2024년 2월

석사학위논문

딥러닝 전이학습 모델과 설명 가능 기법들을 이용한 음성 감정 인식

조 선 대 학 교 대 학 원

전자공학과

김 태 완

딥러닝 전이학습 모델과 설명 가능 기법들을 이용한 음성 감정 인식

Speech Emotion Recognition Using Deep Learning and Explainable Techniques

2024년 2월 23일

조 선 대 학 교 대 학 원

전자공학과

김 태 완

딥러닝 전이 학습 모델과 설명 가능 기법들을 이용한 음성 감정 인식

지도교수 곽 근 창

이 논문을 공학석사학위신청 논문으로 제출함

2023년 10월

조 선 대 학 교 대 학 원

전자공학과

김 태 완

김태완의 공학석사학위논문을 인준함

위원장 염 홍 기 (인)

위 원 신 주 현 (인)

위 원 곽 근 창 (인)

2023년 12월

조 선 대 학 교 대 학 원

목 차

표 목 차

도 목 차

ABSTRACT

Speech Emotion Recognition Using Deep Learning and Explainable Techniques

Kim, Tae-Wan
Advisor : Prof. Kwak, Keun Chang, Ph. D.
Dept. of Electronic Engineering,
Graduate School of Chosun University

Recent advances in technologies utilizing speech, such as user recognition, IoT applications, and emotion classification, have garnered increasing attention. Pre-processing, feature extraction methods, and appropriate models are needed to effectively use speech data in these technologies. In this paper, we propose a speech emotion analysis and recognition using deep learning and explainable method. To implement the above method, we proceed through 5 steps. First, in order to obtain more informaion about emotion, speech data is converted into a spectrogram image in the time-frequency domain by STFT(Short-Time Fourier Transform). Second, use a GDS(Gaussian Data Selection) mechanism, that using gaussian distribution and correlation coefficient to reduce a data volume. Third, higher performance and adapt diverse of speechs, we constructed a late fused to VGGish and YAMNet that learn and extract features independently through each model. Fourth, we apply

three explainable models(Grad CAM, LIME, and Occlusion Sensitivity) to visually confirm which area the trained model focused on while classifying data and which time-frequency domain characteristic make the decision. Finally, we adjust focused area which obtain in Grad CAM method on speech signal.

After convert speech signal to spectrogram image, that is divided for model's input size in time-domain. Unnecessary segments without emotional features such as silence, stutter and noise occur during traditional dividing, which make more spend computational resource for training. In the output of the model, User need to understand about model's decision where is the most crucial area in spectrogram time-frequency. Just output of model, users cant see that results can be sufficiently assist their judgment, and furthermore can't analyze that a certain area is closely related to the corresponding emotion. The data use in the experiment is acquired by Chosun University for emotional classification in 2021 and 2022(CSU2021, CSU2022) and public data in AI-hub for emotion classification.

In the paper, we select heterogeneous data above the threshold by applying correlation coefficients to the gaussian distribution of the segment spectrogram section. And by excluding that data from learning, we effectively reduce the size of the data to reduce computational resource consumption and learning time. Two transfer learning models are fused to design a model with high classification accuracy that can be applied to more diverse speech data. Also about classification result, we applying 3 explainable method(Grad CAM, LIME, Occlusion Sensitivity) to show the crucial area in spectrogram in a variety of ways. That area can explain which frequency band has closely features about emotions and which words and phrases influence the classification by analyzing in the time domain. The concentration area obtain by the Grad CAM is restored to the original signal size, so that the speech signal of the corresponding section could be additionally analyze. Since then, the above methods are expected to enable effective learning and classification for speech research, and to be used for multimodal research using concentrated words obtain through explanable techniques and situational analysis through emotional words.

제1장 서론

제1절 연구 배경 및 목적

음성을 통한 연구는 현대 사회에서 인간-컴퓨터 상호작용과 감정 분석, 음성 기반 서비스 등 사용자의 편리를 위해 다양한 분야에서 주목된다. 특히 음성은 사용자가 자신의 정보, 감정, 생각을 표현하는 대표적인 방법이기에 컴퓨터나 스마트폰, 스마트 워치와 스마트 홈에 사용되는 음성 인식 스피커 등에서 정확하게 신호를 수신받고, 수신된 정보를 분석하여 사용자의 감정, 생각 등을 프로그램에 전달하고 이해시키는 과정이 중요해졌다. 음성 데이터를 통해 얻은 정보는 제품과 서비스의 질을 높이는 데 사용되고, 분석하여 음성 기반 사용자의 정신 상태, 감정 분류, 감성 분석이나 IOT, 음성 인식 시스템 등 다양한 분야에서 활용된다. 또한 개인이나 그룹의 감정 상태를 이해함으로 의사 결정이나 건강 상태 모니터링에도 활용되고 있다[1]. 음성을 데이터로 활용하고 분석하기 위해서는 취득하는 장비의 성능에 따라 발생하는 데이터의 품질 차이와 일상적인 환경에서 발생 되는 주변 환경 잡음, 끊김, 취득 장비의 오류 등을 해결하기 위한 적절한 전처리 과정이 필요하고[2], 음성의 복잡성과 다양성을 효과적으로 분석할 수 있는 특징 추출과 분석 방법이 필요로 하다[3]. 현재의 연구에서는 취득된 음성에 잡음 제거와 전처리 과정을 거친 후, 음성이 가지고 있는 다양한 특징을 확인할 수 있는 형태로 데이터를 변형시키기 위해 주로 음성의 시간 영역 분석뿐만 아니라 음성에서 시간-주파수의 변화를 확인할 수 있는 로그-멜 스펙트로그램 이미지 형태로 변형하여 분류와 예측에 활용하고 있다. 스펙트로그램은 음성이 시간의 흐름에 따라 주파수 성분이 변화하는 것을 한 눈에 파악할 수 있고, 시각적으로 음성 신호를 직관적으로 이해하고 분석할 수 있으며, 기존의 신호로써만 분석하는 게 아니라 이미지의 관점으로 분석할 수 있어서 컴퓨터 비전의 다양한 방법을 적용할 수 있다[4].

설명 가능 모델은 주로 데이터를 딥러닝 모델이나 기계 학습 모델 같은 방식으로 학습할 때, 학습한 모델이 특정 클래스 분류 결정 과정을 확인함으로 개발자들에게 어떠한 부분에서 모델이 분류에 성공하였고, 실패하였을 때는 어떠한 요소가 실패의 원인이 되었는지 확인할 수 있어서 모델의 성능 분석에 활용되며, 모델을 활용한 프로그램을 사용하는 사용자와 기업에 인공지능에 대한 전문적인 지식이 없는 상태에서도 모델의

결정 과정을 확인할 수 있음으로, 사용하는 모델에 대한 신뢰성을 높이고 중요한 의사 결정을 진행할 때 사람의 판단을 보조할 수 있다. 스펙트로그램을 활용한 연구에서도 위와 같은 설명 가능한 기법을 적용하여 어떤 구간에서 모델이 분류와 예측에 관한 결과를 도출했는지, 스펙트로그램에서는 어떤 시간-주파수 영역이 해당 도출에 영향을 주었는지 확인할 수 있다[5].

본 논문은 취득한 음성 신호를 시간-주파수 영역으로 분석할 수 있는 스펙트로그램 이미지로 변형시키고, 모델에 입력 크기에 맞춰 분할하는 기존 음성 전처리 중 발생되는 불필요한 데이터(잡음, 침묵, 끊김 등)를 선별하기 위해 가우시안 분포와 상관 계수를 활용한다. 네트워크는 VGGish와 YAMNet을 Late-Fusion하여 데이터를 독립적으로 학습하여 감정 특징들을 학습하고, 마지막 계층에서 위 정보들을 종합함으로 다양한 음성에서도 감정과 관련된 특징을 추출하고 분류할 수 있는 모델을 설계한다. 또한, 분류 과정에서 모델의 해당 클래스에 대한 특징 초점 영역을 시각화하여 어느 시간-주파수 영역이 클래스 분류에 대한 핵심적인 영역을 확인하도록 설명 가능 기법(Grad CAM, LIME, Occlusion Sensitivity)을 적용 하고, 적용된 영역을 분석하여 어떤 단어와 음정이 감정에 핵심적인 역할인지, 감정과 관련된 주파수 영역을 확인한다.

제2절 연구 내용 및 구성

본 논문에서는 취득한 음성에서 다양한 특징을 확인할 수 있도록 시간-주파수 영역의 스펙트로그램 이미지로 변형시키고, 데이터를 선별하는 가우시안 데이터 선별(Gaussian Data Selection, GDS) 전처리와 YAMNet과 VGGish를 독립적으로 학습하여 분류 이전에 나온 정보를 결합하여 결과를 출력하는 네트워크(Custom YAMNet-VGGish Late Fusion Network, 융합모델)에 설명 가능 모델을 사용하여 분류한 결과에서 집중 영역을 시각화한다. 또한 표현된 모델의 클래스에 대한 스펙트로그램의 집중 영역을 시간 영역의 음성에 적용하여, 집중 영역을 청각적으로 확인할 수 있도록 하였다.

음성을 시간 영역만으로 분석하면 주파수 영역에서 얻을 수 있는 발화자의 감정이나 특성에 대한 정보가 부족해지고, 잡음의 영향에 더 민감해진다. 그래서 주파수 성분의 변화를 통해 음성의 감정 특징을 잘 파악할 수 있도록 시간-주파수 영역을 동시에 확인할 수 있는 형태로 데이터를 변형시킨다. 음성 데이터는 구간마다 일정 부분을 겹치도록 하여 STFT(Short Time Fourier Transform)를 통해 스펙트로그램 이미지로 변환한다. 변환된 데이터를 모델의 입력 크기에 맞춰서 구간별로 나누게 될 때, 학습에 영향을 주는 데이터와 영향을 주지 않는 구간이 생성되기 때문에, 영향을 주지 않는 데이터를 선별하여 효과적인 컴퓨팅 자원관리와 정확도 향상을 위해 GDS를 적용한다. 위 선별 과정은 분할된 스펙트로그램 이미지를 각각 평균과 분산을 사용하여 가우시안 분포를 구하고, 각 가우시안 분포 간의 상관 계수를 통해 데이터를 선별한다. 위 과정은 데이터의 값이 단순히 작거나 큰 데이터를 선별하는 것이 아닌 해당 데이터 내부에서 이질적인 데이터를 찾는 것을 목표로 하고 있다. 상관 계수를 통해 확인된 데이터 구간 중, 이질적인 구간의 경우 다른 구간보다 평균값과 분산이 급격히 변하는 구간으로, 위 구간을 제외하고 학습에 영향을 주는 데이터만으로 선별하여 효과적인 데이터 규모 축소를 진행할 수 있다. 선별된 데이터는 두 모델을 변형, 합성하여 생성한 융합모델을 통해서 분류하고, 모델을 통해 분류한 클래스에 대해 어떤 시간, 주파수 영역을 집중하여 나온 분류 결과인지, 영역의 중요도를 시각화하기 위해 Grad CAM, LIME, Occlusion Sensitivity 방법을 사용한다. Grad CAM으로 출력된 집중 영역은 청각적으로 확인하기 위해 오디오에 STFT 변환 과정에 사용한 파라미터를 고려하여 시간 영역으로 복원하여 신호에 적용하여 진행하였다. 본 논문에서는 제안하는 데이터 선별을 통해 학습에 유효한 데이터를 선별하여 규모를 줄였고, 이를 통해 학습 시간과 소모되는 컴퓨팅 자원을 줄였다. 또한

Late-fusion 모델을 통해 기존보다 높은 분류 정확도를 확인하고, 설명 가능한 기법을 통해 어떤 영역이 분류에 영향을 주었는지 스펙트로그램 이미지에서 시각적으로, 오디오 데이터에서 청각적으로 확인하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 음성을 스펙트로그램 이미지로 변형한 데이터를 활용한 딥러닝 분류와 설명 가능한 기술을 적용한 딥러닝 분류에 관련 연구에 관해 기술한다. 3장에서는 논문에서 음성 데이터를 스펙트로그램 이미지로 변환하는 방법에 대한 설명과 이미지에서 학습에 효과적인 영향을 주는 데이터를 선별하는 과정의 흐름과 방법론을 설명하고, 제안하는 음성을 효과적인 감정 분류를 하기 위해 변형한 합성된 네트워크에 대한 설명과 분류 과정을 분석하기 위한 다양한 설명 가능 기법에 대한 방법을 기술한다. 4장에서는 조선대학교에서 21년도, 22년도에 취득한 음성 데이터베이스와 AI-HUB의 공용 감정 음성 데이터 셋에 대한 설명과 규모를 기술한다. 또한 학습에 사용된 파라미터와 하이퍼 파라미터에 대한 설명과 융합모델의 분류 정확도, 설명 가능 모델로 집중 영역을 확인, Grad CAM을 오디오 데이터에 적용을 위한 방법을 설명한다. 마지막 5장에서는 결론과 향후 연구 방향에 관해 기술하고 마무리한다.

제2장 관련 연구

제1절 스펙트로그램 이미지에 딥러닝을 활용한 연구

Roy의 경우 폐 음성 데이터인 ICBHI 2017 challenge database, Chest Wall Lung Sound Database, RespiratoryDatabase@TR를 활용하여 적은 파라미터로 높은 분류 성능을 보여주기 위한 네트워크를 실험했다. 음성은 로그-멜 스펙트로그램 이미지 형태로 변환하고, 연구의 성능을 검증하기 위해 다른 모델들과 파라미터 수, 모델의 크기, 각 실험에 대한 정확도를 비교하여 목적한 모델이 다른 모델에 비해서 더 적은 파라미터 수를 갖고, 더 높은 분류 정확도를 가질 수 있음을 확인했다[6]. Xi의 경우, IEMOCAP 데이터셋을 활용하여 기존 음성 연구에서 발생하는 음성 길이 불균형으로 발생하는 불필요한 데이터 자원 소모를 부족한 영역에 특정 값으로 패딩하는 방법으로 균형하게 만들어 학습과 분류의 혼란을 줄였다. 각 패딩 된 데이터를 CNN과 LSTM을 활용한 네트워크를 활용하여 학습, 분류를 진행하였고 모델 성능을 평가했다. 데이터의 수가 불균형한 경우 클래스의 샘플 수에 따라 가중치를 부여하는 WA(Weighted Accuracy)와 샘플 수와 관계없이 동등하게 반영하는 UA (Unweighted Accuracy)를 사용하여 평가하였다. 기존의 데이터의 길이가 불균형한 경우의 WA와 UA보다 패딩 할 경우가 감정 분류에서 약간의 정확도 향상을 했다[7]. BADSHAH의 경우 Berlin Dataset 에 있는 7가지 감정 데이터를 효과적인 SER(Speech Emotion Recognition)하기 위해 제안된 3개의 CNN 계층과 FC(Fully Connected) 계층을 가진 네트워크에 학습 후 분류한 결과와 사전 학습된 미세 조정한 AlexNet의 분류 정확도를 비교하였다. AlexNet의 분류 정확도는 특정 클래스에서는 높지만 다른 클래스의 정확도는 낮은 과적합이 발생하였고, 제안한 모델의 경우 AlexNet의 정확도보다 균등하고 효과적임을 확인했다[8]. ZHANG은 MUOC 데이터에 있는 교실 분위기 정보를 가지고 있는 음성 데이터를 활용하여 수업 진행과 교실 분위기가 어떻게 진행되고 있는지 확인하는 연구를 진행하였다. 많은 사람이 모여있는 교실에 적용함으로 다수의 분위기와 환경적 특징을 확인하고 관찰할 수 있도록 하였다. FBank, MFCC, Spectrogram 특징들을 가져와 종합하여 더 정확한 분류와 적은 손실이 일어날 수 있도록 하였고, 각각의 특징은 CNN과 LSTM을 적용한 HNN(Hybrid Neural Network) 네트워크에 병렬로 독립적으로 학습을 진행하였다. FC 계층에서는 각 특징 맵을 독립적으로 진행하여 얻어진

정보를 종합하여 6가지 교실 분위기로 분류하였다[9]. Raja의 경우 Berlin Dataset(EMO-DB) 의 음성 데이터를 활용하여 감정을 인식할 때, 효과적인 데이터의 형태를 확인하기 위해 진행하였다. 데이터는 1차원 음성 형태와 MFCC를 적용한 2차원 이미지 특징 형태의 데이터를 SVM과 1D-CNN과 2D-CNN을 사용하여 학습했다. 성별마다 7가지 감정, 총 14가지의 클래스로 분류를 진행하였을 때, 1차원 음성 데이터를 활용한 SVM이 전체적인 클래스에 대해 가장 낮은 정확도를 보이고, MFCC를 활용하여 2차원 데이터로 학습한 2D-CNN으로 분류를 진행했을 때 가장 높은 정확도를 확인하면서 음성을 통해 분류할 때, 2차원 형태로 변형하는 것이 효과적임을 확인 했다[10]. Zheng의 경우 CASIA 중국 감정 데이터를 활용하여 사용자의 감정을 인식하고, 그에 따른 행동을 취할 수 있는 NAO 로봇에 기능을 학습하기 위해서 음성 스펙트로그램 이미지를 CNN과 RF(Random Forest)를 사용했다. 일상적인 환경 속에서 로봇이 수집할 수 있는 음성은 학습에 사용된 데이터보다 잡음이 섞여 있고 순도가 낮은 데이터가 취득된다. 기존 학습 방법보다 일반화 성능을 높이기 위해서 음성이 가지고 있는 특징을 CNN을 통해 추출하여 특징 맵을 가져와서, RF 분류기를 통해 6가지 감정을 분류하였다[11]. Shalini의 경우 지속적인 감정 모니터링을 통해 해당 사용자의 스트레스와 분노의 원인을 분석하고 그에 따른 해결책을 주기 위해서는 사람의 감정과 정신 상태를 가장 잘 표현하는 수단인 음성에 관한 연구가 필요하다. 위와 같은 연구를 진행하기 위해 공용 데이터인 Toronto Emotional Speech dataset(TESS)와 Berlin Emotional Database(EmoDB), Ryerson speech-Visual Database of Emotional Speech and Song(RAVDESS)를 활용하였으며, 음성을 스펙트로그램으로 표현하는 것 외에 RMS, ZCR, Spectral centroid, Spectral Entropy, Spectral roll-off, mean pitch, max pitch, min pitch, tempo, low energy, spectral irregularity 등의 방법으로 음성에서 다양한 특징을 추출하였다. 추출된 음성 특징들과 스펙트로그램은 독립적으로 네트워크에서 학습하여 특징을 추출하고, 중간 계층에서 얻어진 정보를 공유하여 더 다양한 특징과 정확도 증진을 위하여 연구를 진행하였다[12]. Heng의 경우, IEMOCAP 데이터가 가지고 있는 시간-주파수 특징을 스펙트로그램을 일정 크기로 분할하여 CNN과 LSTM으로 학습하였다. CNN으로 감정과 관련된 특징을 추출하고 LSTM을 통해 시간적인 특징을 학습하고 추출하였다. 위 방법으로 추출된 특징으로 WA, UA 방법으로 6가지 감정에 대한 분류 정확도 평가를 진행하였다[13]. Mohammed의 경우 전 세계적으로 큰 영향을 끼친 COVID-19를 기침, 호흡 소리를 통해 분류하기 위해 다국적 남녀 건강한 피험자와 COVID-19에 감염된 피험자의 기침, 대화, 호흡소리로 구성되어 있는 Coswara

데이터 셋을 활용하여 연구를 진행하였다. COVID-19 질병의 경우 잠복기나 무증상로 인해 확진자가 자신의 감염 여부를 확인하지 못할 수 있어서, 인지하지 못한 경우 자신의 의도와 상관없이 다른 사람에게 질병을 확산시킬 수 있다. 질병 확산을 방지하기 위해 사용자의 음성으로 비접촉 질병 감염 여부 파악 연구를 진행하였다. 건강인과 보균자의 기침과 숨소리를 Mel-scale 스펙트로그램으로 변환하고 ResNet18, 34, 50, 100, 101과 GoogleNet, MobileNetv2 등의 여러 전이 학습 모델로 분류 정확도를 비교하였다[14].

제2절 스펙트로그램에 딥러닝과 설명 가능 기법을 적용한 연구

Yuanyuan의 경우 IEMOCAP 데이터 셋을 활용하여 음성데이터를 스펙트로그램 이미지로 변형시키고, Alexnet을 통해 벡터 형식으로 특징 맵을 추출하였다. 이후 어텐션 계층에서 tanh와 softmax를 활용하여 핵심적인 특징 채널을 강조하는 어텐션을 진행하여 다른 CNN과 LSTM을 적용한 모델들과 비교하였을 때, 정확도가 약간 향상하는 효과적인 분류를 진행하였다. 또한 모델이 어느 영역을 집중하였는지 Grad CAM을 통해서 시각적으로 확인했다[15]. Sobahi의 경우 COUGHVID 데이터 셋과 VIRUFY 데이터 셋에 있는 음성을 ViT(Vision Transformer)를 사용하여 효과적인 COVID-19 질병 검출을 진행하였다. 음성 신호를 스펙트로그램으로 변형 후, YAMNet에 학습시켜서 발화자의 음성에서 기침 부분만을 수집하였다. 수집된 기침 소리를 이미지화 시켜 패치로 나누어 ViT를 진행하였고, 이미지 내의 지역 간의 상관관계를 고려하고 Transformer의 이점 들을 통해 기존보다 더 높은 정확도를 보여준다. 또한 Grad CAM을 활용하여 기침 소리 이미지의 어떤 부분이 COVID-19 질병과 관련된 핵심적인 영역인지 시각적으로 확인할 수 있다[16]. Carofilis는 VCTK 데이터 셋을 활용하여 화자의 목소리를 통해 지역 출신지를 인식하는 것을 목표로 했다. 음성 데이터를 스펙트로그램 이미지로 변형시켜, 다양한 CNN 모델로 학습하였다. 이후 핵심 영역을 Grad CAM을 생성하여 표현하며, 다시 스펙트로그램과 함께 특징을 결합시켜 새로운 특징 벡터를 생성하였다. 이후 특징 벡터를 여러 MLA(Machine Learning Algorithms) 방법을 활용하여 음성 발화자의 출신 지역을 분류하였다. UAR(Unweighted Average Recall)과 MAA(Macro Average Accuracy)를 활용하여 음성 파형을 입력한 방법과, 다른 MLA과 비교하여 성능을 분석하였다[17]. Bicer의 경우 DCASE 2016 Challenge에서 활용한 데이터를 사용하여 음향 신호를 분석하여 해당 환경 또는 장면을 식별하여 주변 상황을 분석하고 관리할 수 있는 음향 장면 분류(Acoustic Scene Classification)을 위해 연구하였다. 음성을 스펙트로그램 이미지로 변환하여 ResNet에 학습하여 3가지 환경으로 분류하고, 모델이 분류를 진행하면서 결정하는데 영향을 준 영역을 Grad CAM으로 확인하면서 스펙트로그램에 응용하여 분류한 클래스에 맞는 시간-주파수 영역을 강조하고 음성으로 다시 변환하여 확인할 수 있도록 했다[18]. Cesarelli의 경우 심장 박동 소리(Phonocardiogram)신호를 다양한 하이퍼 파리미터 값으로 2D-CNN 모델로 정상 박동 소리와 비정상 박동 소리를 학습하고, 이진 분류하였다. 이후 분류된 결과를 바탕으로 Grad CAM으로 심장 박동 소리의 시간 영역 중, 어느 부분을 통해 모델이 정상과 비정상을

분류하였는지 히트맵으로 확인하였다[19]. Pengqi Li의 경우 화자 인식에서 모델의 결정 과정을 시간-주파수 영역에서 나타내기 어려움이 있다. 화자 인식에 효과적인 설명 가능 기법을 확인하기 위해 Grad CAM, Score Cam, Layer Cam을 ResNet34SE모델에 적용하여 확인했다. 해당 연구에서는 Layer Cam이 핵심 영역을 세밀하게 표현함을 확인했다[20]. Henna의 경우 코로나19 감염 여부에 따른 기침 소리 차이를 CNN을 활용하여 분류하였다. 또한 추출된 특징에서 분류에 영향을 주었던 영역을 확인하기 위한 여러 설명 가능 기법(SmoothGrad, Grad CAM, LIME)을 활용하여 활성화했다[21]. LEE의 경우 갑상선 암 수술을 받은 환자의 수술 전과 후의 음성을 스펙트로그램으로 변환하여 EfficientNet과 LSTM을 통해 학습하여 회복 기간을 예측하는 연구를 진행했다. 또한 학습된 모델을 Grad CAM을 활용하여 입력되는 이미지의 시간-주파수 영역에서 분류에 영향을 주는 구역을 확인했다[22].

제3장 딥러닝 전이학습 모델과 설명 가능 기법들을 이용한 음성 감정 인식

3장에서는 본 논문에서 제시하는 기존 음성 전처리 과정 중 발생하는 불필요한 데이터 정리와 다양한 음성 대사에 적용 가능하며 높은 분류 정확도를 가지는 네트워크를 설계하고, 설명 가능 기법들을 이용해서 다양한 측면으로 모델의 결과를 분석하는 데 사용된 기술과 방법론에 관해 설명한다. 그림 1은 논문에서 제시하는 음성 감정에 설명 가능 기법들을 적용한 다양한 분석에 대한 구상도를 보여주고 있다. 음성 데이터를 CNN 기반으로 하는 네트워크에 입력하기 위해서 STFT를 사용하였고, 네트워크의 입력 크기에 따라 시간 영역을 기준으로 데이터를 분할시킨다. 이후 불필요한 데이터를 정리하기 위해 각 분할된 구간에 가우시안 분포와 상관 계수를 이용한 데이터 선별을 진행하여 데이터의 규모와 학습 시간 및 메모리 소모를 줄인다. 정리된 데이터에서 여러 특징을 독립적으로 추출하고, 종합하여 학습 및 분류를 진행할 수 있는 네트워크를 설계한다. 모델이 분류한 결과에 대해서 3가지 설명 가능 기법을 적용하여 다양한 측면으로 모델의 의사 결정 과정을 분석하고, 감정에 따른 시간과 주파수 활성화 영역을 확인한다. 이후 Grad CAM으로 얻어진 집중 영역을 음성 데이터에 적용하여 모델이 집중한 언어적, 음성적 특징을 직접적으로 확인하는 것을 목표로 한다.

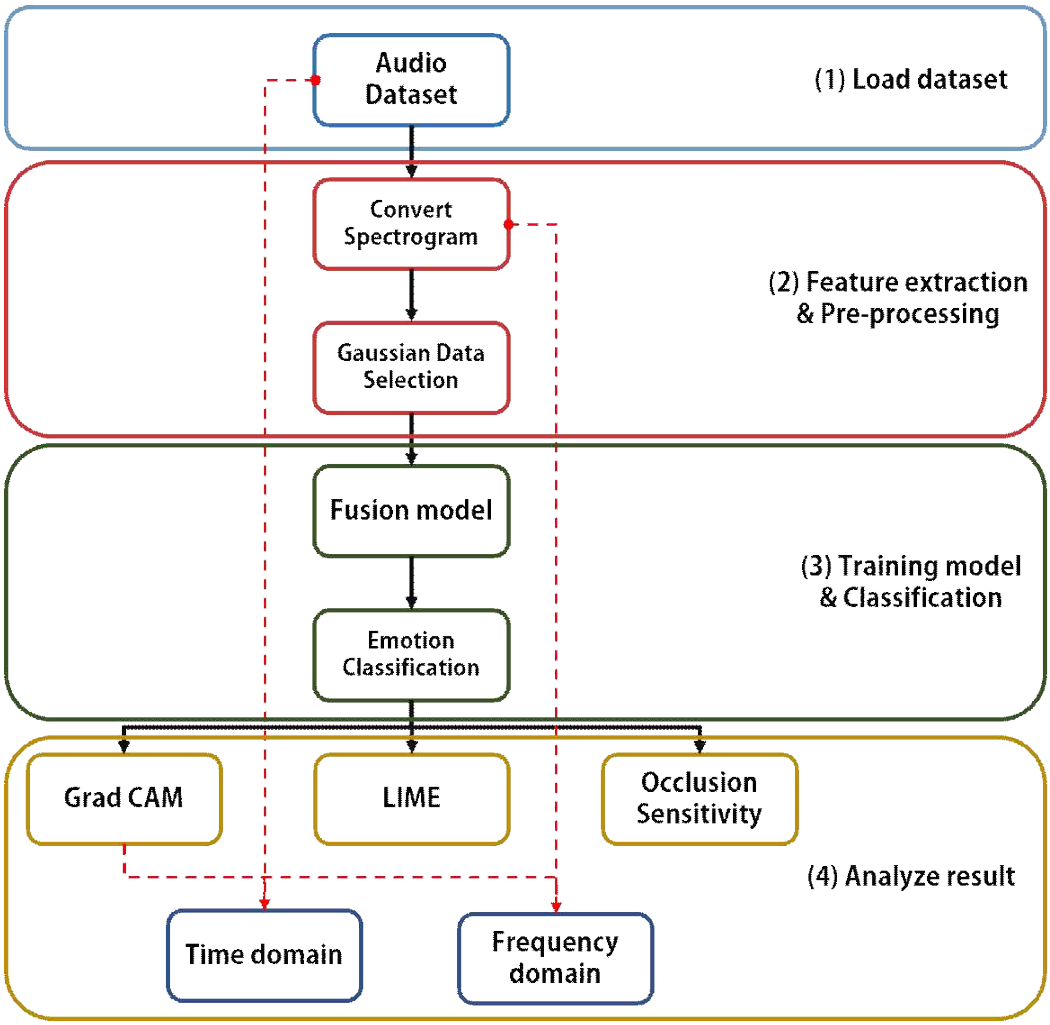


그림 1 음성 감정 인식 연구 구성도

제1절 데이터 전처리

1. 음성 데이터 스펙트로그램 이미지 변환

음성은 감정과 사용자의 정보, 건강과 밀접한 관계를 한 정보임으로, 음성에서 얻을 수 있는 정보를 분석하고 처리하여 스마트 워치, 스마트 카, 음성 인식, 음성 합성, 화자 인식, 의료 진단, 보안 등에 다양하게 활용되고 있다. 하지만 음성은 시간뿐 아니라, 주파수 영역에서도 다양한 정보를 가지고 있어서, 음성 데이터를 그대로 활용하는 것은 음성이 가지고 있는 모든 정보를 활용하기에 형태가 부적합하다. 이러한 정보를 직접적으로 시각화하면서 시간과 주파수 영역에서의 변화를 확인하기 위해 스펙트로그램으로 변환하여 연구 및 분석에 사용하고 있다. 스펙트로그램으로 음성을 표현하게 되면서 신호 처리 모델뿐 아니라 이미지 처리 모델에도 적용할 수 있게 되었고, 그에 따른 다양한 특징 추출 방법을 적용할 수 있게 되었다.

본 연구에서는 음성을 이미지로 변환할 때 자주 사용되는 방법인 STFT(Short-Time Fourier Transform)를 표 1과 같은 파라미터를 통해 이미지로 변환했다. 그림 2와 같이 STFT는 음성 신호를 시간-주파수 영역으로 변환하는 기술 중 대표적인 방법으로, 시간을 윈도우 크기로 나누어서 각 구간마다 FFT(Fast Fourier Transform)를 수행하여 주파수 정보를 얻는다. 위 방법을 통해 시간-주파수 정보를 이미지로 시각화할 수 있으나, 윈도우 크기에 따라서 시간, 주파수 해상도가 달라지며 반비례하기 때문에 적절한 윈도우와 중첩 크기를 선정해야 한다.

	윈도우 사이즈	중첩 사이즈	윈도우	주파수 범위
STFT 파라미터	1,200	720	hann 'periodic'	단방향

표 1 로그-멜 스펙트로그램 변환 STFT 파라미터

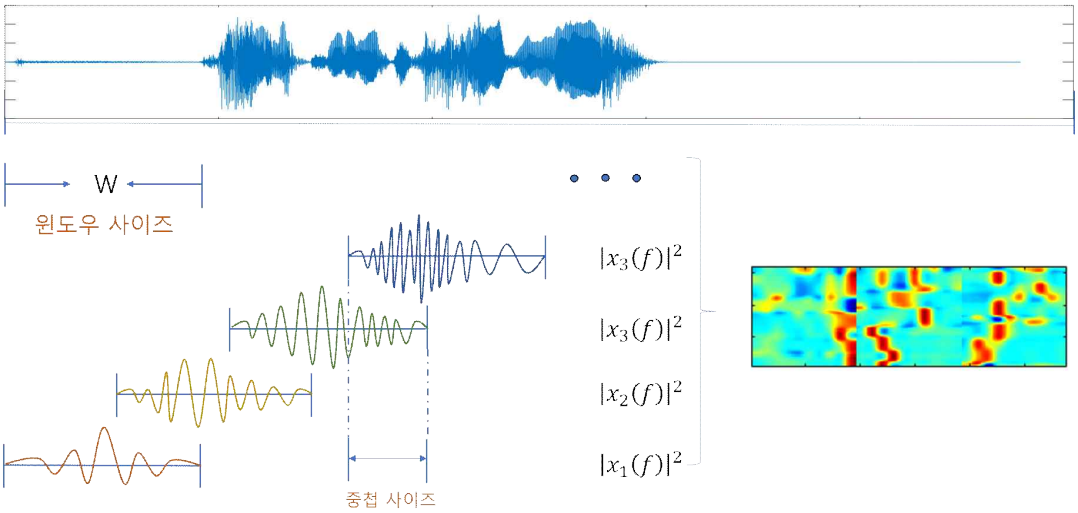


그림 2 음성 데이터 스펙트로그램 변환 과정

음성은 STFT를 통해 스펙트로그램 이미지로 변환한 다음, 표 2의 오디오 필터를 적용하여 원하는 주파수 구간을 가져오고, 모델의 입력 크기에 맞게 전처리하며 데이터의 가짓수를 늘리기 위해 일정 크기만큼 중첩하면서 분할된 데이터를 준비한다.

	밴드 수	샘플 레이트	주파수 스케일	FFT 길이
오디오 필터	64 채널	48,000	Mel	1200

표 2 오디오 필터 파라미터

2. 가우시안 데이터 선별 메커니즘

음성 데이터를 취득할 때, 같은 대사를 사용하여 취득하더라도 화자의 발음 속도에 따라서 데이터 길이가 달라지고, 끊김이나 잡음 등의 이유로 데이터 간의 길이를 일치시키기 어렵다. 또한 입력하려는 모델의 입력 크기에 따라서 조절이 필요하게 되어 음성 데이터는 작은 구간으로 나누어서 데이터의 길이를 통일시키고 부족한 부분은 패딩을 하거나 남는 부분을 자르는 방식을 적용한다. 하지만 구간을 나누게 되면서 감정 분류와 관련 없는 데이터의 구간이 생겨나고, 이러한 구간으로 데이터의 규모가 늘어나서 소모되는 컴퓨팅 자원이 증가하게 된다. 이러한 불필요한 구간을 선별하고 제외하기 위한 다양한 데이터 전처리 방식이 연구되고 있다.

그림 3에서는 데이터 취득 시 발생한 불필요한 잡음이나 침묵이 포함되어 있는 구간을 선별하기 위해 가우시안 분포와 상관 계수 적용되는 과정을 보여주고 있다. 선별 과정은 단순히 데이터의 크기가 낮거나 높은 구간을 제거하는 것이 아닌, 분할된 데이터의 구간의 상관 계수를 통해 이질적인 데이터를 선별하여 제외하는 것을 목표로 한다. 분할된 구간의 평균과 분산을 계산하여 가우시안 분포로 나타내고, 각 분포간의 상관 계수를 계산하고 정규화하여 설정한 임계값을 기준으로 선별한다. 위 방법으로 감정 특징이 포함된 데이터와 상관이 적은 구간, 즉 감정 특징 구간의 가우시안 분포와는 이질적인 구간을 제외하여 학습에서 제외함으로 데이터의 규모와 소모되는 컴퓨팅 자원을 감소시키면서 분류 정확도는 유지한다.

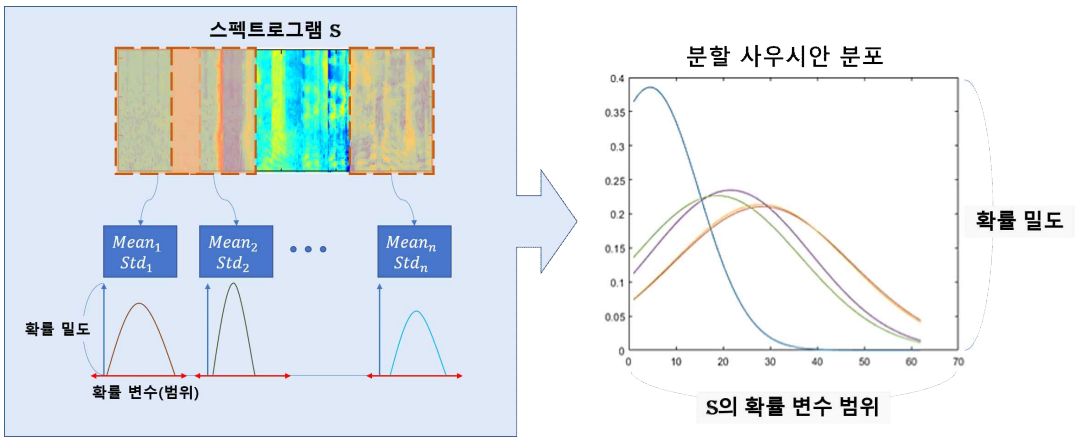


그림 3 음성 스펙트럼 가우시안 데이터 선별 적용 과정

제2절 VGGish와 YAMNet을 활용한 Late-Fusion 모델 설계

1. VGGish와 YAMNet 모델의 특징

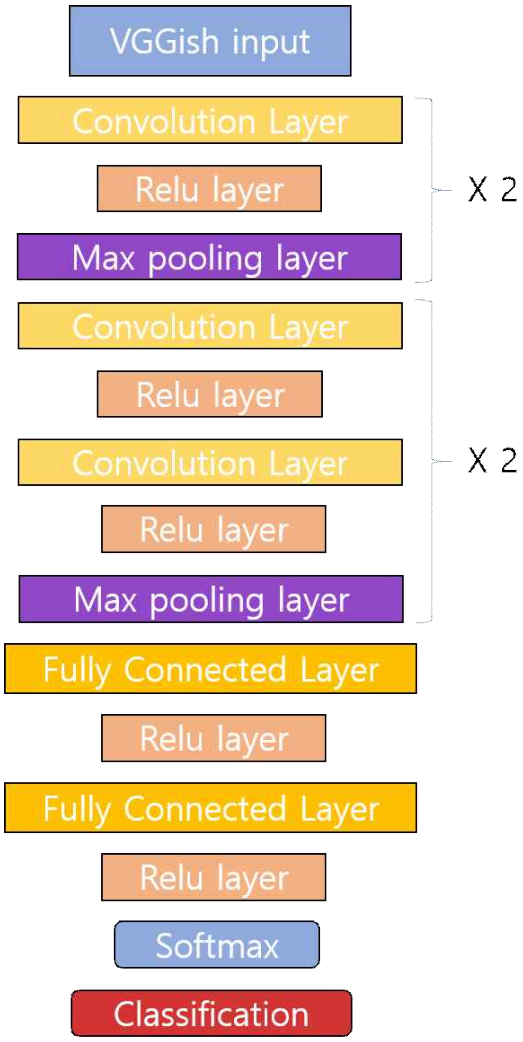


그림 4 VGGish 네트워크 구조도

그림 4는 Google에서 개발한 음성 인식 및 분류에 활용되는 VGGish(Video Game Genre Identification with Spectrograms and Headphones)모델의 구조를 보여준다. 기존 이미지 분류를 위한 VGG 네트워크를 음성 데이터 처리를 위해 변형하여 합성곱 신경망(Convolutional Neural Network)으로 설계되어있다. 시간 영역을 가진 1차원 데이터가 아닌 2차원 데이터에서 특징을 추출하기 위한 목적으로 설계된 위 모델은 입력되는 이미지의 공간적인 패턴과 특징을 가중치를 통해 학습하는 합성곱(Convolutional) 계층과 정규화를 위한 BN(Batch Normalize) 계층, 비선형적인 특징을 적용하여 더 복잡한 패턴을 학습할 수 있도록 하는 활성화 함수(Activation Function) 계층을 활용하여 오디오 이미지의 공간적 특징 벡터를 출력하여 분석과 분류에 활용한다.

입력되는 스펙트로그램의 시간-주파수 영역의 특징 정보를 보존하기 위해 전역 풀링 계층(Global Pooling Layer)을 적용하지 않고, 입력되는 데이터에 비해 큰 특징 맵을 생성하기 위해 완전 연결 계층의 채널 수를 크게 설계하였다. 각 계층 마다 얻을 수 있는 시간-주파수 영역의 위치 특징을 활용하기 위해 출력되는 정보를 종합하여 분류에 활용한다.

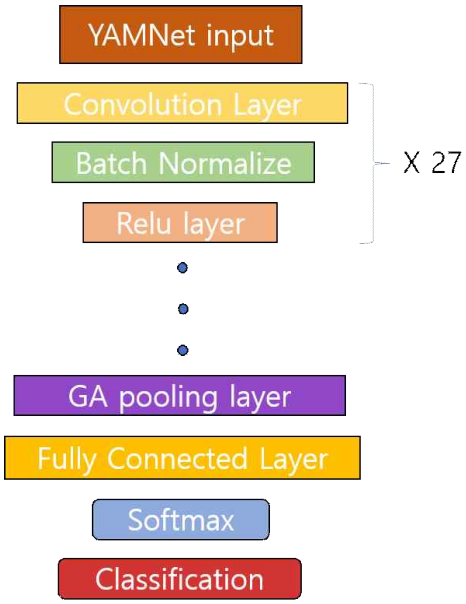


그림 5 YAMNet 네트워크 구조도

그림 5의 YAMNet 또한 VGGish와 마찬가지로 Google에서 개발한 오디오 분류 모델로, 동물의 소리, 자동차, 음악, 자연, 사람의 대화 등 512개의 다양한 음성 인식을 수행할 수 있도록 대규모 데이터 셋을 통해 학습한 전이 학습 모델이다.

위 모델은 합성곱 신경망을 기반으로 하여 여러 이미지의 패턴과 특징을 가져오는 합성곱 계층과 공간 차원을 줄여서 계산 효율성을 높이는 풀링 계층(Pooling Layer), 활성화 함수 계층으로 구성되어 있고 오디오와 관련된 분류 등에 활용되고 있다. 분류 계층 이전에 전역 풀링 계층을 사용하여, 이전 계층에서 입력되는 특징 맵의 시간-주파수 위치적인 특징보다는 추출된 특징 맵의 채널적인 특징에 주목하였다. 이는 입력되는 스펙트로그램 이미지의 전반적인 특징 파악과 위치적인 특징에 의존하지 않기 위해 설계되었다. 그림 6은 전이학습 모델에서 얻을 수 있는 특징 맵의 형태와 특징을 결합하여 분류하는 과정을 보여주고 있다.

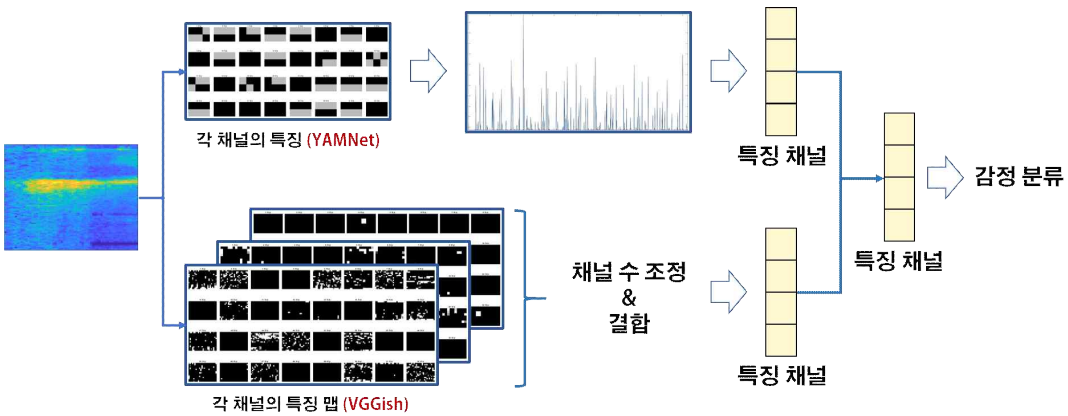


그림 6 융합 전이학습 모델 특징 추출 및 결합

2. Late-Fusion 방식 소개

Fusion 방식은 다른 데이터 소스나 다른 모델로부터 추출한 정보를 결합하여 독립적인 모델의 성능보다 더 나은 분류, 예측 결과를 얻기 위해서 융합하는 기술이다. 다른 모델에서 얻을 수 있는 정보와 다른 특징 추출 방법으로 얻은 정보를 결합하여 더 많은 경우를 학습할 수 있으며, 다중 센서 데이터의 정보를 통합할 때도 활용되는 방법이다. 위 논문에서는 YAMNet과 VGGish에 같은 데이터를 입력하고, 병렬로 학습을 진행하여 얻어진 정보를 적절한 결합을 통해 더 나은 분류 결과를 얻는 Late-fusion 방식을 선택하였다. 그림 7에서는 각 모델에 동일한 스펙트로그램 이미지를 입력했을 때, 분류 이전 계층에서 Grad CAM을 통해 활성화되는 영역을 표현하고 있다. 각 모델은 구조적 차이로 인해 VGGish는 이미지의 위치적인 특징(시간-주파수 영역)을 추출하여 높은 위치 해상도를 가지고 있지만 YAMNet의 경우 이미지의 전반적인 특징을 파악하고 위치적인 특징에 의존하지 않도록 설계하여 정보 분석의 차이가 발생한다. 그림 8에서는 전처리 과정을 거친 데이터를 입력하여 각 모델을 통해 독립적으로 감정 분류를 위한 패턴과 특징을 추출하게 된다. VGGish의 각 계층에서 얻어지는 정보를 더하여 새로운 특징 정보를 생성하여 총 3가지 종류의 데이터를 종합하여 마지막 분류 계층에서 특징 정보를 종합하고 감정을 분류하는 네트워크를 설계하였다. 특징을 결합할 때는 주로 더하는 방식이나 곱하는 방식을 사용하지만 위 네트워크에서는 감정마다 핵심이 되는 특징 맵의 채널이 다르고, 더하거나 곱함으로 그 특징 채널의 영향력이 감소하는 것을 줄이기 위해 깊이 결합을 진행하여 두 모델에서 얻은 정보를 유지하면서 결합하여 이후 완전 연결 계층에서 다양한 정보의 상호 작용을 통해 분류가 진행된다.

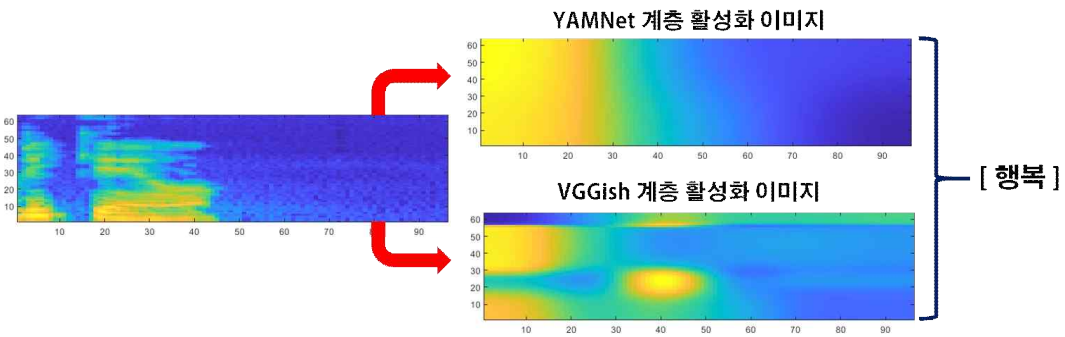


그림 7 YAMNet과 VGGish의 특징 추출 해상도 비교

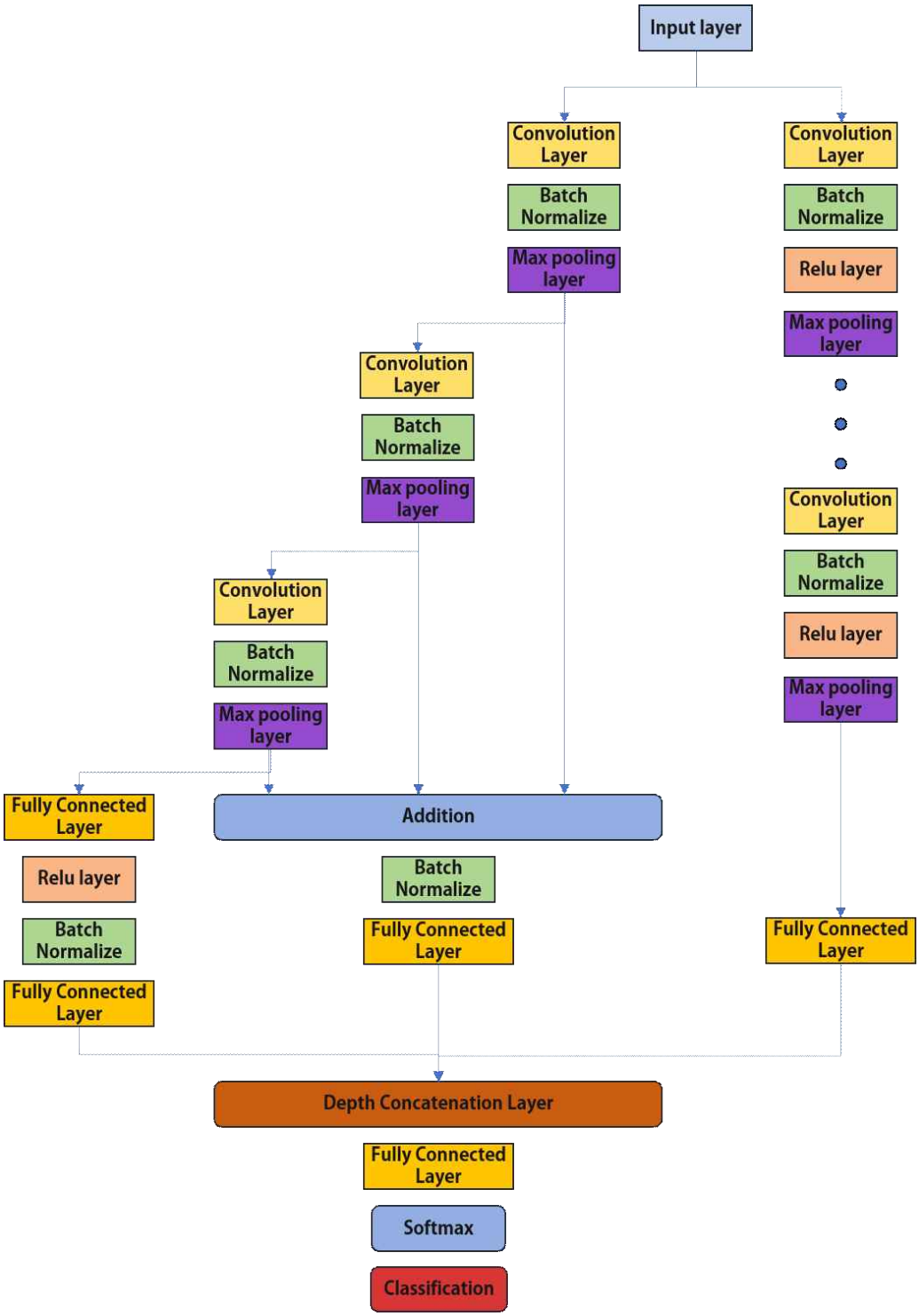


그림 8 VGGish와 YAMNet의 Late-fusion 네트워크

제3절 설명 가능 모델(XAI)

1. 모델의 집중 방향을 시각화하는 설명 가능 기법

자율 주행 자동차, 스마트 워치 등 인공지능을 활용한 다양한 기술들이 발전하여 일상생활과 로봇, 의료, 금융 등 많은 산업에 적용되고 있다. 많은 분야에서 적용되고 있는 인공지능이 사용자의 판단을 보조하고 생산 등의 활동에 활용될 수 있으려면 도출된 결과의 근거와 타당성에 관한 확인이 필요하다. 그림 9은 기계 학습과 딥러닝 모델에 입력되는 데이터와 라벨을 사용하여, 입력되는 데이터에 영향을 미치는 요소와 영역을 이해하고 해석할 수 있는 설명 가능 모델(eXplainable AI)을 보여준다. 위 기술은 모델을 사용하는 사람에게 데이터의 어떤 부분이 출력에 영향을 주었는지 이해하고 신뢰성을 높일 수 있으며, 모델이 도출한 잘못된 분류와 예측을 확인하여 진행한 모델의 성능을 높일 수 있도록 수정하고, 어떠한 이유로 잘못된 출력이 나왔는지를 통해 모델을 사용하는 사용자를 이해시킬 수 있게 되며 신뢰성을 높인다. 의사 결정 트리나 선형 회귀의 경우 모델이 단순하므로 모델 자체에서 의사 결정의 과정을 확인할 수 있고, 그 외에 복잡한 모델들은 후에 Grad CAM, Activation, Occlusion Sensitivity, LIME 등의 처리가 필요하다.

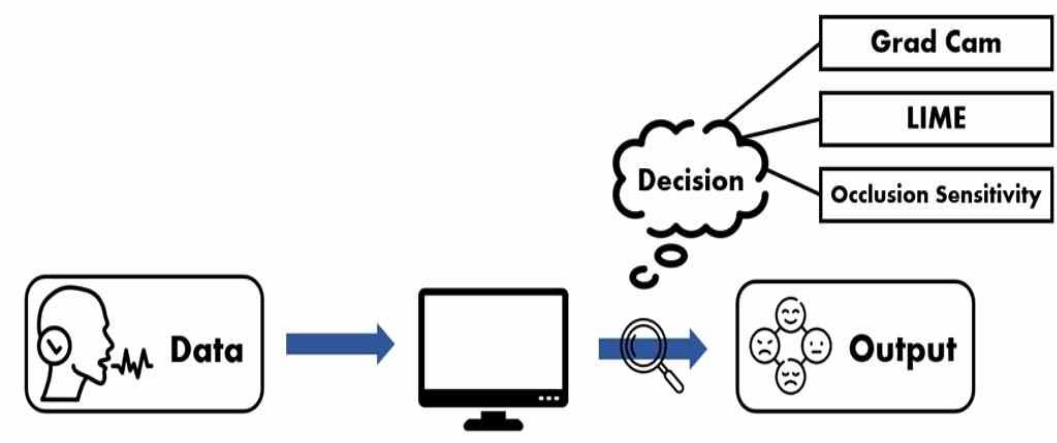


그림 9 설명 가능한 인공지능 설명

가. Grad CAM 기법

Grad CAM(Gradient-weighted Class Activation Mapping)은 합성곱 신경망 모델의 클래스 분류를 설명하기 위해 역전파 그래디언트(Backpropagation Gradient)를 활용하여 분류 클래스에 대한 기울기로 집중 영역을 생성한다. 입력된 이미지는 합성곱 네트워크를 통해 통과하며 특징 벡터를 생성하고, 특징 벡터를 활용하여 클래스가 분류되게 된다. 분류된 클래스에 대한 점수로 이전 계층에서 각 채널에 대한 그래디언트를 각 채널이 클래스 점수에 얼마나 영향을 미치는 지로 가중치로 활용하여 특징 맵에서 클래스와 관련된 공간 영역을 구하고 원본 크기로 복원하여 시각적으로 표현한다. Grad CAM은 그림 10에서 알 수 있는 바와 같이 음성 데이터의 시간과 주파수 영역을 동시에 확인할 수 있는 스펙트로그램 이미지를 입력하여 합성곱 신경망을 통과하게 된다. 이후 이미지에 히트맵으로 모델의 클래스 분류 집중 중요도 정도를 확인하고, 입력한 데이터와 같은 크기로 복원하여 겹쳐서 어떤 시간, 주파수 구간이 클래스 분류에 핵심적인지 확인한다.

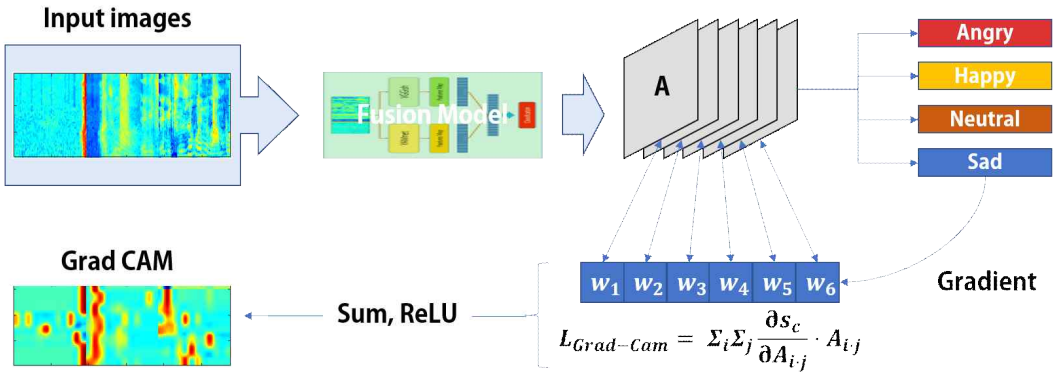


그림 10 Grad-CAM 생성과정

나. LIME 기법

LIME(Local Interpretable Model-agnostic Explanations) 방법은 입력되는 데이터의 픽셀과 같은 작은 영역이 모델의 클래스 분류에 얼마나 영향을 미치는지 확인하는 설명 가능 기법이다. 위 방법은 그림 11에서 알 수 있는 바와 같이 입력 이미지의 특정 영역을 잡음을 추가하는 방법 등으로 변형을 주어서 변형된 영역들을 모델에 입력하여 모델의 예측 결과를 측정한다. 각 픽셀 또는 특정 영역에 대한 측정된 특징값으로 벡터를 생성하여, 위 특징 벡터를 해당하는 이미지들의 예측 결과를 사용하여 선형 회귀와 같은 간단한 모델로 학습을 진행한다. 학습된 모델을 통해 원본 이미지의 각 픽셀이나 특정 영역이 이미지 클래스 분류 영향력을 계산하여 히트맵으로 시각화한다. 스펙트로그램 이미지에서 다양한 시간-주파수 픽셀 영역에 변형을 주어 예측된 결과를 측정하고, 영향력을 학습 시켜서 어떠한 영역이 클래스 분류에 영향을 주었는지 확인한다.

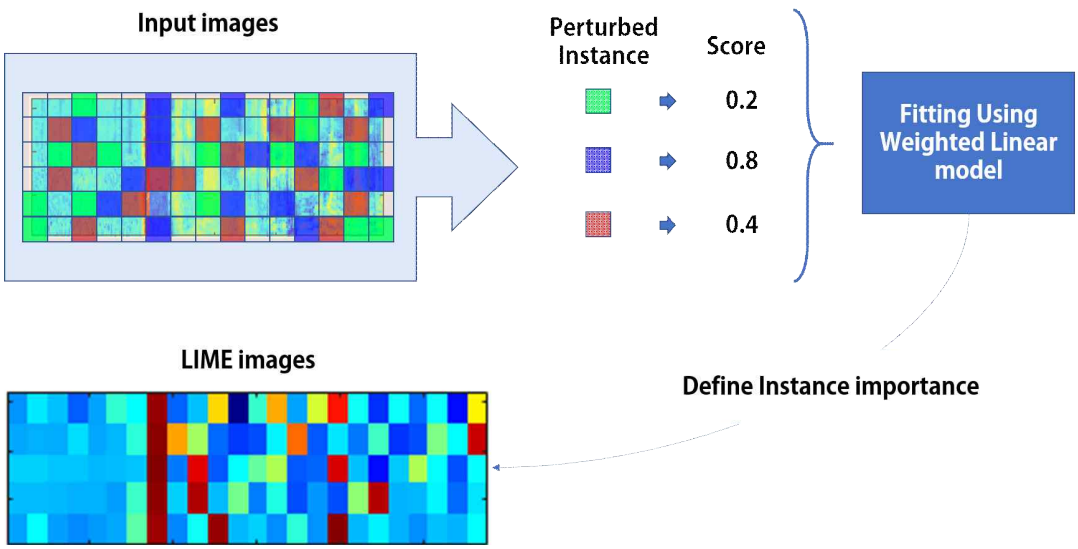


그림 11 LIME 생성과정

다. Occlusion Sensitivity 기법

그림 12는 Occlusion Sensitivity를 진행하기 위해 모델에 입력되는 이미지의 특정 영역을 가려서 모델에 입력하여 결과를 확인하고 클래스 점수와 특징 맵의 활성화 여부를 파악하는 방법을 보여주고 있다. 가려지는 영역을 지속해서 위치를 변경하여 각 위치가 모델의 예측에 얼마나 큰 영향을 주는지 시각적으로 표현하여 준다. 위 방법은 이미지뿐만 아니라 음성이나 텍스트의 특정 부분을 가리면서 해당 영역을 무작위 값이나 평균값 등으로 채우고 모델에 입력하여 정확도의 변화를 확인하는 것을 반복하여 영역의 영향력을 확인한다.

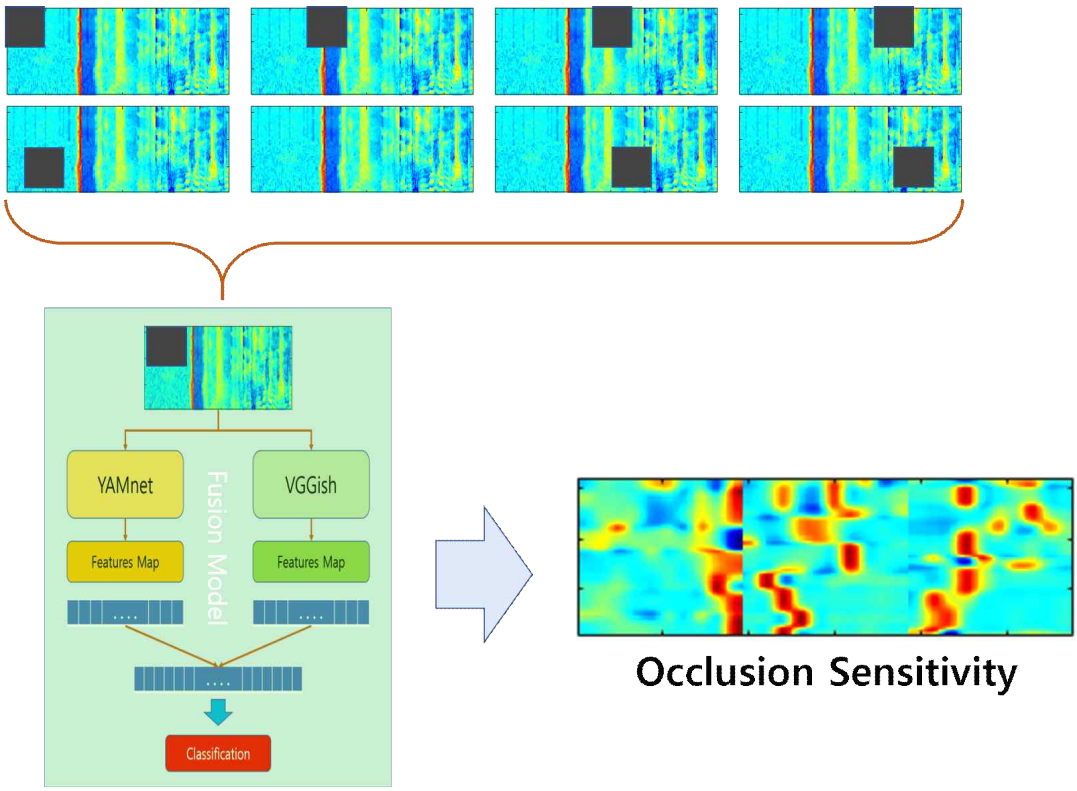


그림 12 Occlusion Sensitivity 생성과정

제4장 실험 및 결과

제1절 데이터 셋 소개

1. CSU 2021 일반인 대상 음성감정 데이터 셋

위 데이터 셋은 조선대학교에서 2021년도에 일반인, 극단 배우, AI 캐릭터를 통해서 행복(Happy), 무감정(Neutral), 화남(Angry), 슬픔(Sad)의 4가지 감정을 총 100명에게서 취득하여 데이터를 구축하였다. 데이터는 드라마 또는 영화에 나오는 4가지 감정에 관련된 대사들을 각각 10개씩 총 40개의 대사를 수집하여 취득자들이 연기하는 목소리를 그림 13과 같이 녹음하였다. 잡음을 최소화하기 위해서 소음이 없는 조용한 공간에서 소니의 스테레오 핀 마이크(ECM-LV1)를 사용하여 극단 배우와 일반인들은 4가지 감정을 10개씩 녹음하여 총 40개의 음성을 녹음하였다. AI 캐릭터는 인공지능 성우 서비스인 타입캐스트(Typecast), 프로소디 프로그램을 사용하여 사람과 같은 환경에서 취득하기 위해 1개의 AI 캐릭터를 통해 기본 음성 10개, 속도와 톤을 조정한 음성 10개를 생성하여, 감정당 20개씩 총 80개의 음성 데이터를 생성하고 음성을 마이크를 통해 다시 녹음하여 그림 14와 같이 구성하였고, 그림 15와 같이 나타내진다.

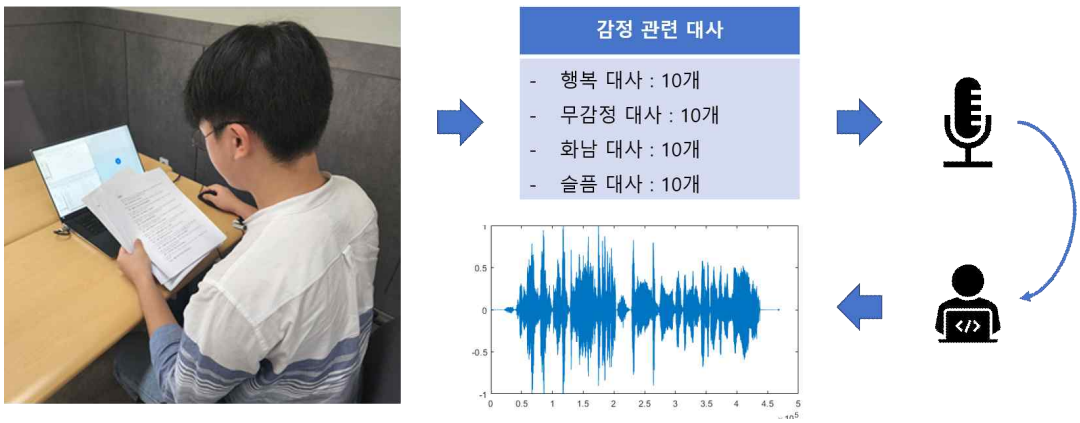


그림 13 데이터 취득 과정

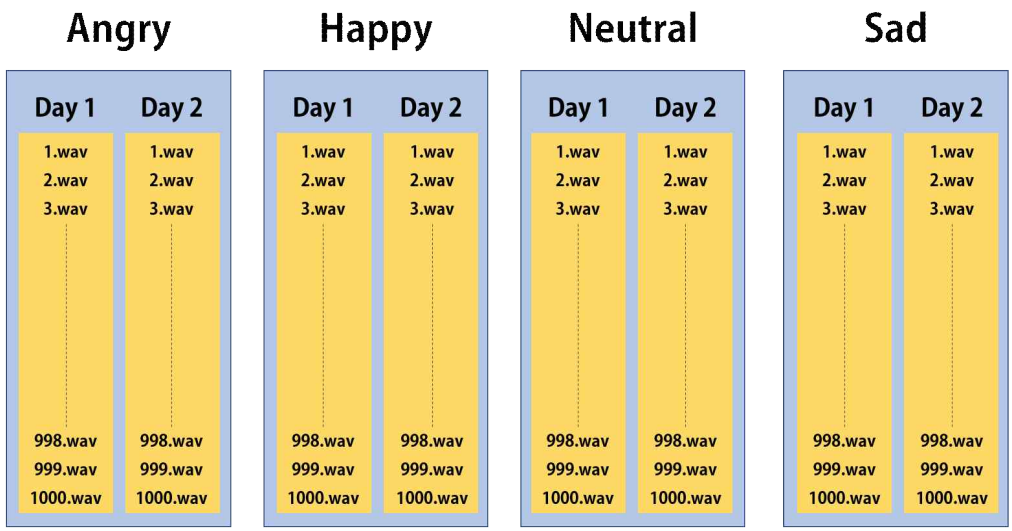


그림 14 CSU 2021 일반인 대상 음성 감정 데이터 셋

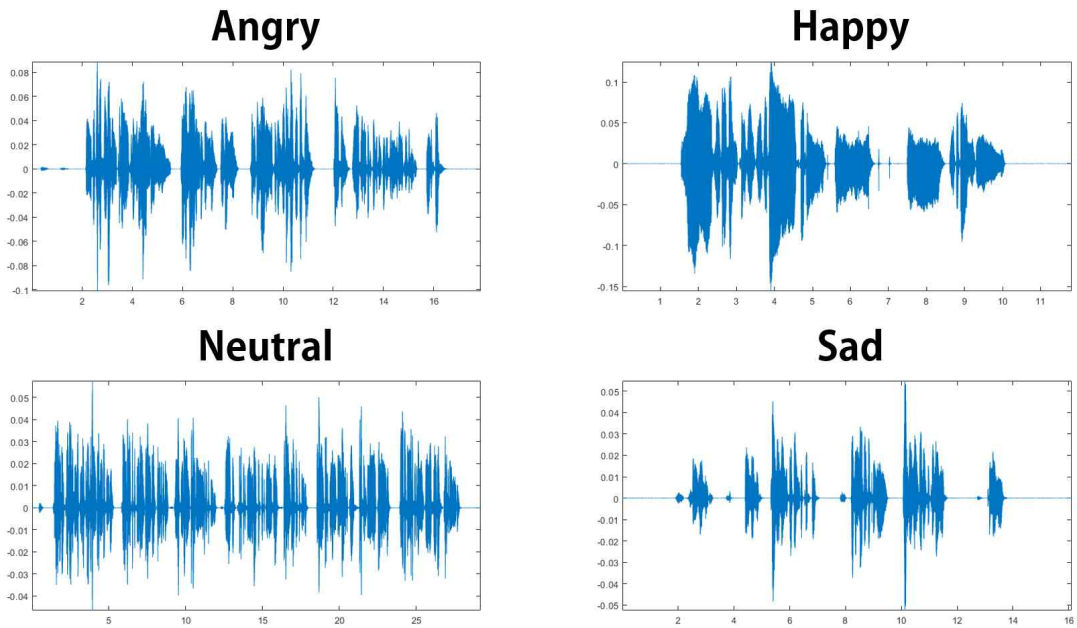


그림 15 CSU 2021 데이터 표본

2. CSU 2022 일반인 대상 음성 감정 데이터 셋

위 데이터 셋은 조선대학교에서 2022년도에 일반인 200명을 대상으로 행복(Happy), 무감정(Neutral), 화남(Angry), 슬픔(Sad), 억울(Chagrin), 역겨움(Disgust), 공포(Fear), 놀람(Surprised)으로 총 8가지 감정 상태 분류용 데이터를 취득하였다. 각 감정은 해당하는 감정에 맞는 상황과 그에 맞는 짧은 대사를 10개씩 선정하여 참가자들의 감정을 담아 연기할 수 있도록 준비하며, 취득 장비로는 소니의 스테레오 핀 마이크(ECM-LV1)를 사용하여 48kHz로 최대한 잡음이 없도록 소음이 없는 조용한 공간에서 혼자 녹음을 하였다. 데이터는 각 참가자가 감정당 10개씩 녹음하여 총 80개, 전체 인원이 16,000개의 wav 형식으로 그림 16과 같이 구성하였고, 그림 17과 같이 나타내진다.

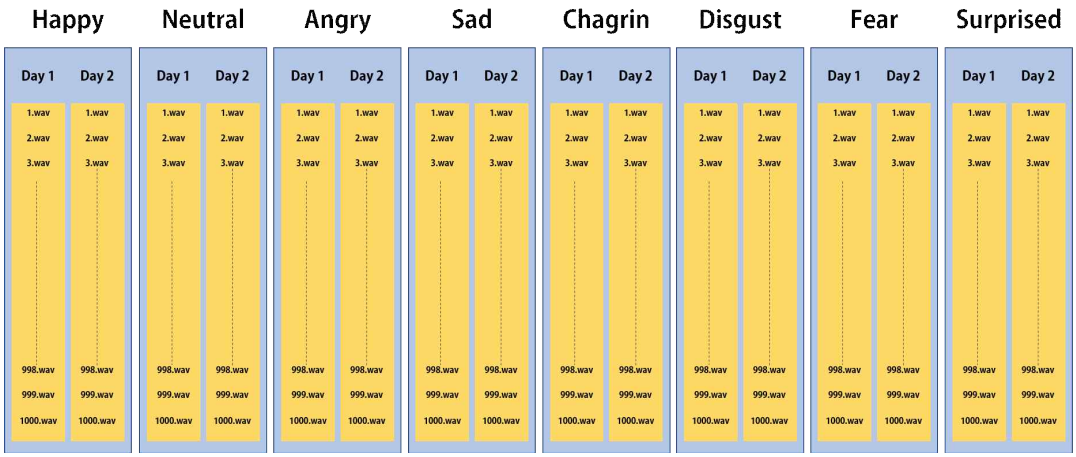


그림 16 CSU 2022 일반인 대상 음성감정 데이터 셋

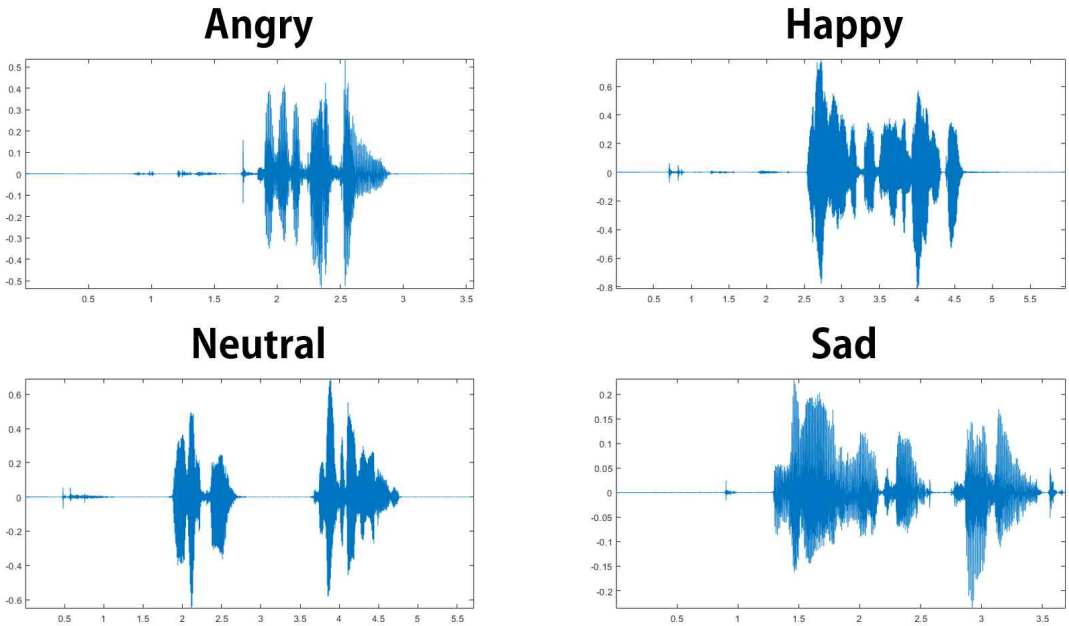


그림 17 CSU 2022 데이터 표본

3. AI-HUB, 감정 분류용 데이터 셋

위 데이터는 AI-Hub의 공용 데이터로, AI-Hub은 AI 기술 및 제품, 서비스 개발에 필요한 AI 인프라(AI 데이터, AI SW API, 컴퓨팅 자원)를 지원함으로 누구나 활용하고 참여할 수 있는 AI 통합 플랫폼이다. 기존 공개 데이터 중 FER 2013은 감정별로 데이터 개수가 큰 차이를 보이며, 얼굴 감정 데이터가 저화질로 취득되었고, SFEW 데이터의 경우 고화질 데이터이지만 2,000건 정도로 규모가 작았다. CMU-MOSI의 경우 발화 도메인이 영화 리뷰로 한정되었고 감정 분류도 긍정과 부정으로만 되어있는 단점 때문에 위 데이터가 구축되었다. 연기 지망생과 연기 전문가 100명을 대상으로 행복(Happy), 놀람(Surprised), 무감정(Neutral), 공포(Fear), 역겨움(Disgust), 화남(Angry), 슬픔(Sadness) 7가지 감정을 약 100번씩 발화 및 연기 수행하여 총 10,351개의 영상과 음성을 그림 18과 같이 구성되었다. 음성 데이터는 그림 19와 같이 나타내어진다.

그림 18 AI-hub 감정 분류용 데이터 셋

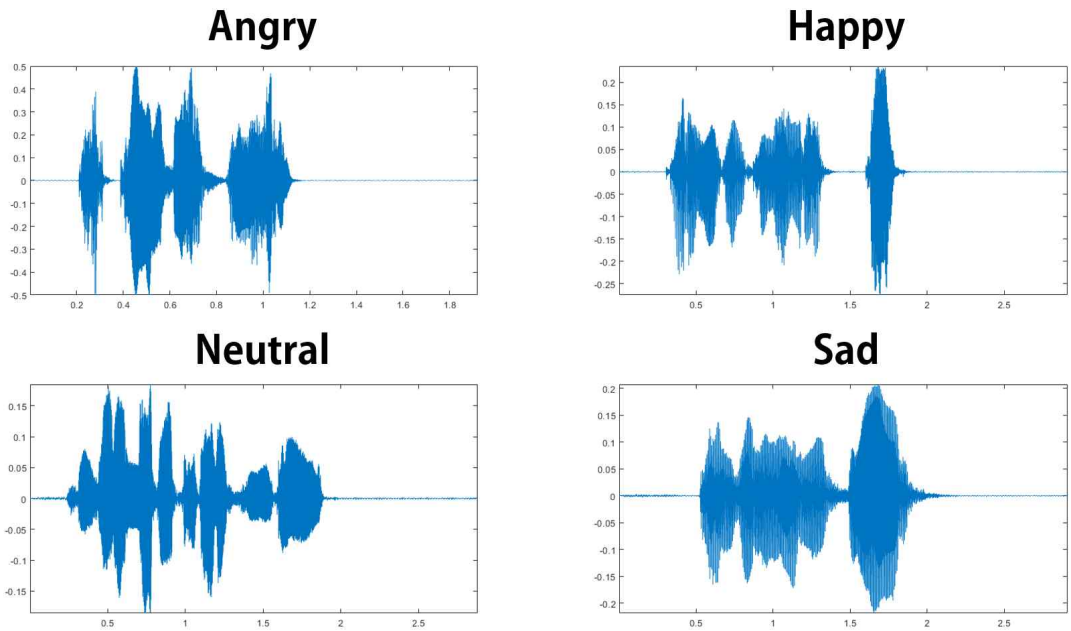


그림 19 AI-hub 감정 분류용 데이터 표본

4. 데이터 전처리 및 데이터 선별

음성을 모델의 입력 크기에 따라 시간 영역으로 분할하게 될 때, 분류와 예측하려 하는 특징과 상관없는 무의미한 구간이 발생하게 된다. 이 구간을 효과적으로 선별 및 제외하기 위해서 본 논문에서는 가우시안과 상관 계수를 이용한 GDS 전처리를 진행하였고, 그림 20는 GDS를 적용하기 전의 각 학습 데이터의 분할 개수와 적용 후의 학습 데이터의 분할 개수를 나타내고 있다. 적용 전의 데이터는 분할되어 총 165,240개의 데이터로 구성되었지만, 적용 후에 불필요한 구간을 제외함으로 140,885개로, 총 24,385개의 데이터를 제외해 15%를 단축했다. 이 결과는 학습 시간과 소모되는 컴퓨팅 자원까지 절약할 수 있음을 확인한다.

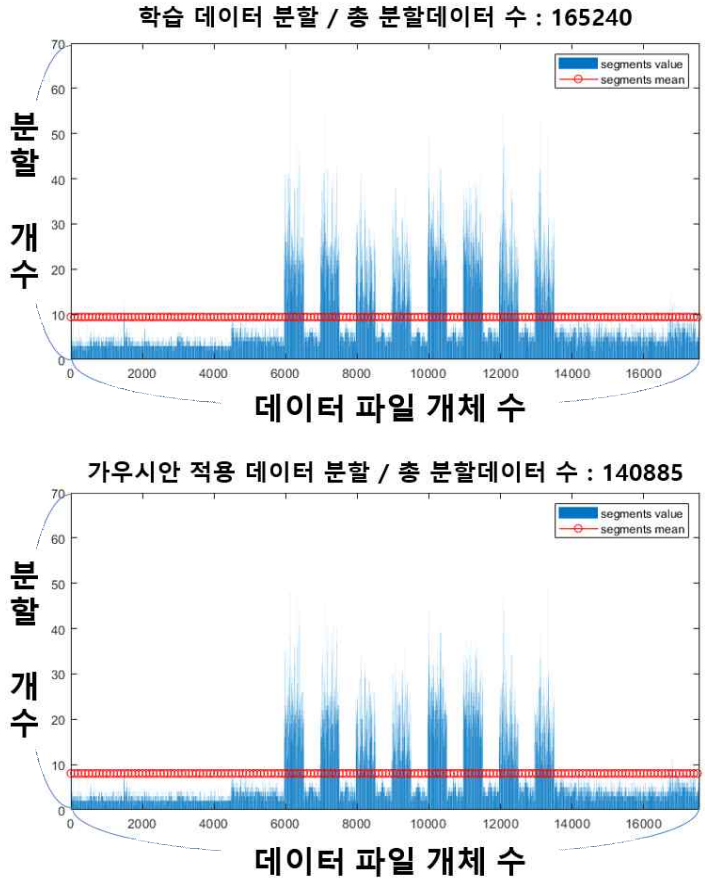


그림 20 데이터 선별 후 변화된 분할 데이터 수

제2절 실험 및 결과

1. 학습 환경 및 파라미터

실험을 위해 사용된 하드웨어는 NVIDIA GeForce GTX TITAN X와 32G RAM, Intel(R) Xeon(R) CPU E5-1650 v3를 사용하고, MATLAB 2023a 환경에서 진행한다. 실험에서는 하드웨어의 성능과 실험 시간을 고려하여 로그-멜 스펙트로그램으로 변환하기 위한 오디오 필터와 STFT의 파라미터를 조절한다. 표 3은 데이터 전처리 이후 설계한 네트워크에 학습하기 위해 사용된 하이퍼 파라미터를 나타낸다. 최적화 함수는 딥러닝에서 자주 사용되는 Adam(Adaptive Moment Estimation)을 사용한다.

	최적화 함수	학습률	반복 횟수	미니 배치 사이즈
학습 옵션	Adam	0.01	20	128

표 3 학습 하이퍼 파라미터

2. 클래스별 정확도 분석

위 논문에서 제시한 모델을 사용하여 분할된 음성 구간에 대해서 분류를 진행하였고, 원본 데이터에 해당하는 구간들의 정확도 값을 종합하여 평균값으로 최종적으로 클래스를 정의한다. 그림 21에서는 제시한 융합 모델의 정확도를 보여주고 있으며, 4가지 핵심 감정에 대한 분류에서 높은 정확도를 보여주고 있다. 그림 22에서는 GDS를 적용하여 학습에 영향을 주지 않는 데이터를 제거하여 학습한 네트워크에 같은 데이터 선별을 적용한 실험 데이터로 정확도를 분석한다. 전체 데이터의 수에는 영향이 없으며, 데이터에 해당하는 분할 구간의 개수가 전체 15% 감소하면서 학습 시간은 감소하면서 정확도가 유사함을 확인할 수 있다.

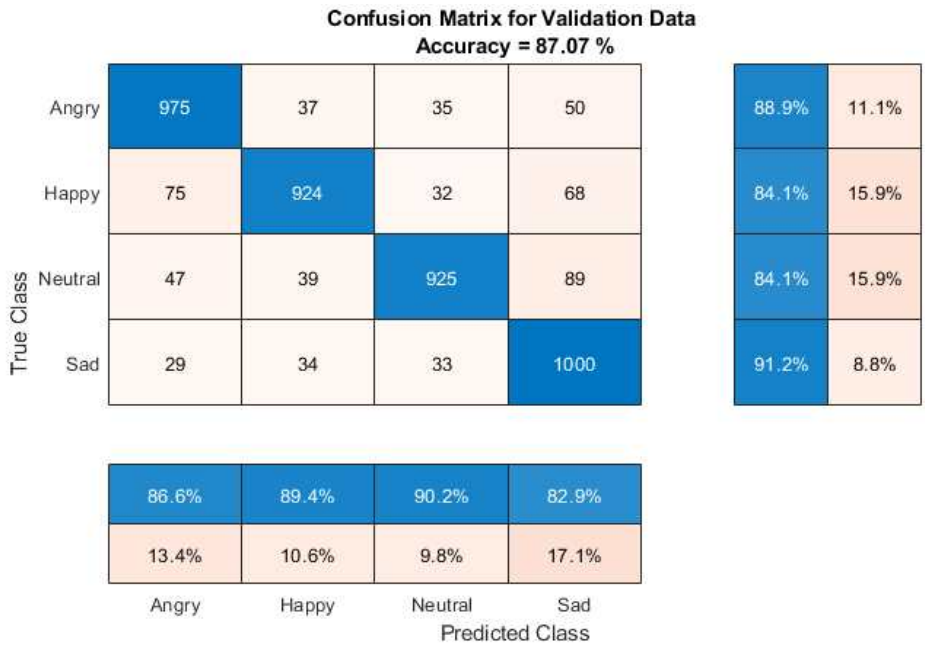


그림 21 Custom YAMNet-VGGish Late Fusion Network, 융합모델 분류 정확도

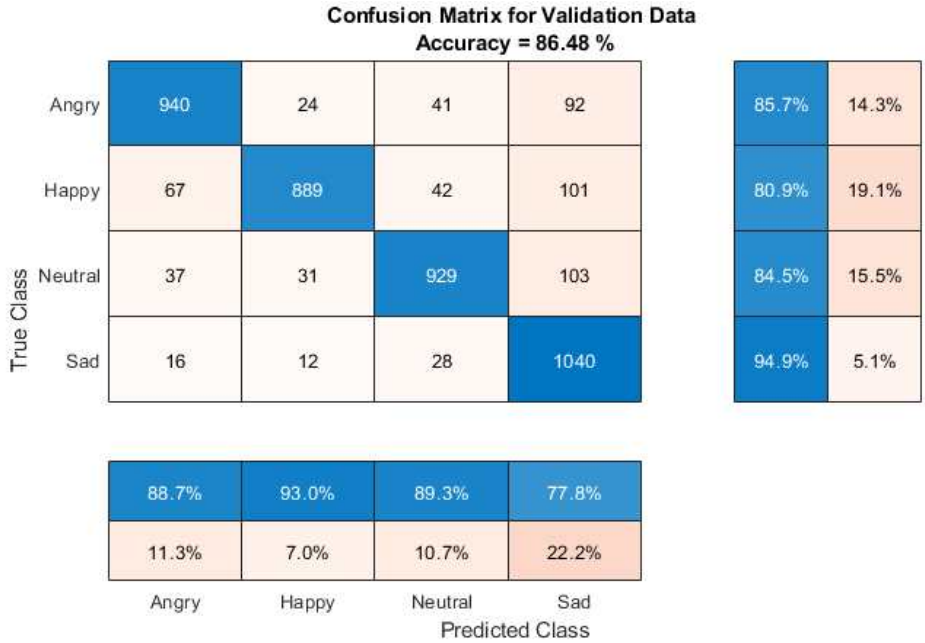


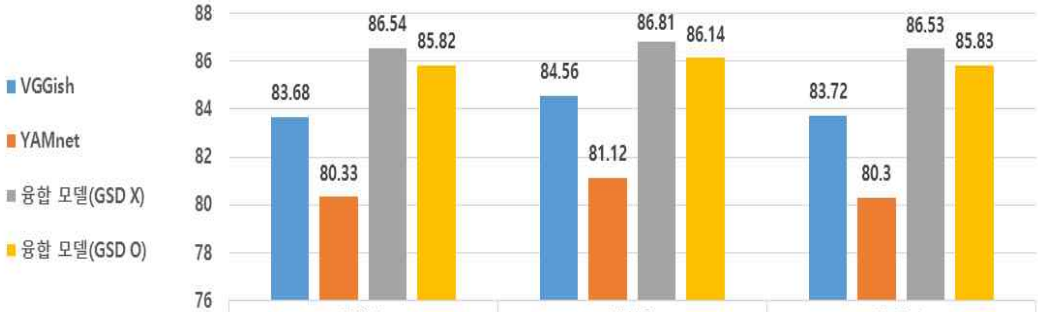
그림 22 가우시안 필터 적용 분류 정확도

표 4에서는 GDS를 적용하기 전과 적용 후의 분류 성능 평가를 확인할 수 있다. 학습한 모델들은 유사한 분류 정확도를 가졌지만, 학습 시간이 약 90분에서 70분으로, 22% 시간과 메모리 소모를 감소한다. 위와 같이 분류한 네트워크에 다양한 설명 가능 모델을 적용하여 집중 영역을 분석하여 어떤 구간을 통해 올바른 분류를 진행하고, 잘못된 분류를 진행했는지 확인하여 모델 사용자에게 이해시키고 향후 모델의 성능을 높이기 위한 작업을 진행할 수 있다.

	정확도	재현율	F1 점수	학습 시간
GDS 적용 전	86.54%	0.8681	0.8653	89.87(min)
GDS 적용 후	85.82%	0.8614	0.8583	70.62(min)

표 4 모델 분류 성능 평가

그림 23에서는 설계한 융합 모델에 GDS를 적용 전후와 기존의 전이 학습 모델인 VGGish와 YAMNet를 정확도와 재현율(Recall), F1 점수를 통해 모델 성능을 비교한다. 재현율은 모델이 실제 양성인 것 중에서 양성이라고 예측한 것의 비율을 나타내는 성능 지표이다. F1점수는 모델의 정밀도와 재현율의 조화 평균으로, 두 가지를 고려하여 성능을 나타내는 지표이다. YAMNet의 경우 가장 낮은 성능을 보여주고, VGGish는 YAMNet보다 높은 성능을 보여준다. 설계한 모델의 경우 기존의 전이 학습 모델보다 모든 성능 지표에서 높아진 성능을 보여주고, GDS를 적용하기 전과 후는 작은 차이를 보여준다.



	정확도	재현율	F1 점수
VGGish	83.68	84.56	83.72
YAMnet	80.33	81.12	80.3
융합 모델(GSD X)	86.54	86.81	86.53
융합 모델(GSD O)	85.82	86.14	85.83

그림 23 기존 모델 성능 비교(Accuracy, Recall, F1score)

3. 설명 가능 모델을 적용한 집중 영역 분석

입력되는 음성 데이터의 시간-주파수 특징을 효과적으로 모델에게 학습시키기 위해 다양한 전처리를 진행하고, 분류를 진행하게 된다. 이때 사용자가 모델이 어떤 특징 요소를 가지고 감정을 분류하였는지 확인하기 위해 다양한 설명 가능 기법이 적용된다. 그림 24에서는 각 감정에 해당하는 데이터를 학습된 모델에 Grad CAM을 적용하여 핵심 영역을 가져오고, 전체 데이터의 통계를 바탕으로 감정별로 활성화되는 주파수 영역의 특징을 알 수 있다. 분노와 행복 감정의 경우 상대적으로 높은 위치의 주파수 영역이 공통으로 활성화되고, 무감정과 슬픔의 경우는 더 낮은 위치의 주파수 영역이 활성화된다.

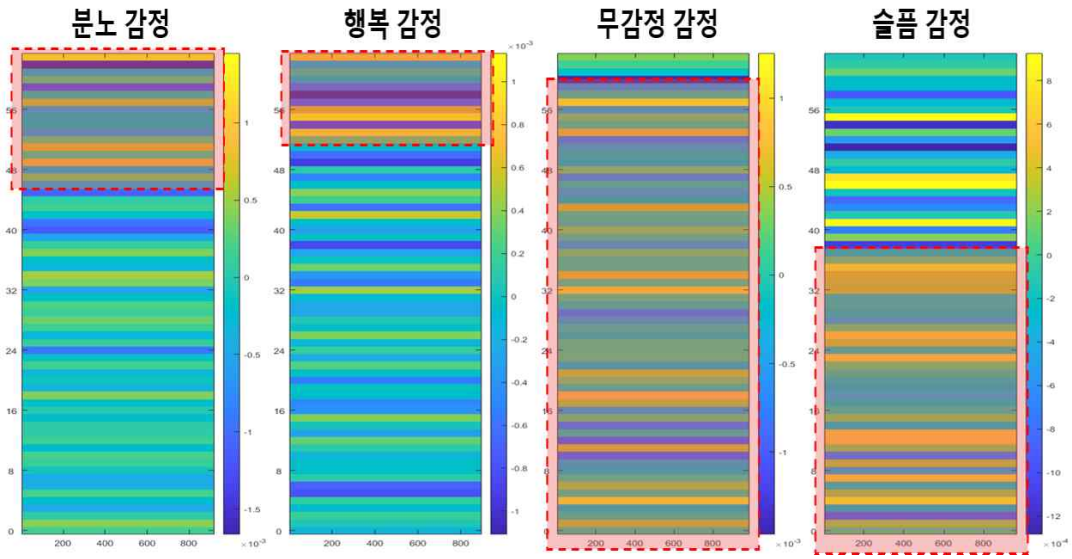


그림 24 감정의 활성화 주파수 영역

그림 25에서는 분노와 관련된 대사를 통해 얻은 음성을 스펙트로그램으로 변환하고, 학습된 모델과 Grad CAM으로 집중 영역을 분석했을 때, 활성화되는 영역을 보여주고 있다. 위 방식을 통해 어떤 시간-주파수 영역이 모델의 의사 결정에 영향을 줬는지 붉은 영역에서 확인할 수 있다. 큰 값을 가진 영역만 모델이 집중한 것이 아닌 다양한 값의 영역을 종합하여 감정의 분류를 진행한 것을 확인한다.

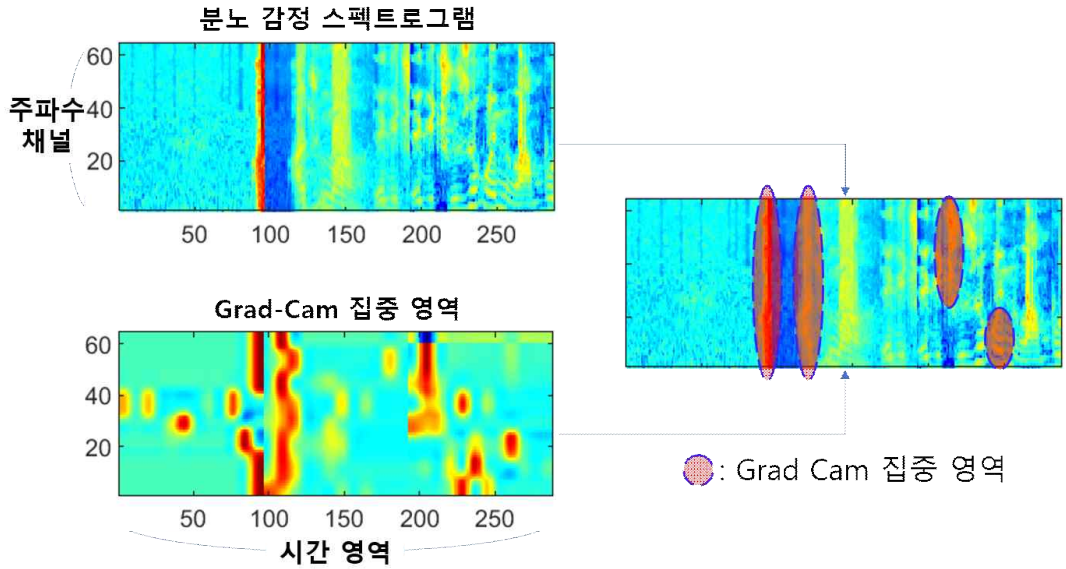


그림 25 분노 감정 스펙트로그램 및 Grad CAM 집중 영역

그림 26~29에서는 음성 스펙트로그램을 Grad CAM과 LIME, Occlusion Sensitivity를 적용하여 모델의 의사 결정 과정을 확인할 수 있다. 설명 가능 기법을 적용한 이미지를 보게 되면, 입력된 음성 스펙트로그램의 어떤 영역이 모델에게 분류 기준에 영향을 주었는지 선명하게 확인할 수 있다. Grad CAM의 경우 모델 내부에서 클래스에 대한 점수를 통해 그래디언트 정보를 활용하므로 어떤 영역이 중요한지에 대한 높은 해석력을 가지고 있다. LIME과 Occlusion Sensitivity의 경우 일정 영역이 모델 분류에 얼마나 영향을 주었는지 확인함으로 Grad CAM과는 다른 측면으로 어떤 영역이 분류에 영향을 주었는지 확인할 수 있다.

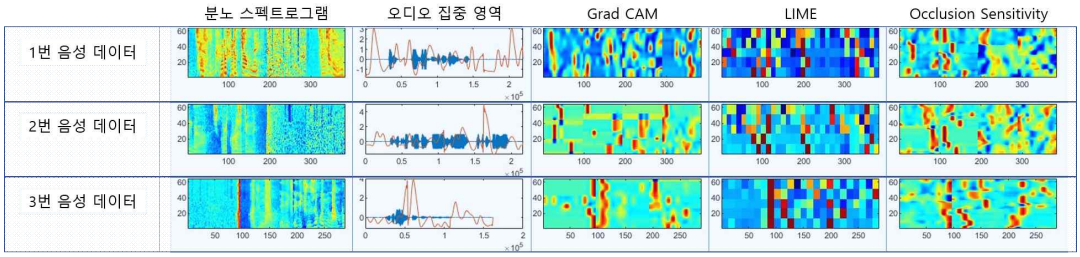


그림 26 분노 감정 스펙트로그램 (원신호, Grad CAM, LIME, Occlusion Sensitivity)

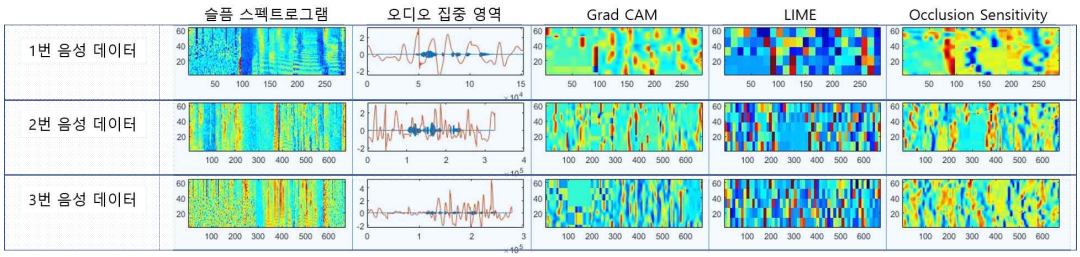


그림 27 슬픔 감정 스펙트로그램 (원신호, Grad CAM, LIME, Occlusion Sensitivity)

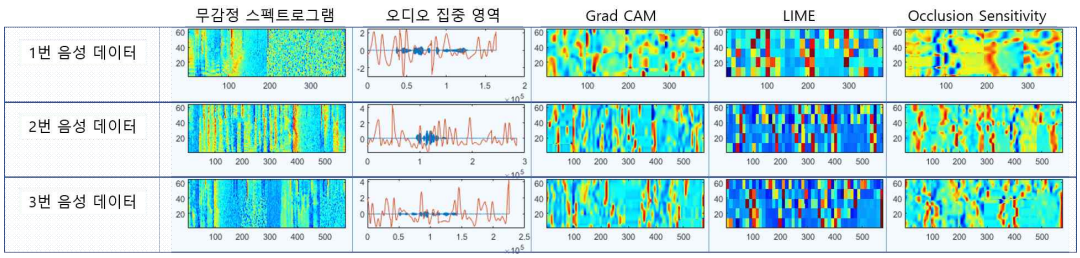


그림 28 무감정 스펙트로그램 (원신호, Grad CAM, LIME, Occlusion Sensitivity)

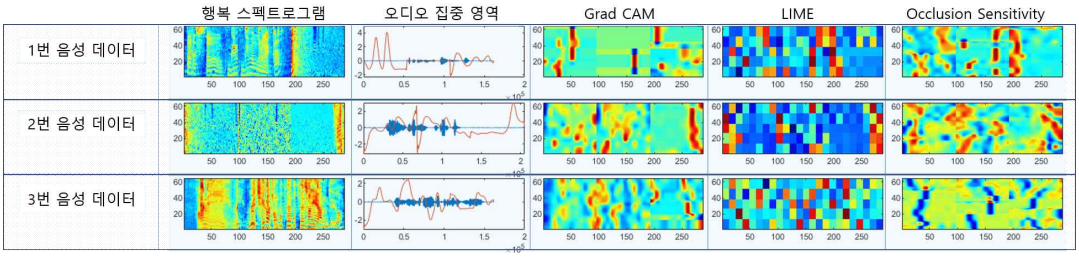


그림 29 행복 감정 스펙트로그램 (원신호, Grad CAM, LIME, Occlusion Sensitivity)

4. 설명 가능 모델의 집중 영역을 적용한 오디오

그림 30에서는 입력되는 이미지에서의 주파수 특징이 아닌 시간 특징을 중심으로 확인하기 위해 Grad CAM에서 얻어진 주파수의 영향력을 시간 영역으로 합하여 진행한다. 특정 시간 영역에서 많은 주파수 영역이 집중되었다는 것은 해당 시간에서의 단어나 억양이 클래스 분류에 중요한 역할을 함을 적용한다. 스펙트로그램으로 변환하기 위해 사용된 파라미터를 역으로 활용하여 원본 신호의 크기로 복원하고 음성에 적용하게 되면, 잡음이 있어서 청각적으로 원활히 파악하기 어렵기 때문에 필터값이 1로 구성된 합성곱을 적용하여 빠르게 진동하는 부분을 부드럽게 안정시켜 음성으로 들을 때 자연스러워질 수 있도록 후 처리하였다. 이후 활성화 값이 양수로 나타나는 영역의 오디오 음성을 추출하여 어떤 단어와 어절, 억양이 분류 의사 결정에 영향을 주었는지 확인한다.

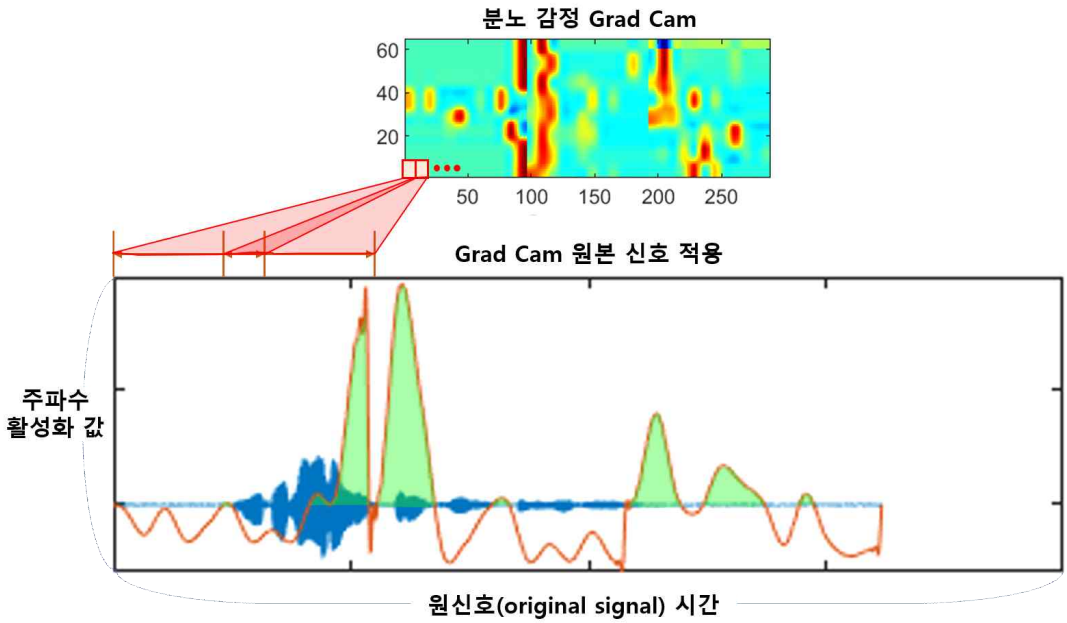


그림 30 Grad CAM 집중 영역 이미지 음성 신호로 복원

그림 31에서는 행복 감정의 대사 중, “와 무조건 콜이지, 나 스키장 너무 좋아” 라는 대사를 모델이 분류하였을 때 집중한 영역을 오디오 해상도로 적용하여 분석한 것이다. 시각적으로 표현되는 활성화 영역을 음성에 적용하여 분석하게 되면 “무조건 콜” 와 “너무 좋아” 라는 구간에서 모델이 집중했다는 것을 들을 수 있다. 감정마다 단어적인 요소에 집중하거나 억양에서 감정 특징이 나타나는 것을 오디오를 통해 확인할 수 있다. 무감정 대사의 경우 전체 대사에 균등하게 집중된 모습을 확인한다.

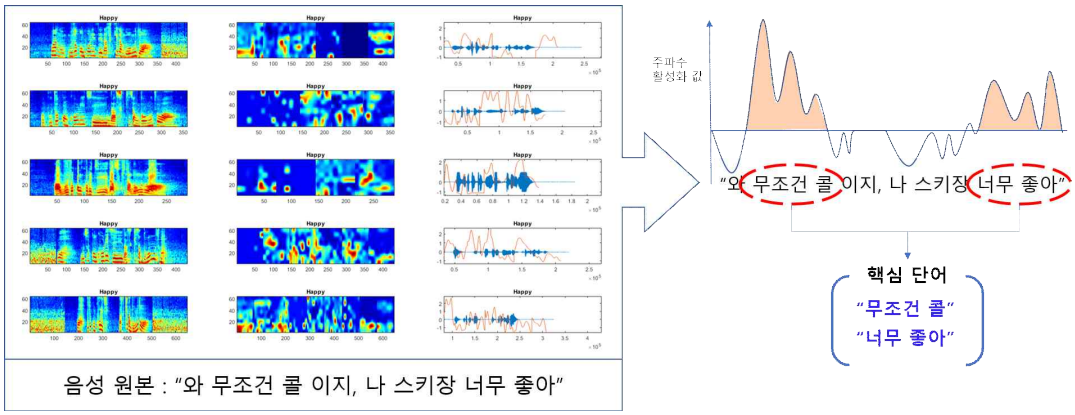


그림 31 행복 감정 Grad CAM 집중 영역 음성 데이터 적용

제5장 결론

본 논문에서는 음성 데이터를 로그-멜 스펙트로그램 이미지로 변환하고 데이터 선별 알고리즘과 Late-fusion 모델을 활용하여 다양한 음성에 적응할 수 있으면서 필요한 데이터만을 학습하여 짧은 학습 시간에 높은 분류 정확도를 갖도록 하였다. 분류된 결과에 대해서 모델의 의사 결정 과정을 시각적으로 확인하기 위해서 다양한 설명 가능 모델을 통해 입력되는 데이터에서 핵심 되는 시간-주파수 영역을 시각적으로 확인했다. 활성화 영역을 주파수와 시간 영역에서 확인하여 감정마다 중심이 되는 주파수 영역과 시간 영역에서 감정에 중점이 되었던 단어와 억양 등을 연구하였다. 음성은 사람의 정신적, 신체적인 상태를 표현하면서 자기 의사를 전달하는 대표적인 방법이다. 이러한 음성을 컴퓨터나 상호 작용이 가능한 웨어러블 장치가 이해하고 분석하기 위해서는 잡음을 제거와 더 많은 정보를 얻기 위한 적절한 전처리 및 특징 추출이 필요하다. 전처리한 음성은 개인의 감정과 의사 표현 분석 등에 활용하여 실시간으로 감정 변화를 관찰하거나 심리 상담, 스트레스 분석, 주변 환경 영향 분석, 스마트 홈과 스마트 카 등 다양한 환경과 상황에서 응용할 수 있다. 위 연구에서는 조선대학교에서 2021년도, 2022년도 감정 분류를 위해 취득된 음성 데이터와 감정 분류를 위한 AI-Hub의 공용 데이터에서 더 많은 정보를 얻으며 음성 특징에서 중요한 주파수의 시간 영역 변화를 활용하기 위해서 시간-주파수 변화를 나타내는 스펙트로그램 이미지로 변환하였다. 이미지는 모델의 입력 크기에 맞게 일정 시간 구간 중첩하여 나누게 되는데, 이러한 과정에서 감정과 관련 없는 의미 없는 구간이 발생하게 된다. 침묵, 끊김, 잡음 등의 학습에 영향을 주지 않으면서 감정과 관련된 정보가 담겨 있지 않은 구간이 발생하게 되면 학습 데이터의 규모가 커지면서 많은 데이터 자원이 소모하게 된다. 영향을 주지 않는 구간을 효과적으로 선별하여 제외하기 위해서 일정 크기로 분할된 구간의 평균과 분산을 활용하여 가우시안 분포로 나타내었고, 각 구간마다의 분포도를 상관 계수를 통해 가장 이질적인 데이터를 선별하여 학습에서 제외함으로 데이터의 규모를 15% 낮추고 학습 시간 또한 20% 단축하게 하고 정확도를 유지할 수 있는 효과를 확인하였다. 다양한 사람과 환경에서 취득한 데이터를 일반화 있게 학습하기 위해서 음성 분류에 사용되는 YAMNet과 VGGish를 Late fusion 하여 활용하였다. 각 모델은 같은 스펙트로그램 이미지를 입력으로 받아서 병렬로 학습하였고, 분류 계층에서 얻어진 정보를 공유하며 더 다양한 음성에서 감정 분류를 가능한 모델로 설계하였으며, 87% 분류 정확도를 보여주었다.

또한, 설계된 모델이 입력받은 스펙트로그램 이미지에 대해서 이러한 의사 결정을 내린 이유를 시각적으로 확인하기 위한 설명 가능 모델(Grad CAM, LIME, Occlusion Sensitivity)을 사용하였고, 얻어진 Grad CAM에서의 활성화 영역을 주파수 영역으로 합하여 감정마다 다른 주파수 영역에 집중하였음을 확인했다. 또한 원신호에 적용함으로 음성 취득할 때 사용한 대사와 음성 데이터에서 감정 분류에 영향을 주었던 단어와 음정 변화 등을 분석하였다. 음성을 통해 다양한 분야에서 활용되고 있으면서, 모델의 판단 결과가 사용자의 의사 결정을 충분히 보조할 수 있는지 사용자에게 신뢰를 주고, 어떤 이유로 이와 같은 판단을 했는지 이해를 시키는 것이 중요해졌다. 위 연구를 통해 입력되는 스펙트로그램 이미지에서 핵심이 되는 시간-주파수 영역을 확인하면서 어떤 주파수 영역이 각 감정에 중심이 되고, 어떤 단어와 음정 변화가 감정마다 다르게 집중되었는지 확인할 수 있었다. 이러한 연구는 향후 음성을 활용한 더 많은 감정 분류 모델에도 적용할 수 있으며, 우울증과 텍스트와 함께 사용하는 멀티 모달 기술에도 활용될 수 있을 것으로 기대한다.

참 고 문 헌

[1] Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering, 14, 342-356.

[2] Cho, J., and Kim, B. (2022). Performance Analysis of Speech Recognition Model based on Neuromorphic Architecture of Speech Data Preprocessing Technique. The Journal of The Institute of Internet, Broadcasting and Communication, 22(3), 69-74.

[3] Lee, S., and Park, H., Deep-Learning-Based Gender Recognition Using Various Voice Features, Proceedings of Symposium of the Korean Institute of communications and Information Sciences, 2021, pp. 18-19.

[4] Fonseca, A. H., Santana, G. M., Bosque Ortiz, G. M., Bampi, S., & Dietrich, M. O. (2021). Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. Elife, 10, e59161.

[5] Lee, Y., Lim, S., & Kwak, I. Y. (2021). Cnn-based acoustic scene classification system. Electronics, 10(4), 371.

[6] Roy, A., & Satija, U. (2023). RDLINet: A Novel Lightweight Inception Network for Respiratory Disease Classification Using Lung Sounds. IEEE Transactions on Instrumentation and Measurement.

[7] Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., Cai, L. (2018) Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. Proc. Interspeech 2018, 3683-3687

[8] Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017, February). Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 international conference on platform technology and service (PlatCon) (pp. 1-5). IEEE.

[9] Zhang, S., & Li, C. (2022). Research on feature fusion speech emotion recognition technology for smart teaching. Mobile Information Systems, 2022.

[10] Subramanian, R. R., Sireesha, Y., Reddy, Y. S. P. K., Bindamrutha, T., Harika, M., & Sudharsan, R. R. (2021, October). Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (pp. 1-6). IEEE.

[11] Zheng, L., Li, Q., Ban, H., & Liu, S. (2018, June). Speech emotion recognition based on convolution neural network combined with random forest. In 2018 Chinese control and decision conference (CCDC) (pp. 4143-4147). IEEE.

[12] Kapoor, S., & Kumar, T. (2022). Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network. Multimedia Tools and Applications, 81(21), 31107-31128.

[13] Li, H., Zhang, X., & Wang, M. J. (2021, October). Research on Speech Emotion Recognition Based on Deep Neural Network. In 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP) (pp. 795-799). IEEE.

[14] Aly, M., & Alotaibi, N. S. (2022). A novel deep learning model to detect COVID-19 based on wavelet features extracted from Mel-scale spectrogram of patients' cough and breathing sounds. Informatics in Medicine Unlocked, 32, 101049.

[15] Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2018, November). Attention based fully convolutional network for speech emotion recognition. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1771-1775). IEEE.

[16] Sobahi, N., Atila, O., Deniz, E., Sengur, A., & Acharya, U. R. (2022). Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. Biocybernetics and Biomedical Engineering, 42(3), 1066-1080.

[17] Carofilis, A., Alegre, E., Fidalgo, E., & Fernández-Robles, L. (2023). Improvement of accent classification models through Grad-Transfer from Spectrograms and Gradient-weighted Class Activation Mapping. IEEE/ACM

Transactions on Audio, Speech, and Language Processing.

[18] Bicer, H. N., Götz, P., Tuna, C., & Habets, E. A. (2022, September). Explainable Acoustic Scene Classification: Making Decisions Audible. In 2022 International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 1-5). IEEE.

[19] Cesarelli, M., Di Giammarco, M., Iadarola, G., Martinelli, F., Mercaldo, F., & Santone, A. (2022, November). Deep Learning for Heartbeat Phonocardiogram Signals Explainable Classification. In 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 75-78). IEEE.

[20] Li, P., Li, L., Hamdulla, A., & Wang, D. (2022). Reliable visualization for deep speaker recognition. arXiv preprint arXiv:2204.03852.

[21] Henna, S., & Alcaraz, J. M. L. (2022). From Interpretable Filters to Predictions of Convolutional Neural Networks with Explainable Artificial Intelligence. arXiv preprint arXiv:2207.12958.

[22] Lee, J. H., Lee, C. Y., Eom, J. S., Pak, M., Jeong, H. S., & Son, H. Y. (2022). Predictions for three-month postoperative vocal recovery after thyroid surgery from spectrograms with deep neural network. Sensors, 22(17), 6387.

[23] AI-Hub emotion classification dataset, online available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=120&topMenu=100&dataSetSn=259&aihubDataSe=extrldata.