 것이 주요 과제로 인식되어 왔다. 본 장에서는 전통적으로 활용된 G.718음성/오디오 분류기와 서브밴드 기반 이진 결정 트리 방식의 분류기를 살펴보기로 한다.

2-1 G.718 음성/오디오 분류기

G.718 음성/오디오 분류기는 ITU-T(International Telecommunication Union Telecommunication Standardization Sector)에서 제안한 음성 코딩 시스템의 일부로, 음성 신호와 일반 오디오 신호를 구분하는 간단하면서도 계산 효율적인 방법을 제공한다[1],[2]. 이 분류기는 MDCT 계수의 에너지 비율을 기반으로 신호를 분류한다. MDCT는 시간 영역의 신호를 주파수 영역으로 변환하는 기술로 주로 음성 및 오디오 신호처리에서 널리 사용된다[8]. G.718 분류기는 입력 신호를 MDCT를 통해 주파수 도메인으로 변환한 후, 특정 주파수 대역의 에너지 비율을 계산하여 이를 바탕으로 신호가 음성인지 일반 오디오인지 구분한다. 음성 신호는 일반적으로 저주파 대역에서 에너지가 집중되는 경향이 있어, 이 특성을 기반으로 효율적인 분류가 가능하다. 반면, 일반 오디오 신호는 더 넓은 주파수 대역에 걸쳐 에너지가 분포하기 때문에 이를 음성과 구분할 수 있다.

이 방식은 계산 효율성이 뛰어나, 실시간 처리 요구사항이 있는 시스템에서 유리하다. G.718은 복잡한 신호 분석이 필요하지 않으며, 단일 특징 매개변수를 이용해 빠르고 간단하게 신호를 분류할 수 있다는 장점이 있다. 예를 들어, 통합 코딩 시스템에서는 처리 속도와 실시간 요구가 중요하기 때문에 G.718 방식이 적합하다. 이 분류기의 주요 장점은 신호 분

석을 위한 복잡한 계산을 최소화하고, 단순한 에너지 비율만으로도 비교적 좋은 분류 성능을 보이는 점이다. 따라서 음성 통화나 방송 시스템, 실시간 코딩 시스템에서 사용되기에 적합하다.

그럼에도 불구하고 G.718 분류기는 몇 가지 중요한 한계를 지닌다[9],[10]. 가장 큰 문제는 복잡한 오디오 신호에 대한 낮은 정확도다. G.718은 단일 특징 매개변수인 MDCT 계수의 에너지 비율만을 사용하여 신호를 구분하는데, 이는 스펙트럼 특성이 중첩되는 복잡한 신호에 대해 오분류를 초래할 수 있다. 예를 들어, 음악 신호는 고조파 구조와 과도 특성을 동시에 가지고 있어, 이를 단순히 에너지 비율만으로 구분하는 것은 매우 어렵다. 또한 음성 신호는 성도(vocal tract)의 공명에 의해 예측 가능한 고조파 패턴을 보이지만, MDCT 계수만으로는 이러한 고조파 구조를 정확히 반영하기 어렵다. 이로 인해 음악과 음성 신호를 명확하게 구분하는 데 한계가 있다. 특히, 복잡한 오디오 신호가 포함된 환경에서는 G.718 분류기의 성능이 크게 저하될 수 있다.

이러한 문제는 통합 코딩 시스템에서 성능 저하를 초래할 수 있다. 예를 들어, 잘못 분류된 신호는 적절한 코덱 설정을 받지 못하게 되어, 음질 열화가 발생하거나 불필요한 데이터 전송이 이루어질 수 있다. 따라서 G.718 분류기의 성능 한계는 실시간 오디오 및 음성 처리가 중요한 응용 프로그램에서 큰 단점으로 작용할 수 있다. 이러한 문제를 해결하려면, 신호의 더 다양한 특성을 반영할 수 있는 방법론이 필요하다.

2-2 서브밴드 기반 이진 결정 트리 방식

서브밴드 기반 이진 결정 트리 방식은 G.718 분류기의 한계를 보완하기 위해 도입된 접근법으로 입력 신호를 여러 주파수 서브밴드로 분할한 후 각 구간에서 추출된 특징을 기반으로 결정 트리를 적용하여 신호를 분류한다[5],[11]. 이 방식은 G.718의 MDCT 계수만을 사용한 단일 특징 분석에 비해 분류 정확도를 개선하였으며, 신호의 세부적인 특성을 좀 더 세밀하게 반영할 수 있다는 장점이 있다. 서브밴드 방식은 신호를 여러 주파수 대역으로 나누어 처리함으로써, 신호의 시간적 및 주파수적 특성을 보다 잘 반영할 수 있다. 이를 통해 G.718 방식의 제한적 구분 정확도를 극복하고, 보다 다양한 신호 환경에서도 성능을 향상시킬 수 있게 된다.

하지만 서브밴드 기반 이진 결정 트리 방식에도 몇 가지 중요한 단점이 존재한다. 첫째, 결정 트리의 결정 경계가 고정적이라는 특성으로 인해 신호의 복잡하고 비선형적인 특성을 효과적으로 처리하는 데 한계가 있다[12],[13]. 이는 특히 스펙트럼 특성이 중첩되거나 빠르게 변화하는 신호에 대해 분류 성능을 저하시킬 수 있다. 예를 들어, 음악 신호나 잡음이 섞인 신호는 빠르게 변화하는 스펙트럼과 고차원적인 패턴을 가지고 있어, 고정된 결정 경계만으로 이를 정확히 구분하기 어려운 경우가 많다 신호의 다채로운 변화를 반영해야 하는 실제 환경에서 분류 성능 저하를 초래할 수 있다.

둘째, 서브밴드 기반 이진 결정 트리 방식은 선택된 특징에 크게 의존한다. 신호의 에너지 비율이나 스펙트럼 엔트로피와 같은 특징은 유용하지만, 신호의 시간적 및 스펙트럼적 의존성을 충분히 반영하지 못할 수 있다[14],[15]. 예를 들어, 신호가 변화하는 패턴을 반영하기 위해서는 더 복잡한 특징이나 시간-주파수 분석이 필요하지만, 단일 특징 기반의 분류는 이를 충분히 반영하지 못한다. 또한, 결정 트리는 과적합(overfitting) 문제에 취약한 경향이 있어, 학습 데이터에 너무 특화된 모델이 만들어질 수 있다. 이는 데이터에 노이즈가 섞이거나 변동성이 클 경우, 모델의 일반화 성능이 크게 저하될 수 있음을 의미한다 . 이러한 문제스펙트럼 콘텐츠와 시간적 구조를 가지는 실제 환경에서 분류기의 효과를 감소시키는 주요 원인으로 작용한다.

결국 서브밴드 기반 이진 결정 트리 방식은 G.718의 한계를 일부 극복하고 분류 정확도를 개선했지만, 여전히 신호의 복잡한 특성을 처리하는 데 있어서 제한적인 요소가 존재한다. 이 방식은 주파수 대역을 나누어 더 많은 정보를 처리할 수 있는 장점이 있지만, 신호의 비선형적이고 빠르게 변화하는 특성을 정확하게 반영하는 데는 한계가 있어, 실제 환경에서는 여전히 오분류가 발생할 수 있다.

2-3 딥러닝 기반 음성/오디오 분류기의 필요성

앞에서 살펴본 바와 같이 G.718 음성/오디오 분류기와 서브밴드 기반 이진 결정 트리 방식은 모두 특정 한계점을 가지고 있다. G.718은 MDCT 계수의 에너지 비율만을 이용하여 음성과 오디오 신호를 구분하지만 복잡한 신호에서 높은 오분류율을 보이며, 신호의 다양한 특성을 충분히 반영하지 못한다. 반면, 서브밴드 기반 이진 결정 트리는 주파수 대역을 나누어 보다 세밀한 분류를 시도하지만, 고정된 결정 경계와 선택된 특징에 의존하여 신호의 비선형적이고 빠르게 변화하는 특성을 처리하는 데 한계가 있다. 이러한 문제들은 실제 환경에서의 복잡한 신호를 정확히 분류하기 어렵게 만든다. 따라서, 신호의 비선형적 관계와 시간적-주파수적 특성을 보다 정교하게 학습할 수 있는 딥러닝 기반의 음성/오디오 분류기가 필요하며, 본 논문에서는 딥러닝 기반 서브밴드 분석 기법을 적용한 새로운 음성/오디오 분류기를 제안한다.

Ⅲ. 제안된 딥러닝 서브밴드 분석 기반 음성/오디오 분류기

본 논문에서는 음성 및 오디오 신호를 정확히 분류하기 위해 딥러닝 서브밴드 분석 기반 음성/오디오 분류기를 제안한다. 입력 신호를 서브밴드로 분할하여 각 주파수 대역의 특징을 독립적으로 분석하고, 이를 CNN과 RNN으로 구성된 딥러닝 모델에 입력하여 복잡한 관계를 학습하도록 설계하였다.

3-1 서브밴드 분석

서브밴드 분석은 신호의 주파수 특성을 다양한 주파수 대역에서 상세하게 표현하며, 이를 딥러닝 모델에 입력하여 신호의 복잡한 관계를 효과적으로 학습할 수 있도록 한다[16]. 제안된 방법에서는 입력 신호를 필터 뱅크를 사용하여 여러 주파수 서브밴드로 분할하고, 각 서브밴드는 서로 다른 주파수 범위의 신호 정보를 포함한다. 이러한 서브밴드 분석은 신호의 스펙트럼 콘텐츠를 보다 정확하게 캡처하여, 복잡한 신호에 대한 분류 성능을 향상시키는 데 기여한다.

입력 신호 $x(t)$ 를 서브밴드로 분할한 i번째 서브밴드 신호 $y_i(t)$ 는 다음과 같이 수학적으로 표현된다.

$$y_i(t) = x(t) * h_i(t) \text{ for } 0 \le i \le N-1 \quad (1)$$

여기에서 $h_i(t)$는 i번째 서브밴드의 필터함수, *는 컨볼루션 연산, N은 서브밴드 수를 나타낸다. 각 서브밴드는 특정 주파수 범위에 해당하는 정보를 포함하고, 신호의 주파수 특성을 독립적으로 분석할 수 있도록 한다. 예를 들어, 음성 신호는 일반적으로 낮은 주파수 대역에서 에너지가 집중되며, 음악 신호는 더 균일한 주파수 스펙트럼을 가진다. 이 방식은 딥러닝 모델이 서브밴드별로 신호의 복잡한 특성을 학습하게 하여, 음성 및 오디오 신호의 구분 정확도를 개선하는 데 중요한 역할을 한다. 특히, CNN은 지역적인 특징을 추출하는 데 효과적이며, RNN은 시간적 의존성을 잘 처리할 수 있어, 두 네트워크의 결합을 통해 더 정교한 분석이 가능하다.

3-2 특징 파라미터

제안된 딥러닝 기반 음성/오디오 분류기의 특징 추출 과정은 서브밴드 분석을 통해 신호의 다양한 특성을 반영하며, 각 서브밴드에서 주요 특징들을 추출하여 CNN, RNN, 그리고 완전 연결 계층을 통해 신호를 분류한다. 각 서브밴드에서 추출된 주요 특징은 스펙트로그램과 시간적 특징으로 나눌 수 있다. 이러한 특징들은 신호의 공간적 및 시간적 특성을 효과적으로 캡처하는 데 중요한 역할을 한다.

1) 스펙트로그램

스펙트로그램은 신호의 시간에 따른 주파수 콘텐츠를 시각적으로 표현한 것으로, CNN 계층의 입력으로 사용된다. 스펙트로그램은 신호의 주파수 특성을 시간적으로 분석할 수 있는 유용한 도구이다. 다음과 같은 과정을 통해 스펙트로그램을 계산한다. 각 서브밴드 신호 $y_i(t)$를 윈도우 크기 512샘플로 분할하고, Hanning 윈도우를 적용한 후 단시간 푸리에 변환(STFT; short time Fourier transform)을 적용한다.

$$y_{i,w}(t) = y_i(t) \cdot w(t) \quad (2)$$

$$Y_i(f,t) = STFT(y_{i,w}(t)) \quad (3)$$

여기에서 $y_{i,w}(t)$는 i 번째 서브밴드신호에 Hanning 윈도우 $w(t)$를 적용한 신호이며, $Y_i(f,t)$는 시간-주파수 스펙트럼이다. 시간-주파수 스펙트럼의 크기를 계산한 후 로그를 취하여 최종 스펙트로그램 $S_i(f,t)$을 구할 수 있다.

$$S_i(f,t) = \log(|Y_i(f,t)|^2) \quad (4)$$

이렇게 계산된 스펙트로그램을 통해 신호의 주파수 스펙트럼을 시간에 따라 분석할 수 있으며, CNN에 적용하여 신호의 공간적 패턴을 효과적으로 추출할 수 있다.

2) 시간적 특징

시간적 특징 추출은 음성 및 오디오 신호의 시간적인 변화를 포착하여 복잡한 패턴을 학습하는 데 중점을 두고 있다. 신호의 시간적 의존성을 학습하기 위해 본 연구에서는 CNN과 Long Short-Term Memory(LSTM)를 결합하여 음성 및 오디오 신호의 시간적 특징을 추출하고 학습하도록 한다[17].

입력 신호는 STFT를 통해 시간-주파수 스펙트럼으로 변환되고 이를 이용하여 생성된 스펙트로그램은 시간 축과 주파수 축의 정보를 모두 포함하고 있으므로 신호의 패턴을 분석할 수 있다. 시간 축에서의 정보 손실을 최소화하기 위해 CNN 계층 설계 시 Stride 값을 1로 설정하고 Average Pooling을 사용하여 시간 축의 연속성을 유지하도록 한다. 이를 통해 각 시간 프레임이 독립적으로 유지되며, 이후 LSTM에서 시계열 데이터의 시간적 의전송을 학습하는 기반을 제공한다.

LSTM은 CNN에서 추출된 특징 맵을 입력으로 받아 시간적 패턴을 학습한다. 여기에서 CNN 출력은 시간 축을 중심으로 Flatten 과정을 통해 시퀀스 형태로 변환하여 LSTM에 입력된다. 이를 통해 LSTM은 CNN에서 축약된 시간적 패턴을 학습하여 음성 신호에서는 억양 변화, 음소 변화, 속도 등을, 음악 신호에서는 리듬, 박자, 주파수 변화 등의 시간적 의존성을 반영하는 특징을 추출할 수 있다. 또한, LSTM의 단기 및 장기 의존성 처리 능력은 신호의 과거와 현재의 변화를 동시에 반영할 수 있게 한다. 예를 들어, 음성 신호에서는 음소의 장기적 변화와 단기적 변화를 동시에 학습한다. 음악 신호에서도 리듬의 장기적 흐름을 추적하면서 빠르게 변하는 트랜지언트 신호의 특성을 포착할 수 있다.

3-3 음성/오디오 분류를 위한 딥러닝 구조

제안된 딥러닝 구조는 CNN, LSTM, 완전 연결 (FC; fully connected) 계층으로 구성된다. 각 계층은 신호의 다양한 특성을 추출하고, 이들을 종합하여 최종 분류 결과를 생성하는 데 중요한 역할을 한다.

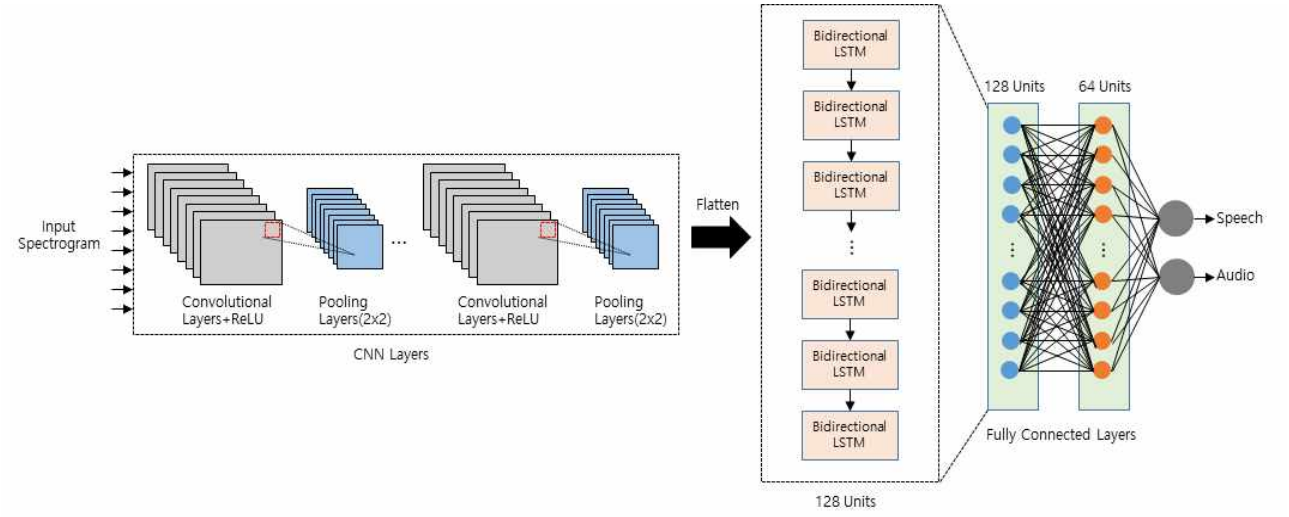

그림 1. 제안된 딥러닝 구조
Fig. 1. Proposed deep learning structure

CNN 계층은 각 서브밴드의 스펙트로그램에서 공간적 특징을 추출한다. 이 계층은 신호의 조화파(harmonics) 및 과도(transient)와 같은 지역적 패턴을 캡처하는 데 유용하며, 음성 신호와 오디오 신호를 구별하는 데 중요한 역할을 한다. CNN 계층의 입력 스펙트로그램은 (T, F, C) 형태로 변환되며, 여기에서 T는 시간 프레임, F는 주파수 대역, C는 채널로 단일 채널을 사용한다. 입력된 스펙트로그램은 3×3 필터와 ReLU 활성화 함수를 사용하여 특징 맵을 생성한다. 2×2 Average Pooling을 통해 시간 및 주파수 축을 축소하여 연산 효율성을 높이며, 채널 수는 64에서 128로 확장된다. CNN 계층의 출력은 다음과 같이 표현되며, 출력은 $(T/2, F/2, 128)$의 형태를 갖는다.

$$F_{CNN} = Conv(S) \tag{5}$$

여기에서 $Conv$는 입력 스펙트로그램 S 에 대해 적용된 컨볼루션 연산을 나타내며, 이 연산을 통해 입력 데이터에서 중요한 지역적 패턴을 추출할 수 있다.

RNN 계열 모델인 LSTM 계층은 CNN에서 추출된 특징 간의 시간적 관계를 학습한다. 특히, 음성/오디오 신호와 같은 시간에 따른 의존성을 모델링하는 데 유리하다. LSTM은 이전의 정보를 기억하고 이를 바탕으로 현재 상태를 예측할 수 있기 때문에 음성/오디오 신호에서 중요한 시간적 역학을 포착하는 데 매우 효과적이다. LSTM 계층의 입력은 CNN 계층의 출력을 시계열 데이터로 학습할 수 있도록 시간 축 중심으로 Flatten하여 $(T/2, F/2\times128)$ 형태로 재구성하여 사용하며, 양방향 LSTM을 사용하여 각 시간 프레임의 이전 및 이후 정보를 모두 학습한다. 입력 게이트, 망각 게이트, 출력 게이트를 통해 중요 정보를 선택적으로 학습하고 유닛 수는 128로 설정된다. LSTM의 출력은 다음과 같이 표현되며, 출력은 $(T/2, 256)$의 형태를 갖는다.

$$F_{LSTM} = LSTM(F_{CNN, Flatten}) \tag{6}$$

여기에서 $LSTM$은 CNN 계층에서 추출된 특징 F_{CNN}을 Flatten한 $F_{CNN,Flatten}$ 에 대해 적용된 순환 연산을 나타낸다. 이 연산을 통해 시간적인 의존성을 모델링하고, 신호의 시간적 변화를 효과적으로 학습할 수 있다.

FC 계층은 LSTM 계층에서 추출된 시간적 특징을 기반으로 최종 분류 결과를 생성한다. FC 계층은 모든 뉴런을 연결하여 각 출력 클래스에 대한 확률 값을 계산하는 역할을 한다. FC 계층에서는 2개의 은닉층을 사용하며, 첫번째 은닉층은 128개의 노드를 사용하고 두번째 은닉층은 64개의 노드를 사용한다. 최종 출력은 다음과 같이 표현되며, 출력 결과는 음성 또는 오디오로 구분된다.

$$P = FC(F_{LSTM}) \tag{7}$$

여기에서 FC는 LSTM 계층의 출력을 기반으로 최종 분류를 수행하는 완전 연결 계층을 나타낸다. 이 계층은 신경망이 학습한 모든 특성을 종합하여 음성 신호 또는 오디오 신호의 종류를 결정하는 역할을 한다.

이와 같이 제안된 모델은 CNN을 통해 신호의 공간적 특성을 추출하고, LSTM을 통해 시간적 특성을 학습하며, FC 계층을 통해 최종적인 분류 결과를 생성하는 다단계 처리 구조로 구성되있다. 그림 1은 제안된 전체 딥러닝 구조를 나타낸다.

3-4 딥러닝 모델 학습

제안된 모델은 교차 엔트로피 손실 함수와 Adam 옵티마이저를 사용하여 학습되었다. 교차 엔트로피 손실 함수는 분류 문제에서 모델의 예측 성능을 평가하는 데 널리 사용되는 손실 함수로 실제 레이블과 모델이 예측한 확률 사이의 차이

를 최소화하는 방향으로 학습을 진행한다[18]. 교차 엔트로피 손실 함수는 다음과 같이 정의된다.

$$L = -\frac{1}{M}\sum_{i=1}^{M} y_i \log(\hat{y_i}) \quad (8)$$

여기에서 y_i는 i번째 샘플에 대한 실제 클래스 레이블로 음성 또는 오디오 클래스를 나타내며, $\hat{y_i}$는 i번째 샘플에 대해 모델이 예측한 클래스에 대한 확률을 나타낸다. 또한, M은 모델이 학습하거나 테스트하는 총 데이터의 개수를 나타낸다.

이 식은 모델이 예측한 확률이 실제 레이블에 가까워질수록 손실 값이 작아짐을 나타내며, 이를 통해 모델은 예측값을 실제 값에 맞추는 방향으로 학습된다.

학습 과정에서 Adam 옵티마이저는 모델 가중치 업데이트를 담당하며, 검증 성능에 따라 학습률을 동적으로 조정하여 최적화의 효율성을 높인다[19]. Adam 옵티마이저는 기존의 확률적 경사 하강법(SGD; stochastic gradient descent)에 비해 학습률의 조정이 더 효율적이며, 이를 통해 모델이 빠르고 안정적으로 수렴할 수 있도록 한다. Adam 옵티마이저는 각 매개변수에 대한 학습률을 개별적으로 조정하면서 모델 가중치를 갱신한다.

Ⅳ. 실험 및 성능 평가

4-1 실험 데이터

제안된 딥러닝 기반 음성/오디오 분류기의 성능을 검증하기 위해 두 가지 주요 신호 유형, 즉 음성 신호와 오디오 신호를 포함하는 데이터셋을 사용하였다. 데이터셋에 포함된 음성 신호는 한국어를 사용하는 남성 10명과 여성 10명, 총 20명의 화자들로부터 수집되었으며, 전체 140분 분량으로 구성되었다. 이 음성 데이터는 한국어 발화, 일상 대화 등을 포함하여 음성 인식과 구분이 필요한 실제 환경을 잘 반영한다. 오디오 신호는 한국 악기 연주곡 총 182분의 데이터를 사용하였다. 이 데이터는 주로 악기 연주로 구성되어 있으며, 보컬 요소는 제외되었다. 음성 데이터와 오디오 데이터는 각각 16 kHz, 44.1 kHz로 샘플링되어 있어 데이터 균일화를 위해 오디오 신호를 16 kHz로 다운 샘플링하여 사용하였다. 또한, 음성 데이터와 오디오 데이터의 크기는 수집 및 녹음 환경에 따라서 다를 수 있기 때문에 정규화 과정을 거친 후 사용하였다. 데이터셋의 70%는 딥러닝 모델 학습에, 나머지 30%는 테스트 데이터로 사용되었다.

4-2 성능 평가 결과

제안된 딥러닝 기반 음성/오디오 분류기의 성능 평가는 10-폴드 교차 검증 방법을 사용하였다[20]. 즉, 음성/오디오 데이터셋을 10개의 부분으로 나누어 각각을 한번씩 검증 세트로 사용하고, 나머지 9개 부분을 학습에 사용하였다.

표 1. 음성/오디오 분류기별 분류 정확도 비교
Table 1. Comparison of classification accuracy by Speech/Audio classifier

Classifier	Accuracy for speech (%)	Accuracy for audio (%)
G.718	71.2	65.3
CART model	80.5	78.2
Proposed	88.4	85.6

표 1은 음성/오디오 분류기별 성능평가 결과를 보여준다. 제안된 딥러닝 기반 모델은 음성 신호에 대해 88.4%, 오디오 신호에 대해 85.6%의 분류 정확도를 기록하여 G.718 분류기의 71.2%, 65.3%와 서브밴드 기반 CART (Classification and Regression Trees) 모델의 80.5%, 78.2%와 비교하여 상당한 성능 개선을 보여준다. 이는 제안된 방법이 음성/오디오의 시간 및 주파수적 특징을 효과적으로 학습하여 신호의 복잡한 패턴을 잘 반영한 결과로 해석된다. 즉, 제안된 딥러닝 모델이 주파수 대역에서 독립적으로 신호의 특징을 추출하면서도 시간적 변화를 포착하고, 복잡한 신호들 간의 상호작용을 학습할 수 있어 기존의 전통적인 분류기들에 비해 뛰어난 성능을 보인 것이다.

따라서, 제안된 딥러닝 모델 기반의 음성/오디오 분류기는 시간 및 주파수적 의존성이 중요한 신호 분류 작업에서 매우 효과적임을 입증하였다. 이는 향후 음성 및 오디오 처리 시스템에서 중요한 기술적 발전을 의미하며, 다양한 실시간 신호 처리 시스템에 적용 가능함을 보여준다.

Ⅴ. 결 론

본 논문에서는 음성과 오디오 신호를 효과적으로 분류하기 위해 딥러닝 기반 서브밴드 분석 방법을 제안하였다. 제안된 방법은 기존의 G.718 및 서브밴드 기반 이진 결정 트리 분류기가 가진 한계를 극복하고, CNN과 LSTM을 활용하여 신호의 공간적 및 시간적 특성을 정밀히 학습하도록 설계되었다. 실험 결과를 통해 제안된 모델이 음성과 오디오 신호 간의 복잡한 패턴과 상호작용을 효과적으로 반영하여 기존의 전통적인 방법보다 높은 분류 정확도를 보임을 확인하였다. 이는 CNN과 LSTM을 결합한 딥러닝 모델이 신호의 공간적 및 시간적 특성을 통합적으로 학습함으로써 음성과 오디오 신호의 복잡한 패턴을 효과적으로 구분할 수 있음을 보여준다. 결과적으로 본 논문에서는 음성과 오디오 신호를 분류하는 문제에서 딥러닝 기반 모델의 설계와 적용가능성을 제시하였으며, 이는 기존의 전통적인 신호 분류 방법들이 처리하지 못했던

신호의 복잡한 특성을 효과적으로 분석할 수 있는 방향성을 제시하는 것이다.

본 논문에서 제안된 방법은 단일 음성/오디오 신호 분류에 집중하고 있어 다중 음성 신호나 혼합 음원 환경에서의 분류 성능을 향상시킬 수 있는 방법에 대한 연구가 필요하다. 이를 위해 딥러닝 모델에 다중 클래스 분류 또는 다중 라벨 분류 기능을 추가하거나 각 신호를 구분하기 위한 분리 기법에 대한 연구를 수행할 계획이다.

감사의 글

본 논문은 2024년도 나사렛대학교 교비학술연구조성비 지원에 의해 연구되었음.

참고문헌

[1] ITU-T (International Telecommunication Union Telecommunication Standardization Sector), Frame Error Robust Narrowband and Wideband Embedded Variable Bit-Rate Coding of Speech and Audio from 8-32 kBit/s, Author, Geneva, Switzerland, G.718, June 2008.

[2] M. Jelinek, T. Vaillancourt, A. E. Ertan, J. Stachurski, A. Ramo, L. Laaksonen, ... and S. Bruhn, "ITU-T G.EV-VBR Baseline Codec," in *Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas: NV, pp. 4749-4752, March-April 2008. https://doi.org/10.1109/ICASSP.2008.4518718

[3] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, New York, NY: Chapman and Hall/CRC, 2017. https://doi.org/10.1201/9781315139470

[4] R. Timofeev, Classification and Regression Trees (CART) Theory and Applications, Master's Thesis, Humboldt University of Berlin, Berlin, Germany, December 2004.

[5] K. Kim, "Enhanced Signal Classifier for Speech/Audio Coding by Binary Decision Tree," *Journal of Knowledge Information Technology and Systems*, Vol. 18, No. 1, pp. 121-128, February 2023. https://doi.org/10.34163/jkits.2023.18.1.013

[6] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 206-219, May 2019. https://doi.org/10.1109/JSTSP.2019.2908700

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, Pittsburgh: PA, pp. 369-376, June 2006. https://doi.org/10.1145/1143844.1143891

[8] B. Yegnanarayana and S. Kumar, "Event-Based Instantaneous Fundamental Frequency Estimation from Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 4, pp. 614-624, 2009.

[9] J. Geiger, B. Schuller, and G. Rigoll, "Large-Scale Audio Feature Extraction and SVM for Acoustic Scene Classification," in *Proceedings of 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1-4, 2013. https://doi.org/10.1109/WASPAA.2013.6701857

[10] K. Zaman, M. Sah, C. Direkogl, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, Vol. 11, pp. 106620-106649, 2023.

[11] M. Gourisaria, R. Agrawal, M. Sahni, and P. Singh, "Comparative Analysis of Audio Classification with MFCC and STFT Features Using Machine Learning Techniques," *Discover Internet of Things*, Vol. 3, No. 1, pp. 1-10, 2024.

[12] P. Markopoulos, G. Karystinos, and D. Pados, "Optimal Algorithms for L_1-Subspace Signal Processing," *IEEE Transactions on Signal Processing*, Vol. 62, No. 19, pp. 5046-5058, 2014.

[13] Q. Li, B. Liao, L. Huang, C. Guo, G. Liao, and S. Zhu, "A Robust STAP Method for Airborne Radar with Array Steering Vector Mismatch," *Signal Processing*, Vol. 128, pp. 198-203, November 2016. https://doi.org/10.1016/j.sigpro.2016.04.006

[14] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, Vol. 55, No. 10, pp. 78-87, October 2012. https://doi.org/10.1145/2347736.2347755

[15] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Proceedings of the 13th European Conference on Computer Vision (ECCV 2014)*, Zurich, Switzerland, pp. 818-833, September 2014. https://doi.org/10.1007/978-3-319-10590-1_53

[16] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio Scene Classification with Deep Recurrent Neural Networks," arXiv:1703.04770, 2017. https://doi.org/10.48550/arXiv.1703.04770

[17] H. Sak, A. W. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large

Scale Acoustic Modeling," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, Singapore, pp. 338-342, September 2014. https://doi.org/10.21437/INTERSPEECH.2014-80

[18] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 2023, pp. 1-28, 2023.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980v1, December 2014. https://doi.org/10.48550/arXiv.1412.6980

[20] Y. Bengio and Y. Grandvalet, "No Unbiased Estimator of the Variance of K-Fold Cross-Validation," *Journal of Machine Learning Research*, Vol. 5, pp. 1089-1105, December 2004.

김광기(Kwangki Kim)

2004년 : 한국과학기술원 (공학석사)
2011년 : 한국과학기술원
(공학박사-정보통신공학)

2012년～2013년: 삼성전자 DMC연구소
2013년～현 재: 나사렛대학교 IT인공지능학부 부교수
※관심분야 : 신호처리, 3D 음향, 디지털콘텐츠 등