Ordinal Learning for Emotion Recognition in Customer Service Calls

Wenjing Han, Tao Jiang, Yan Li, Bjorn Schuller, Huabin Ruan

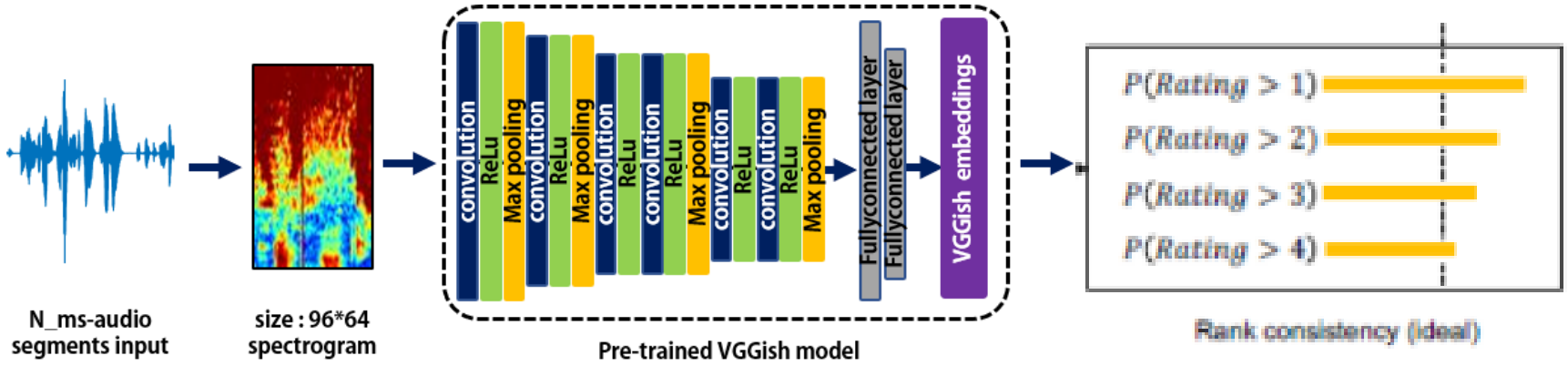
ICASSP 2020

2022. 12. 20

Introduce

- Call center 직원은 제품에 대한 전문적인 설명을 해야 함.
- 간혹 고객의 negative emotion 을 잘 감소시켜야 함.
- 더 좋은 서비스를 위해 emotion recognition이 필요함.
- 이전연구에서는 emotion의 순서적인 특징을 활용하지 않음.
  - negative or positive
- 본 연구에서는 emotion 의 순서적인 특징을 활용하여 음성 분류 모델에 CORAL 을 적용하여 성능을 향상시킴.

Introduce

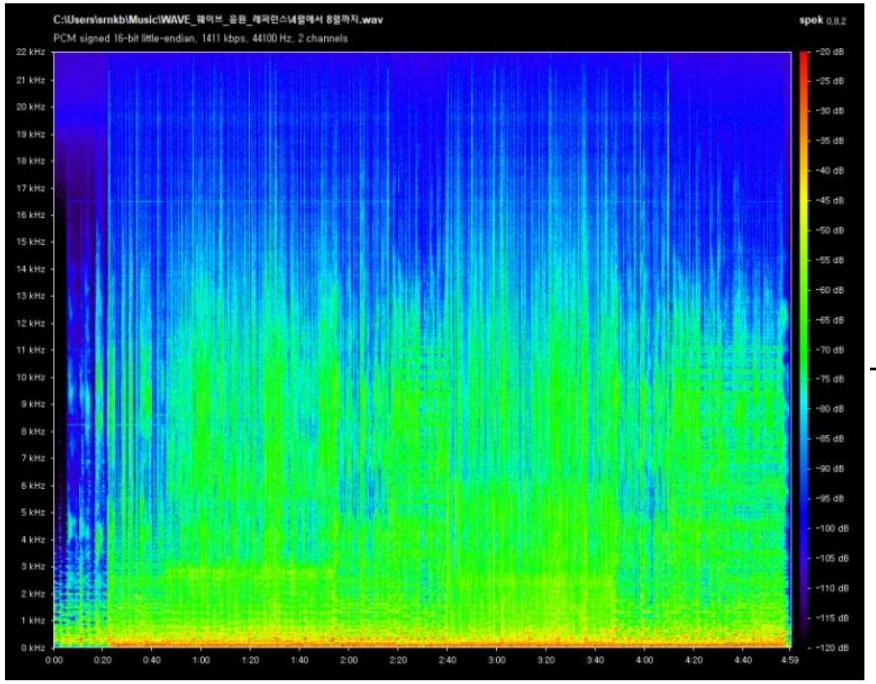


VGGish

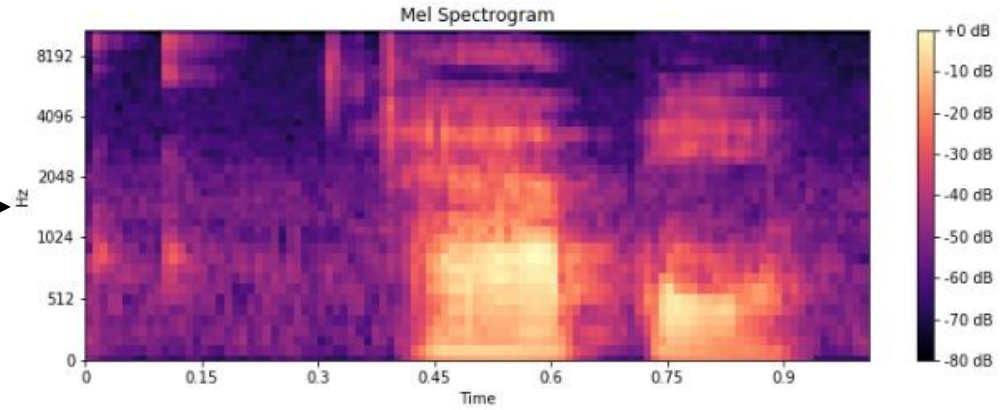
CORAL

Embedding

wav 파일로 기록된 음성을 Mel-spectrogram으로 변환



wave



speech

wav: 시간에 따른 소리의 크기와 높이를 기록한 파일

Mel-spectrogram

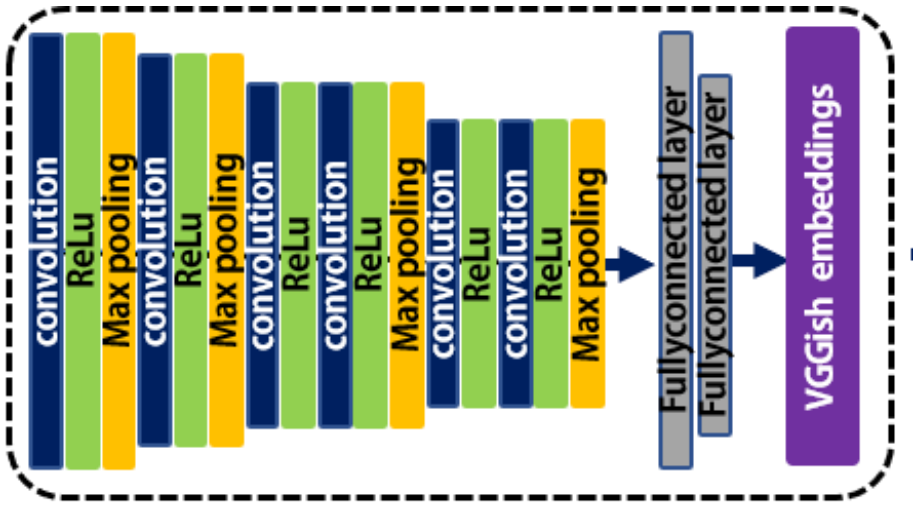
Mel-spectrogram: 고주파대역은 여리게, 저주파대역은 강하게 듣는 인간의 달팽이관의 특성에 따라 음성 feature를 추출한 것.

[DSBA] Mel-frequency cepstrum(MFCC) : https://www.youtube.com/watch?v=jYgMMy8KAc0&t=3287s

VGGish Model

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					



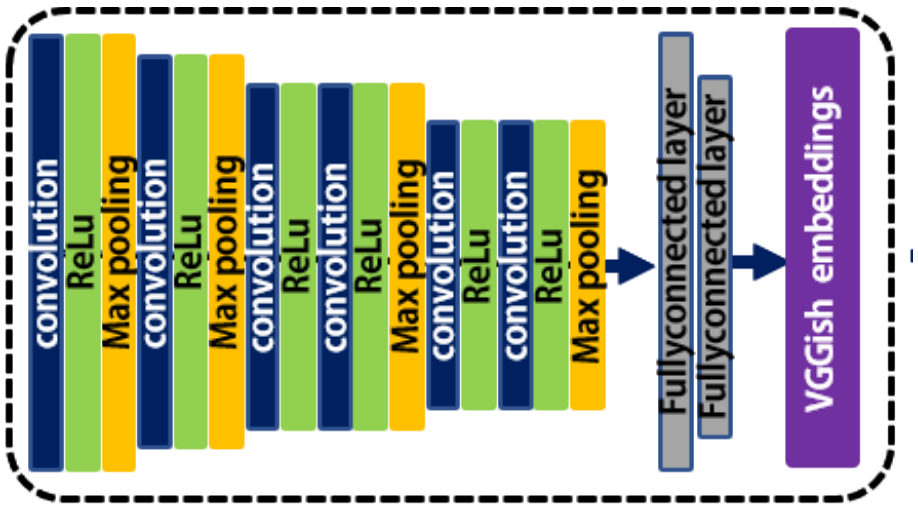
Pre-trained VGGish model

VGG + train audio data

VGG: Very Deep Convolutional Networks for Large-Scale Image Recognition

VGGish: CNN ARCHITECTURES FOR LARGE-SCALE AUDIO CLASSIFICATION

VGGish Model



Pre-trained VGGish model

VGG + train audio data

DATASET

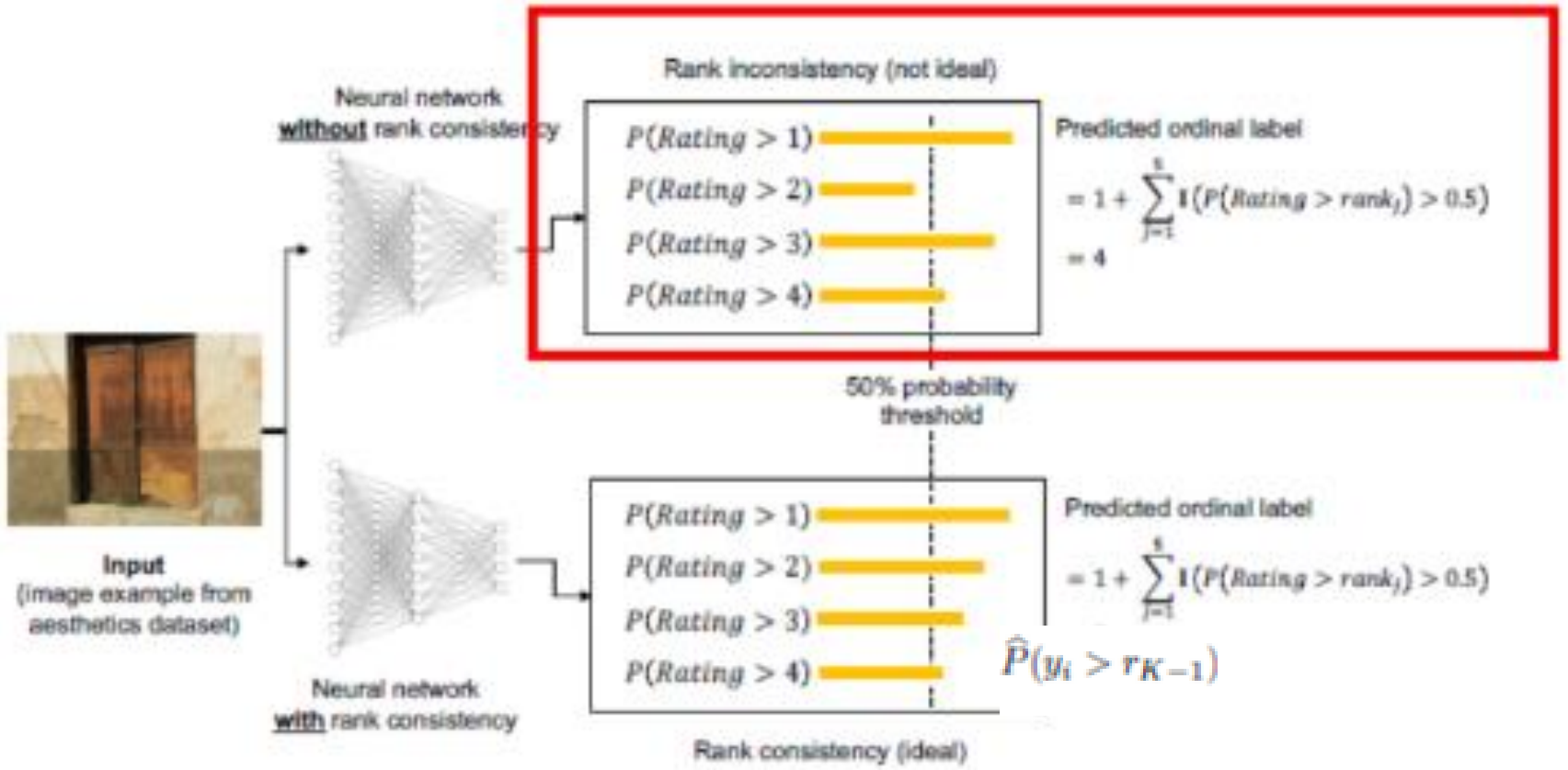
- YouTube-100M dataset (1억개의 비디오)
  - train: 70M (7천만)
  - validation: 10M(1천만)
  - test: 20M(2천만)
- 평균 4.6분의 비디오
- 30,871개의 label

Table 1: Example labels from the 30K set.

Label prior	Example Labels
0.1 . . . 0.2	Song, Music, Game, Sports, Performance
0.01 . . . 0.1	Singing, Car, Chordophone, Speech
$\sim 10^{-5}$	Custom Motorcycle, Retaining Wall
$\sim 10^{-6}$	Cormorant, Lecturer

- 각 비디오를 960ms(0.96초)로 segmentation
- train video -> 200억 개의 example이 나옴.

CORAL


Rank inconsistency (not ideal)
Neural network
without rank consistency
$P(Rating > 1)$
$P(Rating > 2)$
$P(Rating > 3)$
$P(Rating > 4)$
Predicted ordinal label
$= 1 + \sum_{j=1}^{5} \mathbb{1}(P(Rating > rank_j) > 0.5)$
$= 4$
50% probability threshold
Input
(image example from aesthetics dataset)
$P(Rating > 1)$
$P(Rating > 2)$
$P(Rating > 3)$
$P(Rating > 4)$
Predicted ordinal label
$= 1 + \sum_{j=1}^{5} \mathbb{1}(P(Rating > rank_j) > 0.5)$
$\hat{P}(y_i > r_{K-1})$
Neural network
with rank consistency
Rank consistency (ideal)

CORAL

Experiment - Dataset

Call center dataset

- labeled by 10 annotators on a 3-point scale:

  1: Non-negative, 2: Somewhat-negative, and 3: Obviously-negative

- labeling 절차

(i) Annotator 대다수가 **Non-speech** 또는 **이해할 수 없는 발언**으로 표시한 발언을 폐기한다.

(ii) 발언 집합에 대해서는 평균 annotation과의 **상관관계가 0.6 미만인 주석자의 주석을 삭제**한다.

(iii) 나머지 발언 각각에 대해 잔류의 **평균 및 표준 편차를 계산**한다. 주석이 평균에서 하나 이상의 표준 편차인 경우 이 주석은 삭제됩니다.

(iv) 다수결이 없는 발언도 삭제합니다. (annotator들의 의견이 분분한 데이터 삭제)

- 129개의 대화 -> 각 발화로 나누어 5,270개의 예제 얻음
- 5,270개의 예제 -> label 절차 진행 후 4,537개로 감소함

Experiment - Dataset

Call center dataset

- labeled by 10 annotators on a 3-point scale:

  1: Non-negative, 2: Somewhat-negative, and 3: Obviously-negative
- labeling 절차
- 129개의 대화 -> 각 발화로 나누어 5,270개의 예제 얻음
- 5,270개의 예제 -> label 절차 진행 후 4,537개로 감소함
- dataset distribution

Label	Count	Duration [h]	κ
Non-negative	2,317	1.5	0.93
Somewhat Negative	1,701	1.1	0.68
Obviously Negative	519	0.4	0.75
In Total	4,537	3.0	0.79

Experiment - Dataset

Call center dataset (8kHz)

- labeled by 10 annotators on a 3-point scale:

  1: Non-negative, 2: Somewhat-negative, and 3: Obviously-negative

- labeling 절차
- 129개의 대화 -> 각 발화로 나누어 5,270개의 예제 얻음
- 5,270개의 예제 -> label 절차 진행 후 4,537개로 감소함
- dataset distribution

Label	Count	Duration [h]	κ
Non-negative	2,317	1.5	0.93
Somewhat Negative	1,701	1.1	0.68
Obviously Negative	519	0.4	0.75
In Total	4,537	3.0	0.79

Fleiss's Kappa(κ): categorical 데이터를 3명 이상의 annotator가 labeling을 수행했을 때 신뢰도(일관성) 분석

<Fleiss's>

<0.4 poor

0.40-0.75 fair

<0.75 excellent

Experiment - Dataset

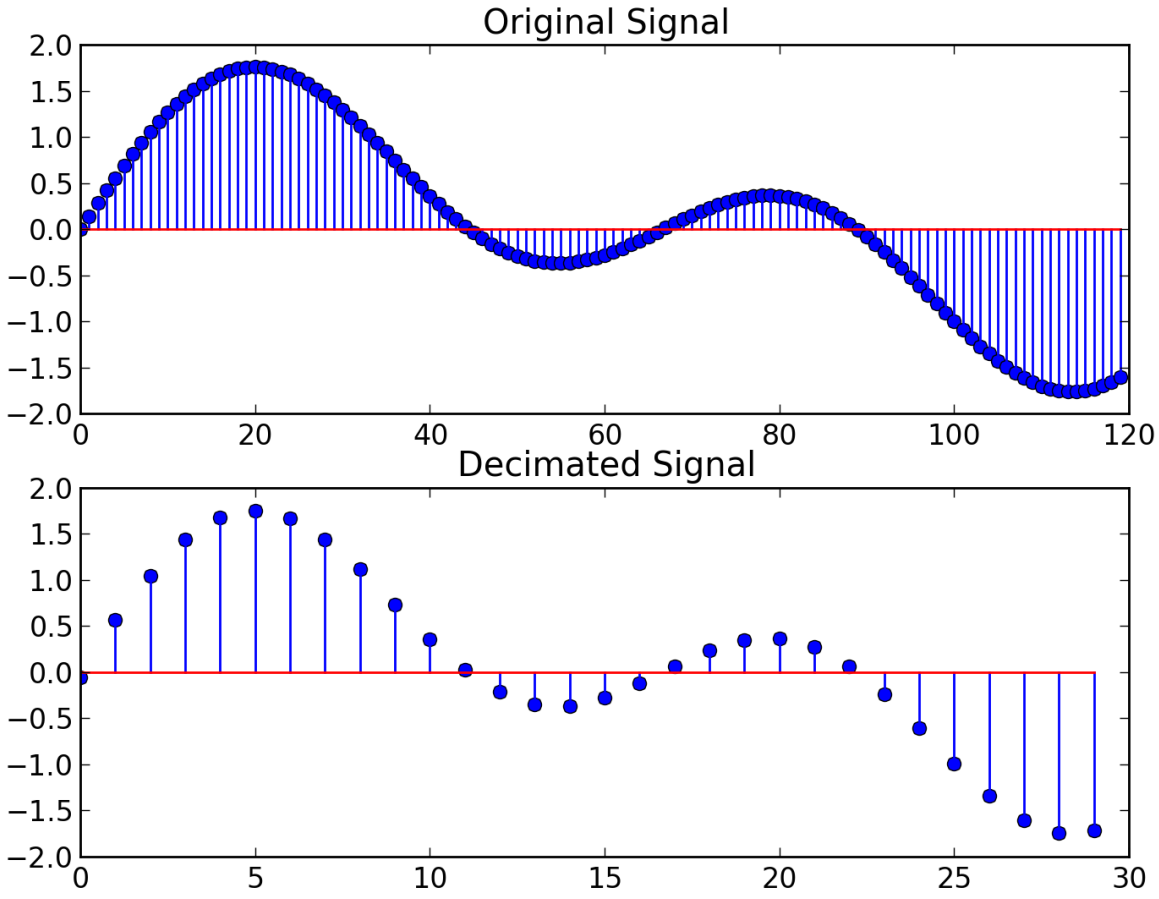
IEMOCAP

- multimodal dataset
- 2명의 화자가 대화한 151개의 비디오로 구성됨.
  - 각 화자로 나누어 302개의 비디오로 분리되어 있음.
- 9개의 감정이 레이블링되어 있음. (happy, angry, ...)
- 여기서 Angry, Excited, Happy, Neutral, Sad 5가지 감정만 사용
- IEMOCAP 데이터셋은 call center dataset 과 맞추기 위해서 downsampled(8kHz)함.

Experiment - Dataset

IEMOCAP

- multimodal dataset
- 2명의 화자가 대화한 151개의 비디
  - 각 화자로 나누어 302개의 비
- 9개의 감정이 레이블링되어 있음.
- 여기서 Angry, Excited, Happy, Ne
- IEMOCAP 데이터셋은 call center

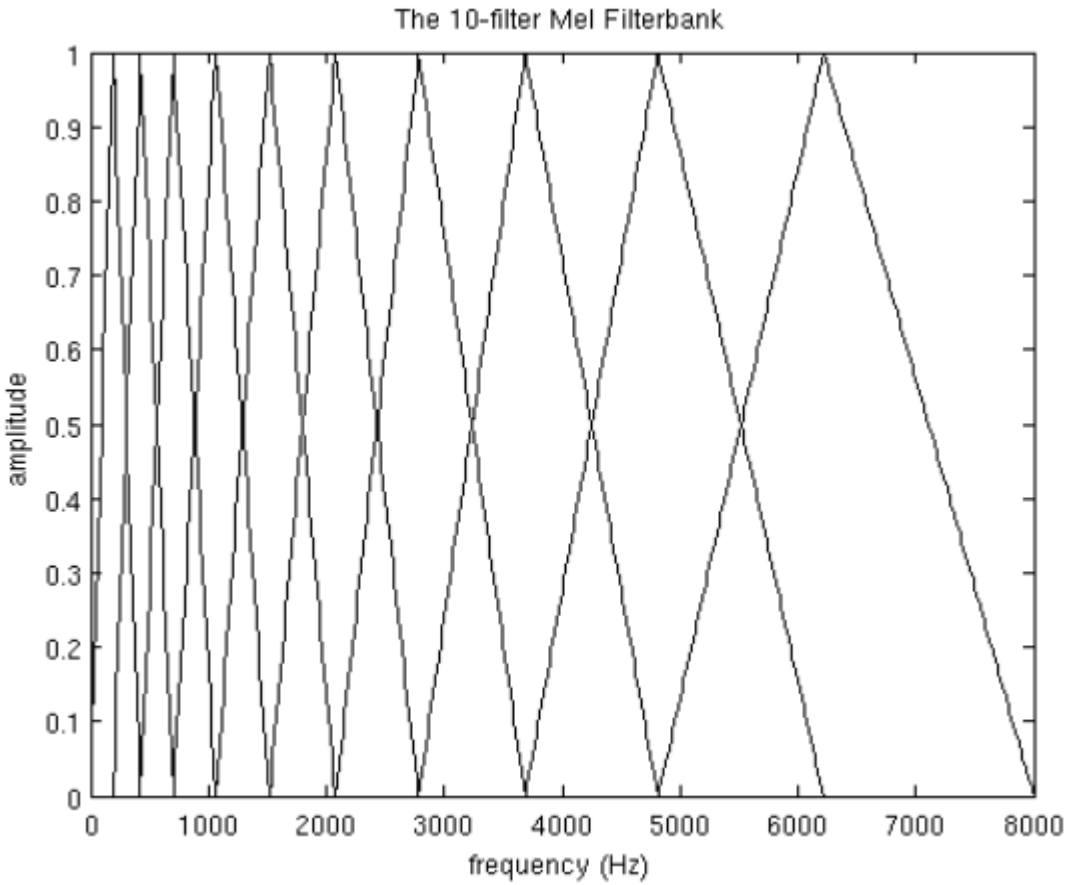


Experiment - Setup detatils

- ❖ Input
  - ➢ Mel-spectrogram: using Librosa. https://kaen2891.tistory.com/39
    - window size: 256 frame length, 한번에 볼 크기(256ms)
    - hop size: 128 음성의 magnitud를 얼만큼 겹친 상태로 잘라서 칸으로 보여줄 것인가?
    - Mel bands=96
    - 8kHz sampling rate of raw signal
    - input vector shape: 96*501 (for each utterance)
- ❖ Model
  - ➢ VGGish, VGGish-CORAL, VGGish-CORAL-TIW(Task Importance Weighting)

Experiment - Setup detatils

- ❖ Input
  - ➢ Mel-spectrogram: using Librosa. https://kaen2891.tistory.com/39
    - window size: 256 fra
    - hop size: 128 음성으
      칸으로 보여줄 것인
    - Mel bands=96
    - 8kHz sampling rate
    - input vector shape:



- ❖ Model
  - ➢ VGGish, VGGish-CORAL, VGGish-CORAL-T

Experiment - Setup detatils

- ❖ Input
  - ➢ Mel-spectrogram: using Librosa. https://kaen2891.tistory.com/39
    - window size: 256 frame length, 한번에 볼 크기(256ms)
    - hop size: 128 음성의 magnitud를 얼만큼 겹친 상태로 잘라서 칸으로 보여줄 것인가?
    - Mel bands=96
    - 8kHz sampling rate of raw signal
    - input vector shape: 96*501 (for each utterance)
- ❖ Model
  - ➢ VGGish, VGGish-CORAL, VGGish-CORAL-TIW(Task Importance Weighting)

Experiment - Setup detatils

- Model
  - VGGish, VGGish-CORAL, VGGish-CORAL-TIW(Task Importance Weighting)

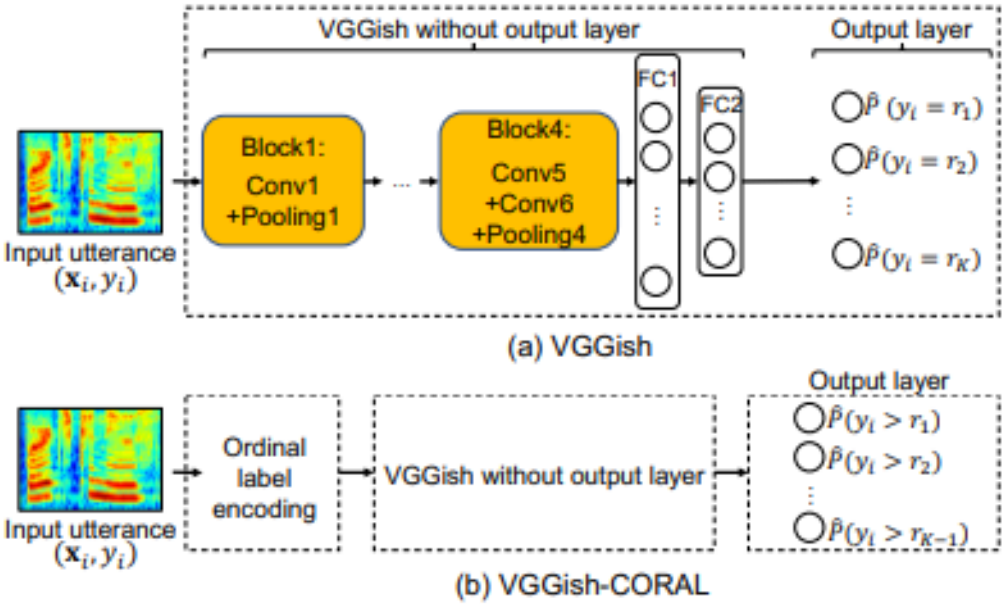


Fig. 1. Framework evolution from the VGGish based classifier (a) to our proposed VGGish-CORAL ranker (b).

$$\lambda^k = \frac{\sqrt{M_k}}{\max_{1 \leq i \leq K-1} \sqrt{M_i}}.$$

$$L = -\sum_{i=1}^{N} \sum_{k=1}^{K-1} \lambda^k [log(s(g(\mathbf{x}_i, W) + b_k)) y_i^k + log(1 - s(g(\mathbf{x}_i, W) + b_k))(1 - y_i^k)],$$

Results and Analysis

Methods	Call Center Dataset			IEMOCAP-Val		
	UAR	RMSE	EER	UAR	RMSE	EER
VGGish	71.4	0.22	23.1	56.5	0.33	33.3
VGGish-CORAL	72.3	0.18	21.8	57.1	0.26	32.2
VGGish-CORAL-TIW	72.6	0.15	21.2	57.3	0.23	31.6

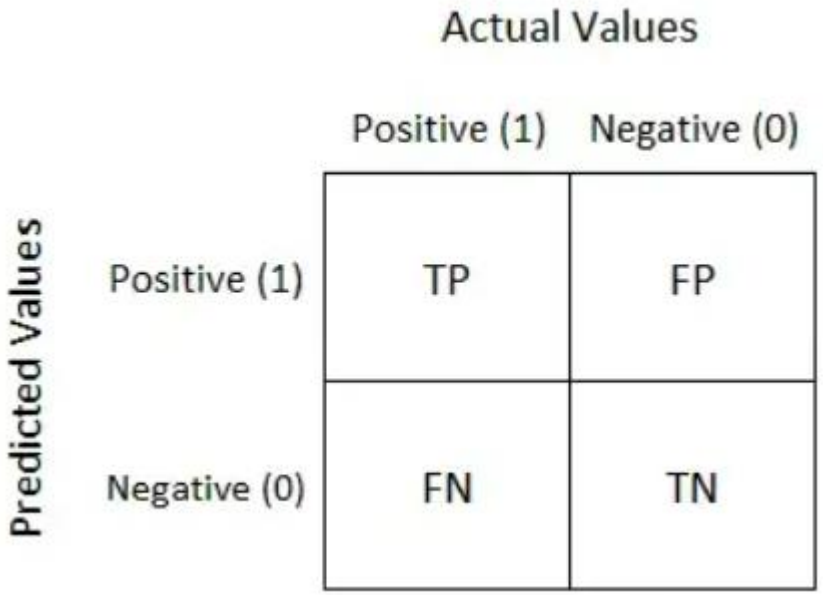
Evaluation Metric

- UAR(Unweighted average recall)[%]: https://ogunlao.github.io/blog/2021/04/24/consider_uar_accuracy.html
- RMSE(Root Mean Squared Error)
- EER(Equal Error Rate)[%]: EER for obviously Negative detection

Results and Analysis

Evaluation Metric

- UAR(Unweighted average recall)[%]



$$accuracy = \frac{total\ correct\ predictions}{total\ number\ of\ predictions}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$accuracy = \frac{tp + tn}{p + n}$$

$$accuracy = \frac{tp}{p + n} + \frac{tn}{p + n}$$

$$accuracy = \frac{tp}{p} . \frac{p}{p + n} + \frac{tn}{n} . \frac{n}{p + n}$$

$$accuracy = Sensitivity . \frac{p}{p + n} + Specificity . \frac{n}{p + n}$$

0.5 or $\frac{1}{no\ of\ classes}$

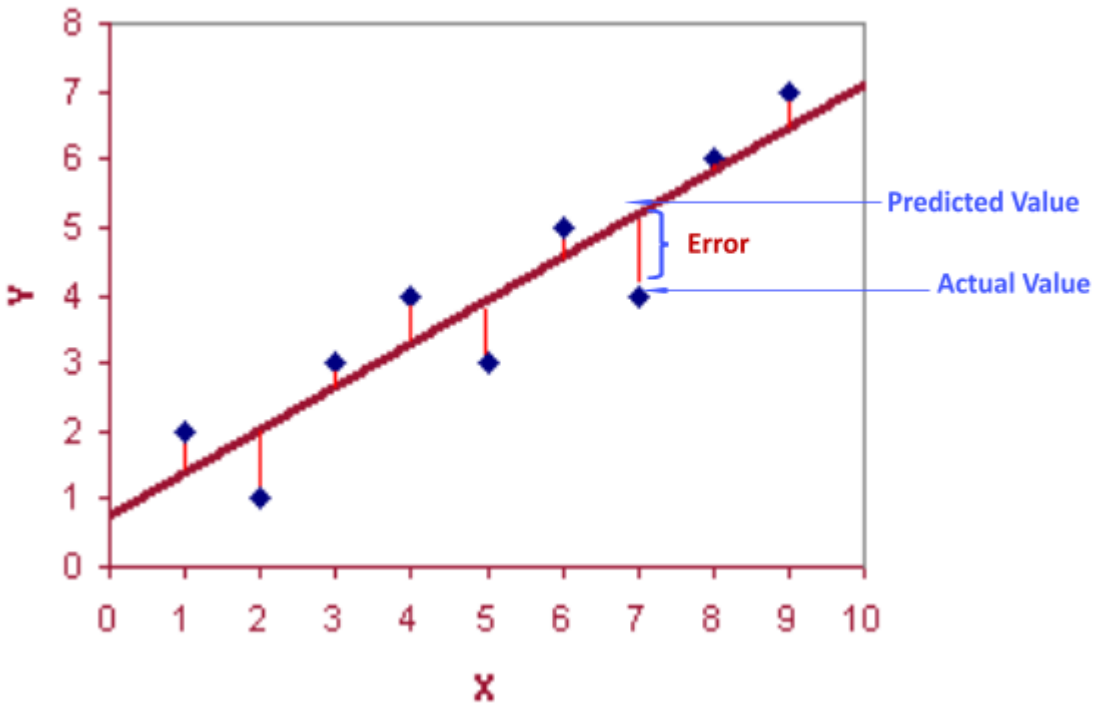
$$accuracy_{bal} = Sensitivity \times 0.5 + Specificity \times 0.5$$

Results and Analysis

Evaluation Metric

- RMSE(Root Mean Squared Error)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$



- 1(부정적이지 않음) -> 3(명백히 부정적)
- 2(약간 부정적) -> 3(명백히 부정적)

=> RMSE값이 낮을수록 순서적인 특성을 잘 학습한 것.

Results and Analysis

Evaluation Metric

- EER(Equal Error Rate)[%]: EER for obviously Negative detection
- 블랙리스트 고객 관리를 위해 추가적으로 조사함.

Conclusion

- 순서형 레이블을 갖는 Call Center Dataset 구축
- 이전 연구에서 사용하지 않았던 레이블의 순서적인 특성 활용
- 순서형 회귀모형 CORAL 을 활용하여 음성감성인식을 했을 때 성능이 향상되는 것을 실험을 통해 확인함
- CORAL 을 사용할 때 Task Importance Weight 를 활용하여 성능 향상을 보임