❓ Domestic Paper

상태 참고자료

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/7552acfe-281d-47d2-9165-65308a6bc098/CNN_기반_전이학습을_이용한_음성_감정_인식.pdf

- 레이블이 다른 일반 소리 데이터(DCASE2018)를 이용한 전이학습 적용
  - feature extractor
    - CNN 구조의 마지막 층인 분류층 이전 층들을 소스 데이터로 학습
      - 학습된 층들은 특징 추출기가 됨.
    - 분류층은 타겟 데이터로 학습하여 타겟 레이블에 대한 분류 작업
  - fine tuning
    - 분류층 이전 층들을 소스 데이터로 학습시킨 후, 타겟 데이터로 분류층을 포함한 모든 층 학습
- 사용된 데이터

표 1. 음성 감정 데이터

Table 1. Speech emotion data

Database	Language	Number of data	Labels
IEMOCAP	English	5,531	Neutral, anger, happiness, sadness
EMO-DB	German	466	Neutral, anger, happiness, sadness, disgust, boredom
SAVEE	English	480	Neutral, angry, happy, sad

- CNN 네트워크

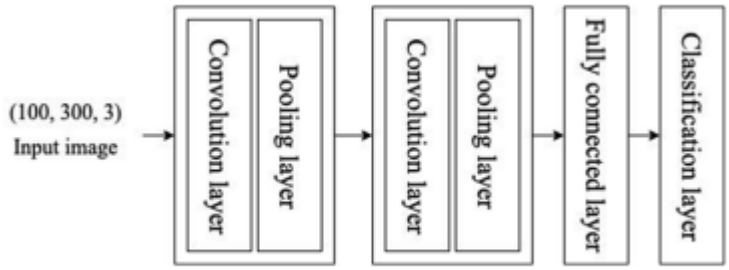


그림 3. 합성곱 신경망 구조

- ACC

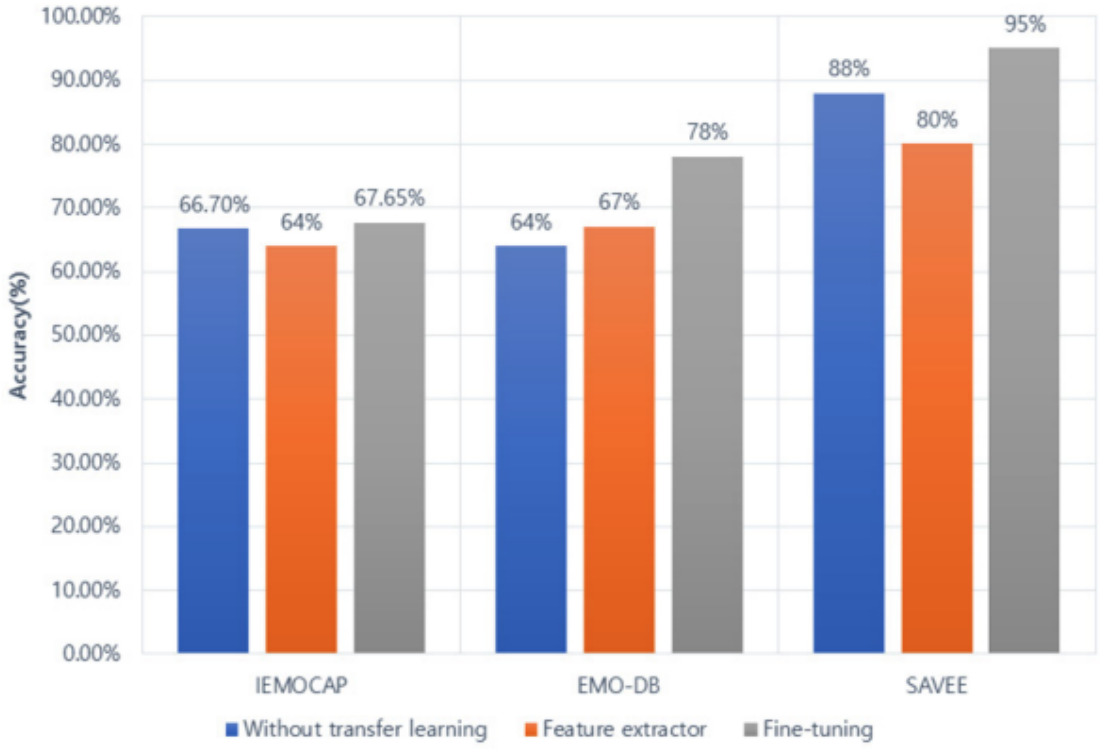


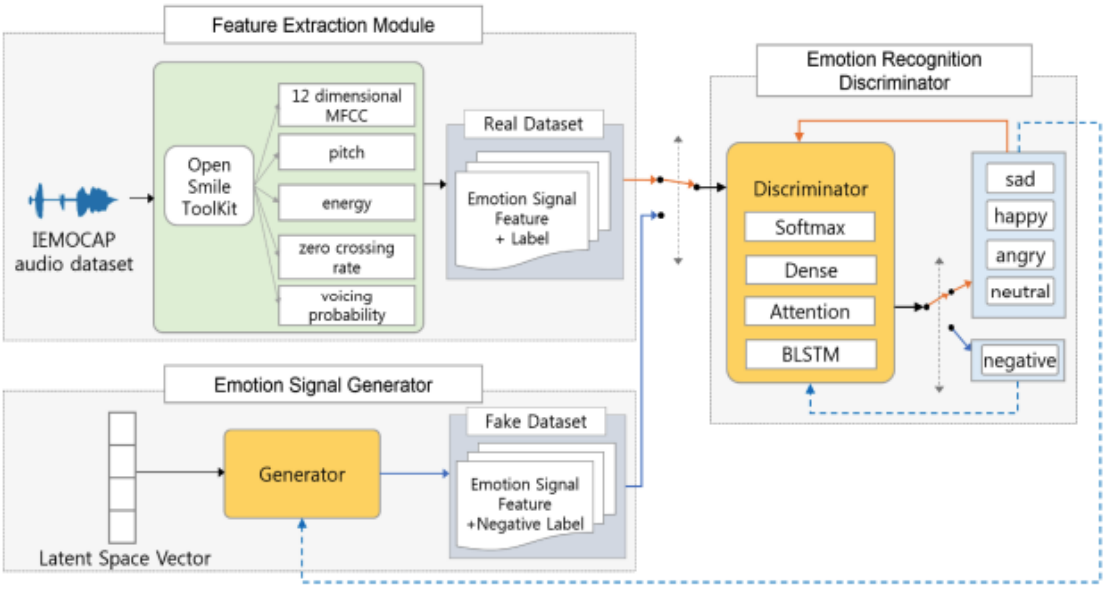
그림 4. 정확도 결과

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/da3346fc-527f-4807-8acd-73399740b943/DNN_을_이용한_End-to-End_한국어_음성_감정_인식.pdf

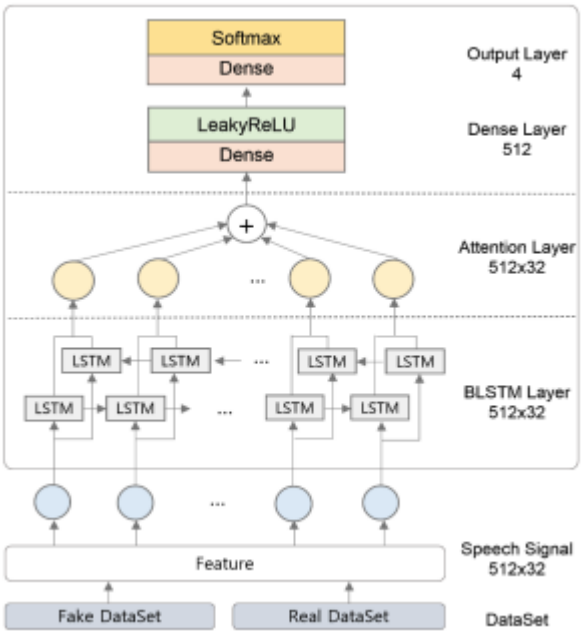
- CNN, CNN+Bi-LSTM, Bi-LSTM+Attetion, CNN+Bi-LSTM +Ateention

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/039d583b-75b5-4f4d-93b9-7d43406bcf03/GAN을_이용한_음성_감정_인식_모델의_성능_개선.pdf

- GAN을 이용해 부정 데이터를 생성하여 추가 학습
  - 특징 추출 모듈로 IEMOCAP 음성 데이터의 실제 데이터셋 생성
  - G를 이용하여 부정 감정 신호 생성 후 부정 레이블을 추가하여 가짜 데이터셋 생성
    - 100의 1차원 latent vector를 입력받아 512x32 부정 감정 신호의 특징 벡터열 생성
    - 부정 라벨 추가
    - Dense(256)-Dense(512)-Dense(1024)-Dense(512)-FC(32)



  - D를 이용한 실제, 가짜 특징 벡터 열에 대한 확률 값 계산



https://s3-us-west-2.amazonaws.com/secure.notion-static.com/ba44d40b-2272-44fc-a2da-24d64ee6dd94/KCI_FI002460260.pdf

- 머신 러닝 분류기의 혼합 모델을 통해 음성 데이터의 7가지 감정(분노, 슬픔, 평온, 행복, 지루함, 공포, 역겨움)을 판별
- MFCC 특징을 SVM, Random Forest, XGBoost 분류기에 나타난 감정 확률에 가중치 융합
  - 해당 음성이 갖는 감정에 대한 확률 값 추출
- 자체 감정 판별 확신도 사용

$$emotion_k(i) = p_k(i_{th} emotion)$$

수식 1. 감정 확률

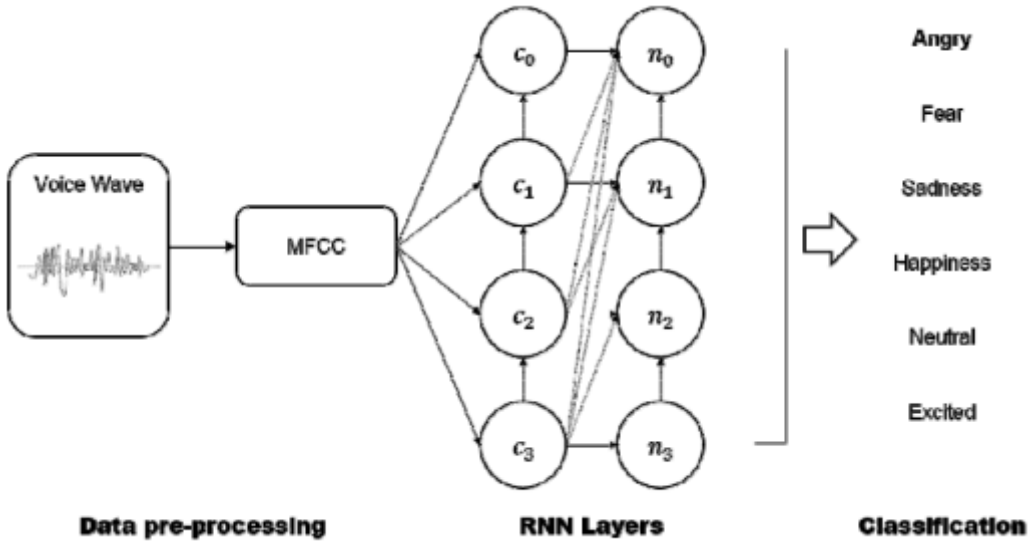
$$weight_k(i) = \frac{emotion_k(i)}{\frac{\sum_{i=0}^{N} emotion_k(i)}{N}}$$

수식 2. 감정에 따른 가중치

$$confidence(i) = \sum_{k=0}^{Z} (weight_k(i) * emotion_k(i))$$

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/f3ccc832-9c3a-4783-b687-598d5ab579d5/RNN_기반_음성_감정인식_기계학습_알고리즘.pdf

- 6가지 감정(화남, 기쁨, 흥분, 중립, 슬픔, 공포) 분류
- 음성 데이터를 MFCC로 변환하고 16000Hz로 샘플링
- 단순 RNN 사용



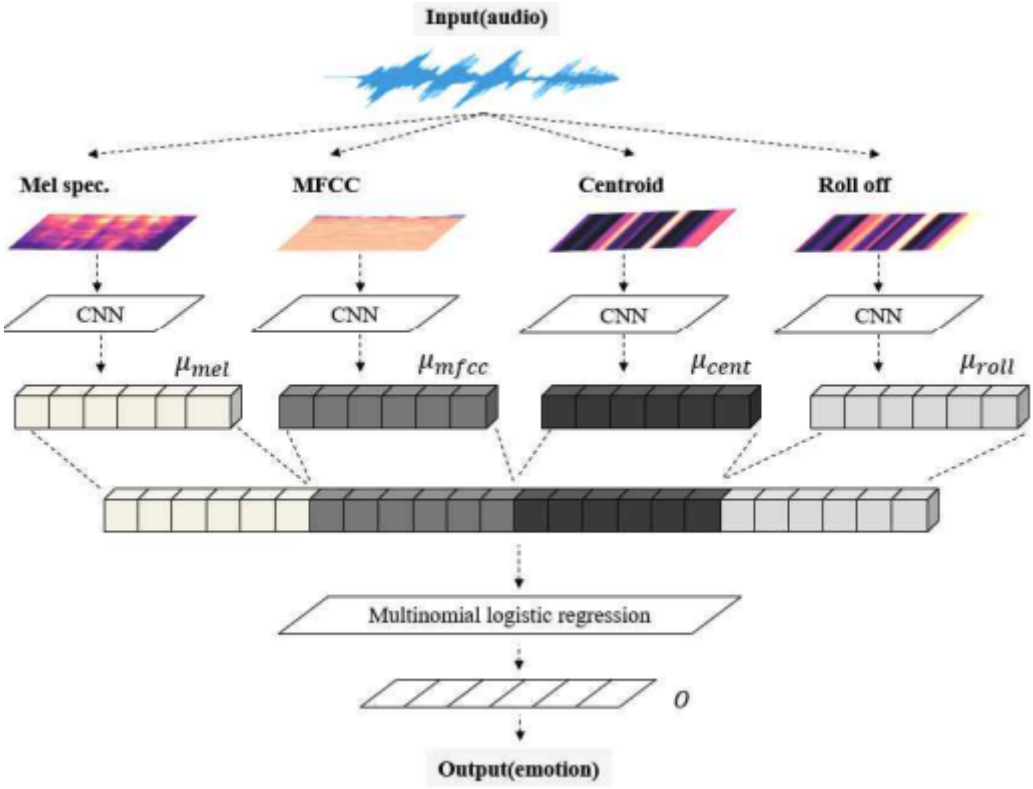
https://s3-us-west-2.amazonaws.com/secure.notion-static.com/942095b8-995a-47db-b54b-f434187db90e/Wav2vec_특징_기반의_한국어_음성감정인식.pdf

- Self-supervised Learning 특징과 기존의 음향적 특징 비교를 위해 wav2vec, IS10 활용
  - wav2vec
    - CNN으로 구성된 encoder, context 네트워크로 구성
    - low-level feature를 사용하지 않고, raw wav를 25ms로 하나의 프레임으로 설계
      - 이를 encoder 네트워크의 입력 특징으로 사용하여 특징 벡터 추출
    - 복수의 특징 벡터 receptive field를 context 네트워크의 입력으로 사용

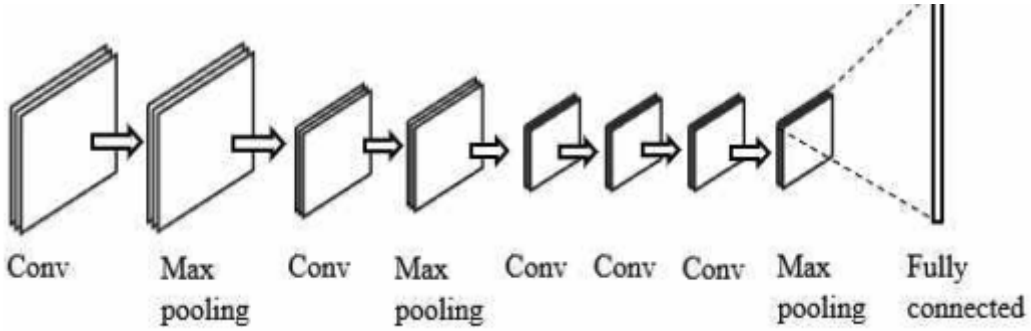
        - contrastive loss의 최소화로 모델 학습
        - Librispeech-960h로 학습된 wav2vec 모델의 context 네트워크의 출력을 mean-pooling한 512차원을 특징으로 사용
    - IS10
        - zero-crossing rate, energy, pitch, MFCC 등 low-level featrue를 프레임 단위로 추출하고 high statistical function을 사용하여 1582차원으로 변환해내는 음향 특징 추출 기법

> https://s3-us-west-2.amazonaws.com/secure.notion-static.com/288bb3fd-d916-4f25-a4d0-6319f83ef6e7/다양한_음성_특징값을_이용한_CNN_기반의_감정_인식_모델.pdf

- 음성으로부터 4가지 특징을 512x400으로 추출하여 입력으로 사용
    - 각 특징은 서로 다른 정보 포함
    - 하나의 CNN이 하나의 특징을 학습
    - 소프트맥스 제거
- 출력값을 concat해서 Multinomial Logistic Regression 적용



- 모델에 사용된 CNN 구조



https://s3-us-west-2.amazonaws.com/secure.notion-static.com/4f0f7e54-6a91-4570-8fe5-0848a2bb5b6d/전이_학습과_어텐션(Attention)을_적용한_합성곱_신경망_기반의_음성_감정_인식_모델.pdf

- 전처리
  - Mel-Spectogram, MFCC의 그래프를 224x224 크기의 이미지로 추출
  - 윈도우 슬라이딩이 아닌 전체 시간에 대한 한번의 그래프 이미지 추출
- 전이 학습에 사용된 CNN

- ImageNet으로 사전 학습된 VGG-19 모델 사용
- block 4, 5까지 미세 조정
  - 1, 2, 3은 가중치 고정
- 분류기를 제거하고 다른 분류기 구성

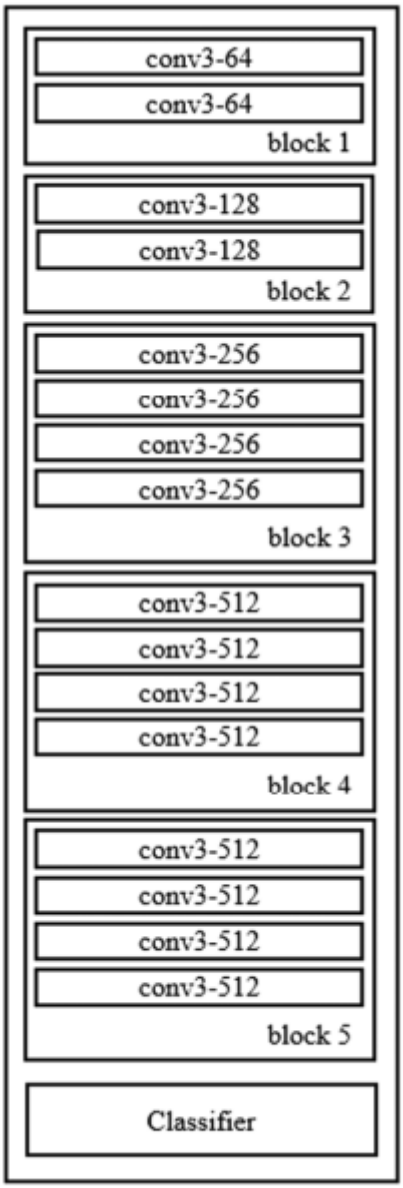
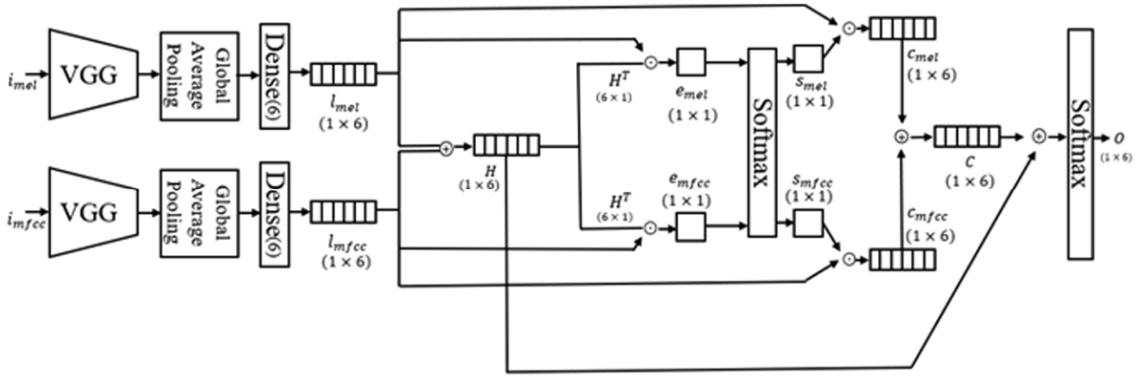


그림 9 VGG-19[15]의 구조

- Attention
  - 앙상블 기법의 다변량 선형 회귀(Multivariate Linear Regression)가 아닌 어텐션 사용
    - 파라미터 수 감소, 데이터의 중요한 부분 강조

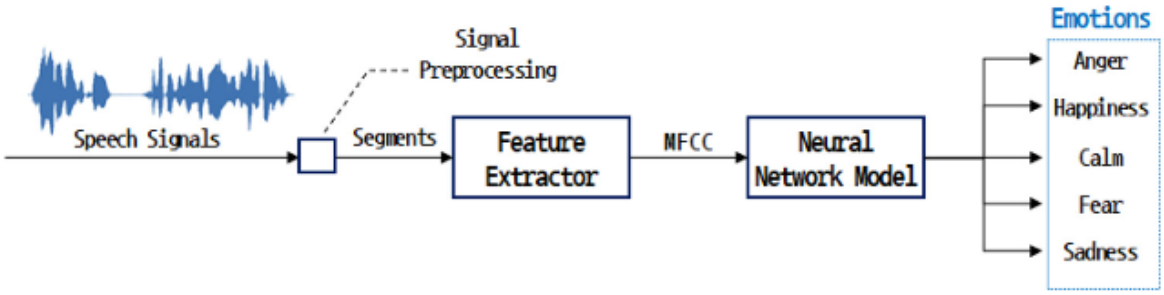
- 모델 구성
  - Mel-Spectogram, MFCC를 VGG19의 입력으로 사용
    - 7x7x512의 크기 벡터 추출
  - GAP을 적용하여 하나의 pooling 값을 구해 1x512 크기의 벡터 추출
  - Dense를 적용하여 1x6 크기의 벡터 H 추출
    - 이는 mel, mfcc의 정보를 모두 가진 일종의 잠재 벡터
  - mel, mfcc의 유사도 추출
  - 유사도에 소프트맥스를 적용하여 attention score 추출



https://s3-us-west-2.amazonaws.com/secure.notion-static.com/f1d6b404-7096-4d58-94bd-b367e68697de/특징_선택과_융합_방법을_이용한_음성_감정_인식.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/6240dbd3-8d6e-435b-8518-eca77734595b/히스토그램_등화와_데이터_증강_기법을_이용한_개선된_음성_감정_인식.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/95860f10-facc-4994-9fa4-c2e3fa132fe1/Speech_Emotion_Recognition_Using_2D-CNN_with_Mel-Frequency_Cepstrum_Coefficients.pdf



- MFCC
  - 44100 샘플링
  - number of MFCC = 50
  - window length = 2048 samples
  - hop length = 512
  - windows = hann
- 사용된 모델 및 정확도
  - 모델 구성은 B장에 나옴.

Model	Accuracy (%)
FCN	72.22
LSTM	70.14
Bi-LSTM	63.54
1D CNN	87.47
1D CNN + LSTM	81.94
1D CNN + Bi-LSTM	86.46
2D CNN	**88.54**
2D CNN + LSTM	84.03
2D CNN + Bi-LSTM	78.47