軍 운송수단 음향신호 분류를 위한 인공지능 모델 설계 및 비교 실험

최연희[1,3], 하은택[2,3], 안병현[1,3], 김민재[4], *백준기[4,5]
대한민국 육군[1], 해군[2]
중앙대학교 국방 AI대학 핵심인재과정[3], 영상학과[4], AI학과[5]
e-mail : {gregoyo, het117, anpos1}@naver.com, {imkbsz, paikj}@cau.ac.kr.

Design and Evaluation of AI Models for Acoustic Classification of Military Vehicles

Yeon hee Choi[1,3], Eun Taek Ha[2,3], Byung Hyum Ahn[1,3], Min-jae Kim[4], *Joonki Paik[4,5]
ROK Army[1] and Navy[2]
Defense AI Core Talent Program[3], Department of Image[4],
and Department of AI[5], Chung-Ang University

Abstract

In this paper, we present AI models and evaluate their performance to classify the operation and engine sound of 23 types of military vehicles. This paper examines six audio feature extraction methods and trains five machine and deep learning models on the audio signals recorded using an ordinary microphone. Experimental results show that the lightweight linear model can achieve more than 80% accuracy when sufficient information is provided in the feature extraction step.

I. 서론

민간 분야에서 촉발된 인공지능 기술의 획기적인 발전을 참고하여[1-3], 최근 軍은 지능화된 무기 및 전력지원체계의 개발을 위해 다양한 R&D 및 전투 실험을 추진하고 있다. 그동안 軍 인공지능 연구의 주요 관심사는 자율주행, 무인감시, intelligent surveillance reconnaissance (ISR) 기반 객체 탐지/식별 등 영상 이해와 자연어 처리(natural language processing; NLP) 모델에 비교적 편중되어 왔다. 이에 본 논문은 軍 장비 중 전차, 장갑차, 각종 전투차량 등 총 23종의 운송수단에서 발생하는 소리를 구분하는 기계학습 및 딥러닝 모델을 학습하여 실험한 결과를 제시한다. 전처리 방법과 모델 구조에 대한 폭넓은 실험을 통해 적절한 전처리 과정이 수반되는 경우, 휴대용 단말기에 장착이 가능한 경량화된 모델을 이용해서도 80% 이상의 정확도로 軍 운송수단을 분류할 수 있음을 확인할 수 있었다.

II. 제안하는 방법

2.1 데이터셋 구성

본 논문에서는 AI-Hub[4]에서 제공하는 자연 및 인공적 발생 非 언어적 소리 데이터의 軍 운송수단 데이터 중 2,369개의 음향 신호를 활용하여 연구를 수행하였다. 활용된 음향 신호의 소음원은 총 23종의 장비로, 이들은 표 1과 같이 세 가지 종류로 구분될 수 있다. 활용된 음향은 클립별 30초 분량의 단일 채널(Mono) 신호이며, 샘플링 속도는 44.100Hz, 비트율(Bit rate)는 128kbps, 정수 비트 깊이는 16bit이고, 파일 포맷은 MP3 (FFMPEG)이다. 본 연구에서는 전체 2,369개의 음향 데이터를 6:2:2로 나누어 1,421개는 학습용, 474개는 검증용, 그리고 474개는 평가용으로 설정하였다.

2.2 軍 운송수단 음향신호 특징 분석

모델 설계에 앞서, 원신호가 갖는 음파의 특성을 이해하기 위해 Spectrogram 분석을 수행하였다. 그림 1을 보면, 운송수단 분류별로 주파수-시간 특성에 유의미한 차이가 있음을 알 수 있다. 특히 그

차이는 고주파 영역에서 두드러졌는데, 이는 음향 신호의 전처리 과정이 모델의 정확도에 큰 영향을 끼칠 수 있음을 암시한다고 볼 수 있다.

구분	운송수단
전차(2)	K-1, K-1a1
궤도·장갑차 (10)	K-56, K-77, K288a1, K-200, K800, 화생방정찰차, K10탄약운반차, Km9ace, 교량 전차, 장애물개척전차
차륜 전투차량(11)	2.5t, 9.5t, 5t, 10t, 27t, 다목적 굴착기, 살수차, 대형버스, 부식수송차량, 승용차, 통신 가설 차량

표 1. 본 논문에서 활용된 음향 신호원종류

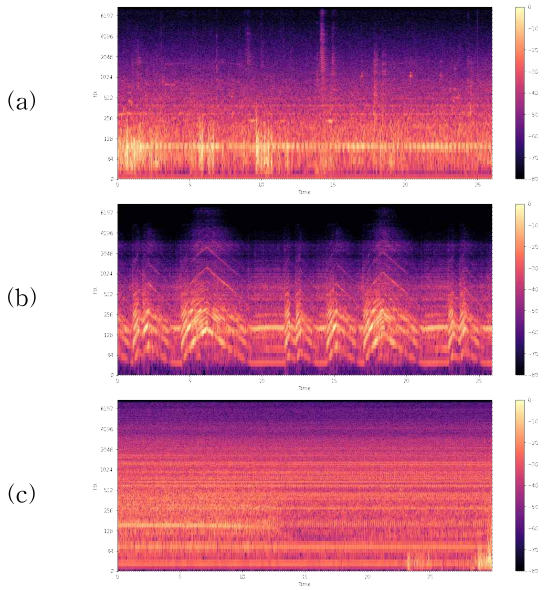

그림 1. Spectrogram을 이용한 軍 운송수단 종류별 음향 신호의 시간-주파수 시각화:
(a) 전차, (b) 궤도·장갑차, (c) 차륜 전투차량

2.3 음향 신호 전처리 및 특징 추출

본 논문에서는 네트워크의 입력 신호를 구성하기 위해, 여섯 종류의 음향 신호 전처리 기술들을 적용하였다: (1) Zero-crossing rate, (2) Spectral centroid, (3) Spectral roll-off, (4) MFCC(Mel-Frequency Cepstral Coefficient), (5) Chroma frequencies, (6) Mel spectrogram.

Zero crossing rate는 오디오 신호 파형의 위상이 중심축을 통과하는 횟수를 나타내며, 음성인식에서 유·무성음의 판별에 주로 사용된다. Spectral centroid는 소리를 주파수로 표현했을 때, 주파수의 가중평균을 계산하여 소리의 무게 중심이 어디에 있는지를 나타내는 지표이다[5]. Spectral roll-off는 스펙트럼의 형태와 낮은 주파수 영역에 신호의 에너지가 얼마나 집중되어 있는지를 나타내는 지표이다. MFCC는 인간의 청각 특성을 모델링 하는 방법으로 오디오 신호의 스펙트럼을 Log-scale로 변환한 후 주파수 특성을 그룹화하여 인간의 청각 특성에 맞는 Mel-Frequency 스케일로 변환한 것이다[6]. 본 논문에서는 MFCC의 초매개변수 n_{mfcc}를 40으로 설정하였다. Chroma frequencies는 음향 신호를 짧은 길이의 프레임 단위로 나누어 그 안에서 스펙트럼을 구한 후, 옥타브 차이가 나는 스펙트럼 성분들을 가산하여, 전체 스펙트럼 성분들을 하나의 옥타브 안으로 접어 표현하는 것이다. Mel spectrogram은 음성 데이터에 STFT(Short-Time Fourier Transform)을 적용하여 계산한 값에 Mel filter bank를 적용하고, 각 필터의 에너지 값을 합한 후 채널별 데이터를 2차원으로 나타낸 것이다.

2.4 軍 운송수단 음향신호 분류를 위한 기계학습 및 딥러닝 모델

본 논문에서는 두 종류의 기계학습 모델과 세 가지 합성곱 신경망(Convolutional Neural Network; CNN) 모델을 이용하여 음향 신호를 분석하였다: (1) LR(Multinomial Logistic Regression), (2) SVM(Multi-class linear Support Vector Machine), (3) LeNet, (4) VGG16, (5) ResNet50.

LR은 선형 모델로서 확장된 로지스틱 함수인 SoftMax 함수를 적용한 뒤 조건부 최대 엔트로피를 학습하는 모델이다[7]. SVM은 일반화된 힌지(Hinge) 손실 함수를 이용하는 선형 모델이다[8]. LeNet, VGG, ResNet은 합성곱 신경망의 이른바 토대 모델들(Foundation models)이다[1,2,9]. LeNet은 인간의 시각 특성을 모방하는 CNN 모델을 처음 실용화한 방법으로, LeNet5는 세 개의 콘볼루션 계층과 두 개의 완전연결 계층으로 구성된다[9]. VGG는 단순한 3x3 콘볼루션 계층을 반복적으로 쌓아 올린 깊은(Deep) 네트워크가 우수한 성능을 보인다는 것을 증명한 방법이고, 그중 VGG16은 13개의 콘볼루션 계층과 3개의 완전연결 계층을 쌓은 것이다[1]. ResNet은 각 유닛의 입력과 출력 사이의 잔차(Residual)를 계산하여 더해주면, 훨씬 깊은 모델의 학습이 가능해진다는 것을 보여준 방법이다[2]. ResNet50은 총 50개의 콘볼루션 및 완전연결 계층들로 구성된다.

여섯 종류의 전처리 및 특징 추출 결과는 수의 재배치를 통해 동일한 $H \times W$ 크기의 2D 배열로 변환하여 딥러닝 모델의 입력 값을 구성하였다. 선형 모델인 LR과 SVM은 HW 크기의 1D신호를 입력으로 설정하였다. 여기서 추출된 입력 벡터(텐서)의 크기는 $H = 40$, $W = 1292$로 설정하였다. 모든 모델의 입력은 학습 데이터의 최소/최댓값에 대한 통계적 특성을 활용하여 [0, 1]의 범위로 표준화하였다. 딥러닝 모델의 Stride는 입력 텐서의 크기에 따라 적절하게 조절하였고, 그밖에 언급되지 않은 사항은 원본의

설정을 따랐다.

Ⅲ. 실험 및 결과

3.1 세부 설정

전처리 방법과 모델 구조 외에도, 본 논문에서는 Optimizer와 학습 Epoch의 수를 초매개변수로 설정하였다. 학습에 고려한 Optimizer는 SGD와 ADAM[10] 이다. 그 결과, 표 2에 나타난 것처럼 총 120개 조합의 실험을 구성할 수 있었다.

알고리즘의 구현은 Keras[11]와 Tensorflow[12] 라이브러리를 이용하였다. 실험은 Nvidia Geforce RTX 3070 Laptop GPU의 구성을 갖는 노트북에서 GPU 가속화 연산을 활용해 수행하였다.

특징 추출: 6종류		모델: 5종류		학습 조건: 4종류	
Zero-crossing rate	×	LR	×	Optimizer	Adam
Spectral centroid		SVM			SGD
Spectral roll-off		LeNet5		Epoch	100
MFCC		VGG16			300
Chroma frequencies		ResNet50			
Mel spectrogram					

표 2. 초매개변수 설정 (총 120개의 조합)

3.2 실험 결과 및 분석

초매개변수 조합에 따른 정확도 측정 결과는 그림 2에 나타나 있다. 표 3은 모델과 전처리 방법에 따른 분류 정확도를 나타낸다. 모델을 막론하고, 여섯 가지 특징 추출 방법 중에서 MFCC를 이용한 경우가 가장 우수한 성능을 보여주었다. 또한 Adam optimizer가 SGD보다 좋은 학습 결과를 보여주었다. 표 4는 학습 Epoch 설정에 따른 모델별 정확도를 보여준다. 대부분 학습을 많이 할수록 성능이 좋았지만, 더러는 과적합(Overfitting) 문제가 발생하기도 하였다.

Model		Accuracy (%)	Model		Accuracy (%)
LR	ZCR	20.04	LeNet5	MFCC	83.33
	Centroid	14.56		Chroma	53.8
	Roll-Off	17.09		Mel	43.67
	MFCC	86.92	VGG16	ZCR	21.1
	Chroma	44.30		Centroid	14.56
	Mel	56.96		Roll-Off	24.05
SVM	ZCR	14.56		MFCC	67.72
	Centroid	16.03		Chroma	42.62
	Roll-Off	17.30		Mel	54.01
	MFCC	87.55	ResNet 50	ZCR	25.74
	Chroma	43.46		Centroid	21.73
	Mel	61.81		Roll-Off	25.53
LeNet5	ZCR	22.78		MFCC	76.79
	Centroid	21.52		Chroma	51.69
	Roll-Off	23.84		Mel	48.87

표 3. 모델과 전처리 방법에 따른 분류 정확도
(네 가지 학습 조건 중 가장 좋은 결과로 표기,
모델별 가장 좋은 결과는 파란색 글씨로 표시)

그림 2. 분류기의 초매개변수 설정에 따른 정확도, 정밀도, F-score 측정 결과

Model	LR		SVM		LeNet5		VGG16		ResNet50	
	100	300	100	300	100	300	100	300	100	300
Accuracy (%)	84.18	86.92	83.54	87.55	83.33	82.07	62.24	67.72	76.79	71.31

표 4. 학습 Epoch 설정에 따른 정확도
(여섯 가지 전처리 방법 중 가장 좋은 결과로 표기)

일반적인 통념과는 다르게, 본 실험에서는 선형 모델인 SVM과 LR이 딥러닝 모델보다 높은 정확도를 보여주었다. 오히려 전처리 방법의 설정이 성능에 큰 영향을 주었다. 이러한 현상이 어떻게 발생했는지 이해하기 위해, 가장 정확도가 높게 나타난 MFCC 특징을 t-SNE(t-distributed Stochastic Neighbor Embedding) 방법[13]으로 시각화하였다. 그림 3의 시각화 결과를 보면, n_{mfcc}를 40으로 설정하여 고주파수 성분을 충분히 확보한 경우 특징 벡터가 클래스별로 상당히 군집화해있는 것을 볼 수 있었다. 반면 $n_{mfcc} = 20$의 경우, 클래스별 군집화 정도가 덜하였다. 이에 별도의 실험을 수행해본 결과, n_{mfcc}를 40보다 작게 설정하면 모델의 정확도가 표 3에 언급된

것보다 떨어지는 것으로 나타났다.

모델별 학습과 평가 시간은 표 5와 같이 측정되었다. 모든 경우, 노트북 환경에서 30초 분량의 음향 신호를 분석하는데 0.1초 이내의 시간이 소요되었다. 가장 빠른 LR과 SVM의 경우, 초당 약 2,500개의 음성 신호를 분류할 수 있었다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 음향 신호를 해석하여 軍 운송수단을 자동으로 분류하는 기계학습 및 딥러닝 모델을 설계하고 학습하였다. 다양한 특징 추출 방법 중에서는 MFCC가 가장 우수한 결과를 보여주었으며, 의외로 선형 모델이 딥러닝 모델을 능가하는 성능을 보여주었다. 이는 그림 1과 그림 3에서 볼 수 있는 것처럼 고주파 성분이($n_{mfcc} \geq 40$) 충분히 보존된 상황에서 軍 운송수단의 음향 신호는 선형 분리가 가능하기 때문이라고 이해할 수 있다. 만약 분류할 클래스의 수가 더 늘어나거나, 주변 잡음의 세기가 증가하고, 학습용 데이터의 수가 더 많아지게 되면, 딥러닝 모델의 성능이 더 향상될 것이다.

본 논문에서는 적절한 음향 신호 전처리 과정을 적용한다면, 휴대용 단말기에서 구동이 가능할 정도로 가벼운 선형 모델로도 80%이상의 정확도로 軍 운송수단을 분류할 수 있음을 확인할 수 있었다. 향후 연구에서는 딥러닝 모델의 성능을 이끌어내기 위해 데이터 증강, 정칙화, 사전학습 등의 방법 등을 시도할 계획이다. 본 논문에서 제시하는 음향 기반 軍 운송수단 분석 결과가 추후 유사 연구에 도움이 되기를 바란다.

Model	LR	SVM	LeNet5	VGG16	ResNet50
Train	0.0237	0.0229	0.0637	3.8202	4.3865
Test	0.0004	0.0004	0.0007	0.0124	0.0065

표 5. 모델별 평균 학습 및 평가 시간 (sample/seconds)

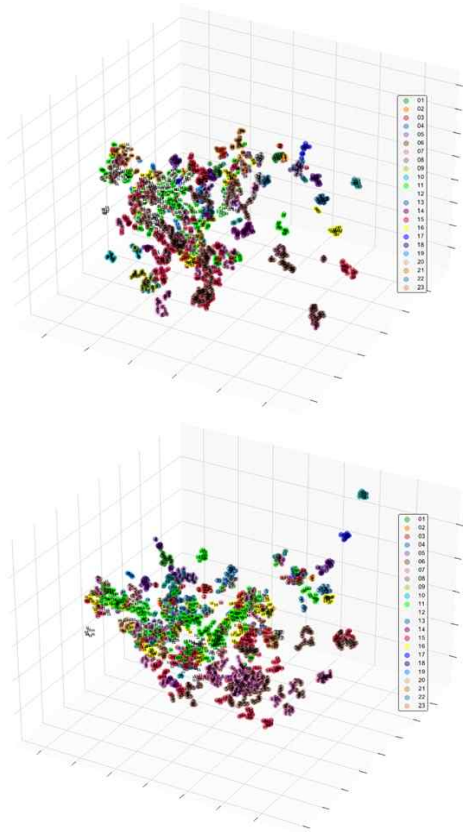

그림 3. 음향 신호 MFCC 특징 벡터의 t-SNE 시각화 (위: $n_{mfcc} = 20$, 아래: $n_{mfcc} = 40$)

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원[No.2014-0-00077, (2세부) 대규모 실시간 비디오 분석에 의한 전역적 다중 관심객체 추적 및 상황 예측 기술 개발]과 정보통신기획평가원의 지원(No.2022-0-00601, 군 특화 AI 교육과정 개설·운영)을 받아 수행된 연구임.

참고문헌

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, 2015.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, 2016.
[3] T. B. Brown et al., “Language Models are Few-Shot Learners,” NIPS, 2020.
[4] AIHub Homepage. Available online: http://www.aihub.or.kr/ (accessed on 31 August 2022).
[5] J. M. Grey and J. W. Gordon, “Perceptual effects of spectral modifications on musical timbres,” J. Acoust. Soc. Am., vol. 63, no. 5, pp. 1493-1500, 1978.
[6] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, “HMM-based audio keyword generation,” Pacific-Rim Conference on Multimedia, Springer, Berlin, Heidelberg, 2004.
[7] W. David and R. X. Stanley Lemeshow, Applied logistic regression, vol. 398. John Wiley & Sons, 2013.
[8] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” IEEE Trans. Neural Netw., vol. 13, no. 2, pp. 415-425, 2002.
[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv 2014.
[11] A. Gulli and S. Pal, Deep learning with keras. Birmingham, England: Packt Publishing, 2017.
[12] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” arXiv, 2016.
[13] L. van der Maaten and G.E. Hinton, “Visualizing High-Dimensional Data Using t-SNE,” J. Machine Learning Research, vol. 9, pp. 2579-2605, 2008.