Global KU
Frontier Spirit

WAV에서 MFCC로의 지식 증류를 이용한 음성감정인식

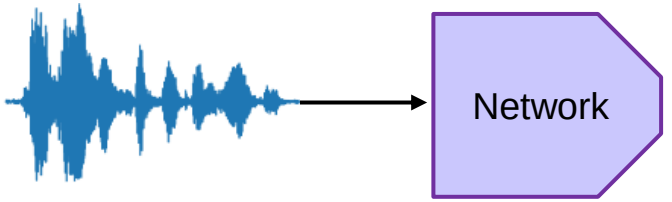
SPEECH EMOTION RECOGNITION USING KNOWLEDGE DISTILLATION WITH WAV TO MFCC

고려대학교 전기전자공학과
홍윤아, 이보경, 구본화, 고한석

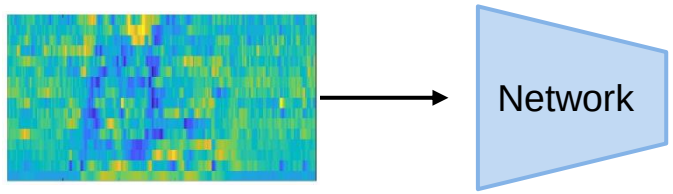
2023.11.02

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Introduction
  - Transformer
    - audio의 정보를 모두 사용할 수 있다. 정확도가 높다.
    - 모델 크기가 크다. Inference시간이 느리다.



  - CNN (MFCC 사용)
    - 모델 크기가 작다. Inference 시간이 빠르다.
    - Audio의 모든 정보를 사용하지 못한다. 정확도가 낮다.



- Knowledge distillation을 이용해서 mfcc의 정보와 함께 raw audio signal에서 추출한 정보를 사용할 수 있도록 하여 가볍고 정확도가 높은 모델을 만들고자 한다.

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Teacher-Student models
  - Teacher model – WAV2VEC 2.0 - base[1)]
  - Student model – TIM-Net[2)]

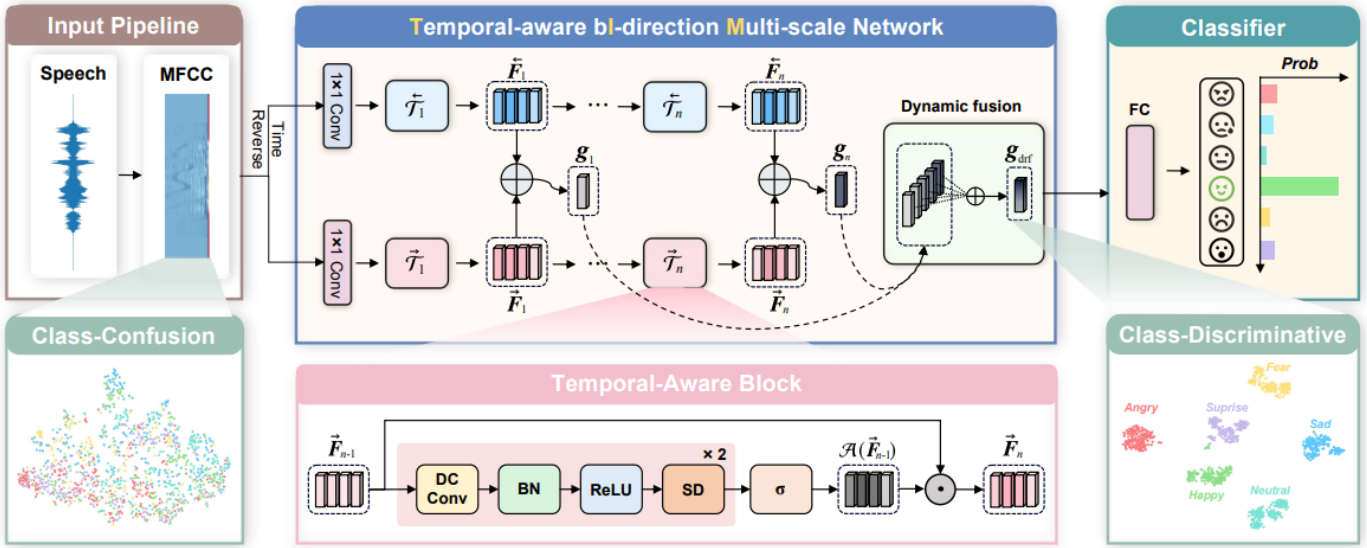


Fig 1. TIM-Net논문의 fig 1 개요도

1) Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.
2) J. Ye, X. -C. Wen, Y. Wei, Y. Xu, K. Liu and H. Shan, "Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition," *ICASSP 2023*

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Proposed knowledge distillation method
  - Combine logit knowledge with relational knowledge

Total Loss

$$\alpha\, L_{CE}\left(\sigma(Z_s), \hat{y}^{LS}\right) + (1-\alpha) L_{KL}\left(\sigma(Z_T/T), \sigma(Z_s/T)\right) + \beta * (L_{RKD-D} + L_{RKD-A})$$

- Logit distillation[3)]
  - KL-divergence loss - Teacher와 Student의 logit을 이용하여 soft label, soft prediction를 사용
  - Cross-entropy loss - Student의 logit을 이용하여 만든 hard prediction과 label smoothing된 soft label을 사용

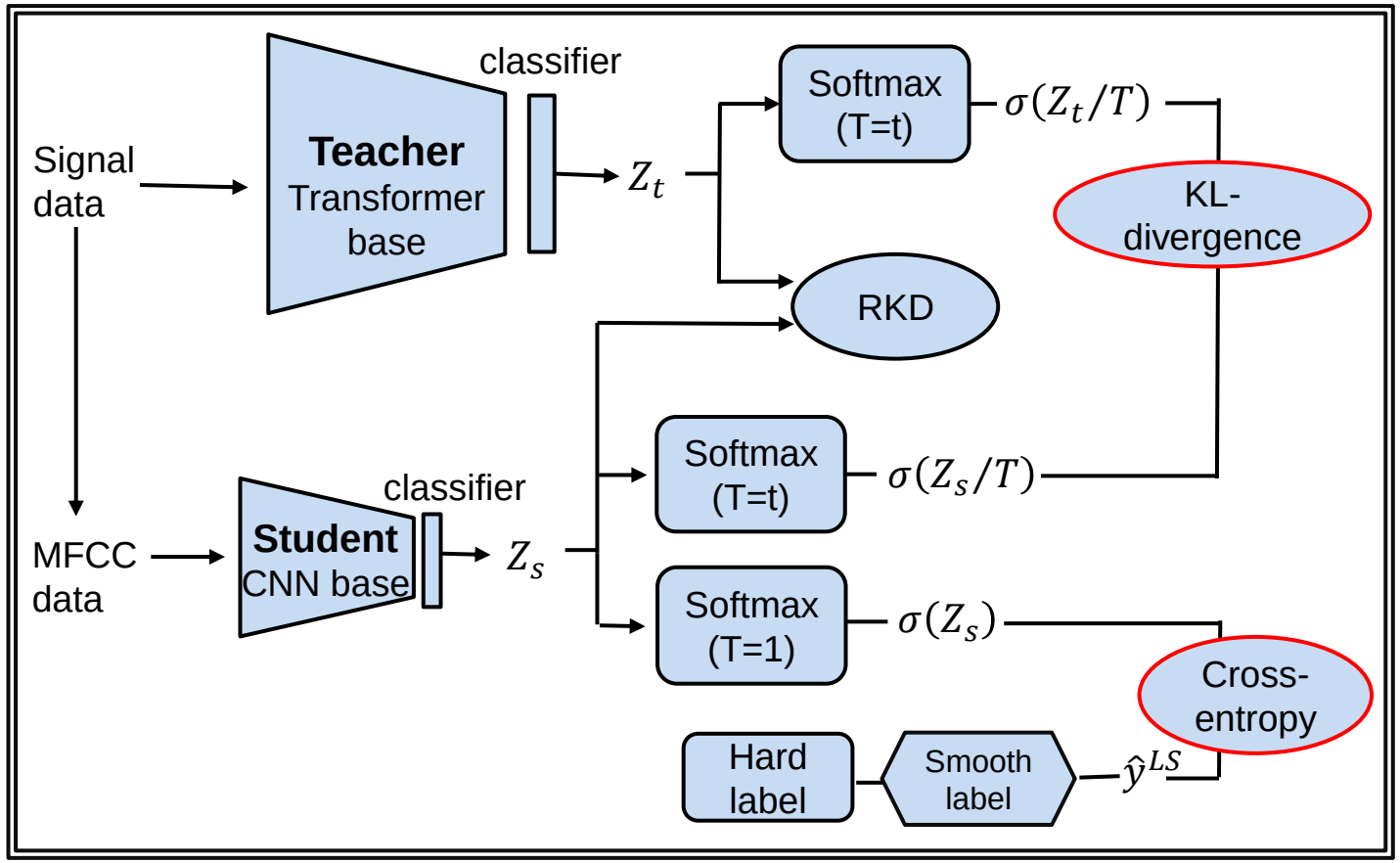


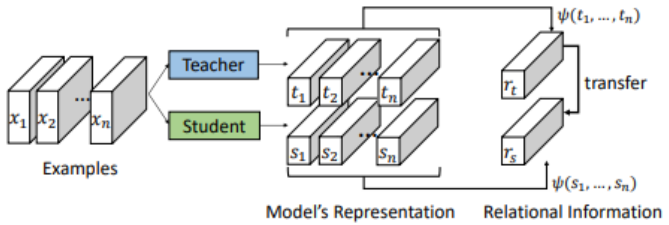
Fig 2. 전체 과정 개요도

3) Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

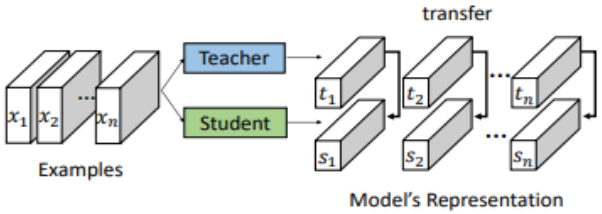
WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Proposed knowledge distillation method

  - Relational knowledge distillation[4)]



Relational knowledge distillation



logit knowledge distillation

Fig 3. Relational knowledge distillation
(Relational Knowledge Distillation논문의 fig 2)

Total Loss

$$\alpha\, L_{CE}(\sigma(Z_s), \hat{y}^{LS}) + (1-\alpha) L_{KL}\big(\sigma(Z_T/T), \sigma(Z_s/T)\big) + \boldsymbol{\beta} * (\boldsymbol{L_{RKD-D}} + \boldsymbol{L_{RKD-A}})$$

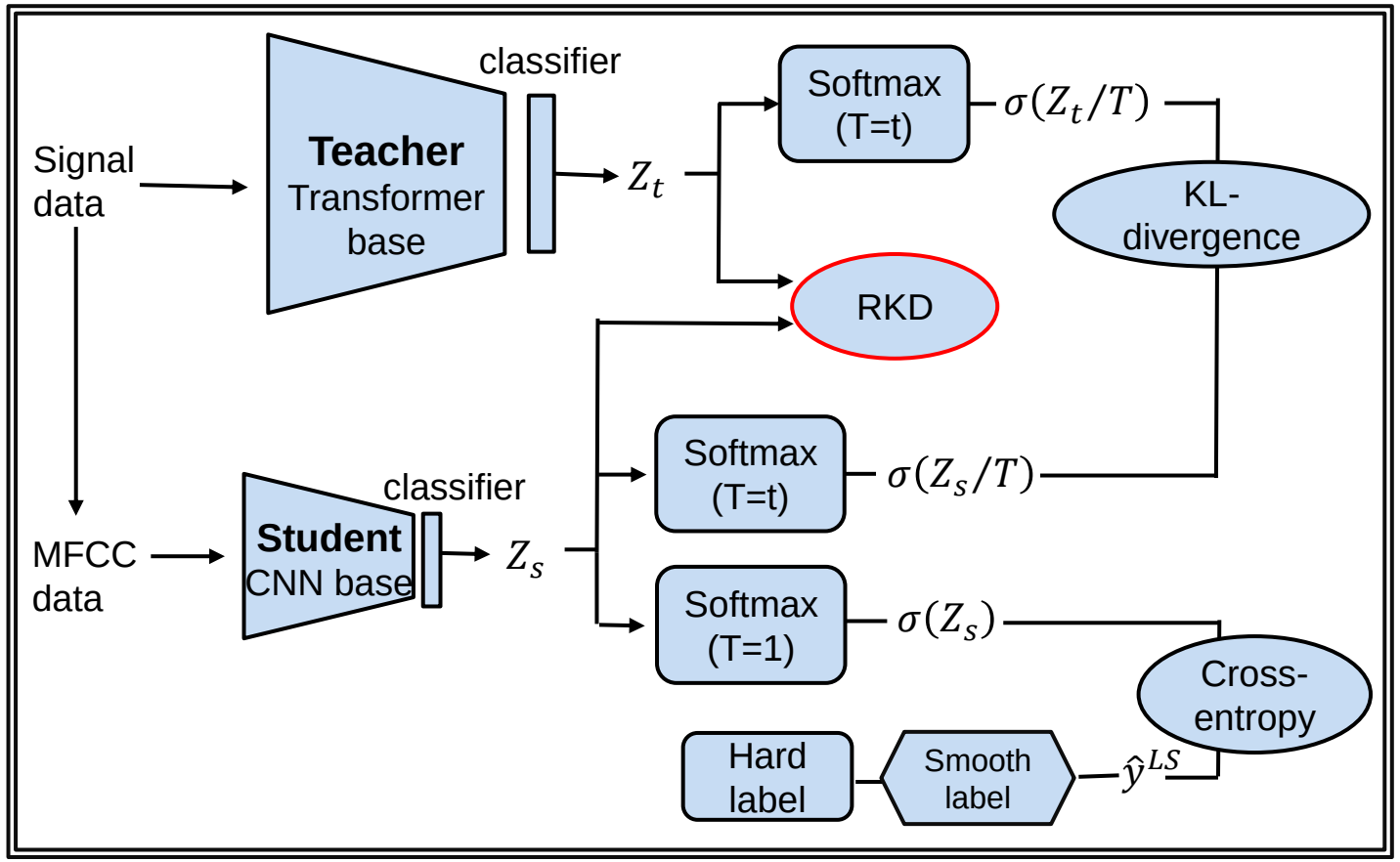


Fig 2. 전체 과정 개요도

4) Park, Wonpyo, et al. "Relational knowledge distillation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Proposed knowledge distillation method

  - Relational knowledge distillation[4)]

$$L_{RKD-D} = \sum_{(x_i,x_j)\in Z_t} l_\delta(\psi_D(t_i,t_j),\psi_D(s_i,s_j))$$

$$L_{RKD-A} = \sum_{(x_i,x_j,x_k)\in Z_t} l_\delta(\psi_A(t_i,t_j,t_k),\psi_A(s_i,s_j,s_k))$$

l_δ: huber loss
ψ_A: dot product
ψ_D: normalized distance

Total Loss

$$\alpha\, L_{CE}(\sigma(Z_s),\hat{y}^{LS}) + (1-\alpha)L_{KL}(\sigma(Z_T/T),\sigma(Z_s/T)) + \boldsymbol{\beta} * (\boldsymbol{L_{RKD-D}} + \boldsymbol{L_{RKD-A}})$$

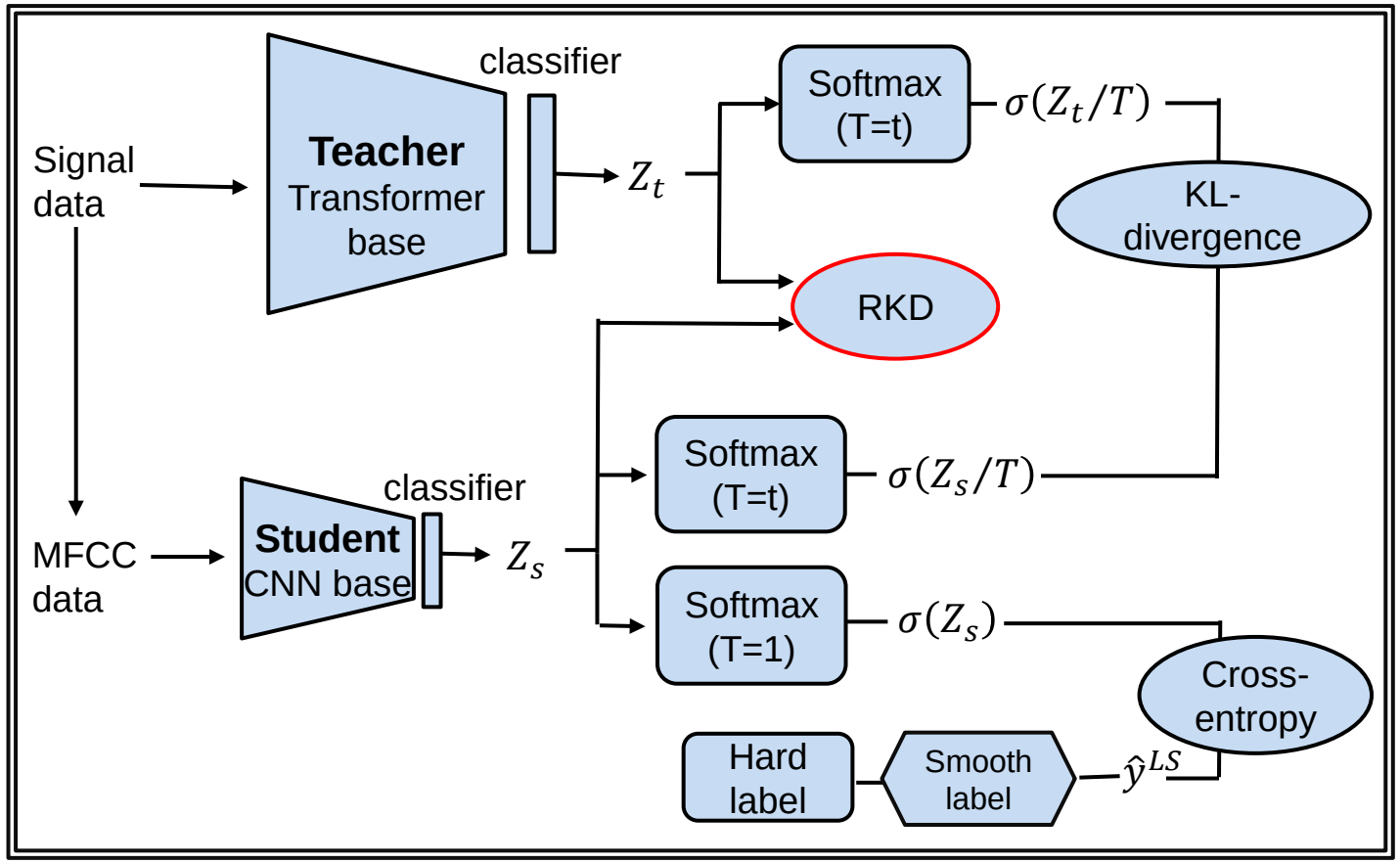


Fig 2. 전체 과정 개요도

4) Park, Wonpyo, et al. "Relational knowledge distillation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Experiments
  - Dataset
    - RAVDESS (5 fold)
    - 8 classes: Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprise
    - 24 actors (남자 12명, 여자 12명) / 1440개 data


Neutral-1


Calm-1


Surprise-2

    - Augmentation : gaussian noise, scaling amplitude, shift, scaling pitch (50%) / padding (84351로 전체 통일)

- Fold 0: (2, 5, 14, 15, 16);
- Fold 1: (3, 6, 7, 13, 18);
- Fold 2: (10, 11, 12, 19, 20);
- Fold 3: (8, 17, 21, 23, 24);
- Fold 4: (1, 4, 9, 22).

Fig 4.Actor기준 fold

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Experiments
  - Performance of teacher model

Wav2Vec2.0 - base

TRAIN	TEST	Fold 0	
		5-CV accuracy	Weighted F1-Score
RAVDESS(7)	RAVDESS(7)	84.00%	83.47%

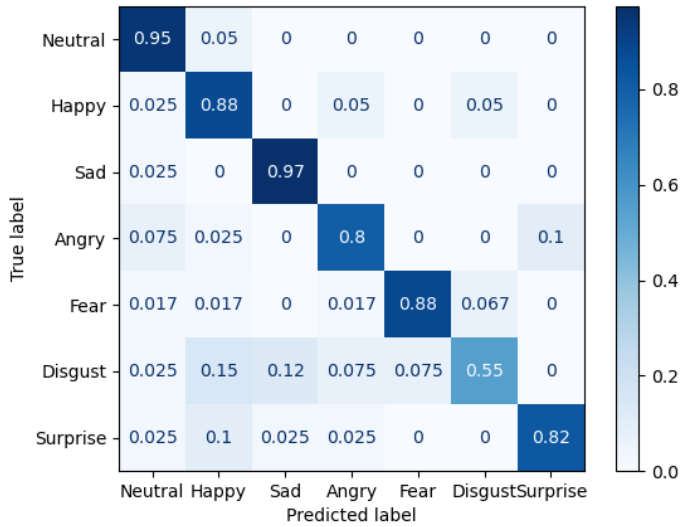


Fig 5. Teacher confusion matrix

-> disgust class에 대한 정확도가 다른 클래스들에 비해서 떨어짐.

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Experiments

SER Model	CPU latency	GPU latency	Acc
Teacher Model (Wav2Vec2.0 - base) + ONNX	0.2507s	0.0239s	84.00%
Student Model (Timnet) + ONNX	0.0018s	0.0016s	78.67%

SER Model	Parameters	CPU latency	GPU latency	FLOPs	Acc
Teacher Model (Wav2Vec2.0 - base)	94,967,687	—	0.0862s	26.8621G	84.00%
Student Model (Timnet)	245,235	0.0444s	0.0342s	0.0883 G	78.67%

Original model

TRAIN	TEST	Fold 0	
		5-CV accuracy	Weighted F1-Score
RAVDESS(7)	RAVDESS(7)	71.67%	70.15%

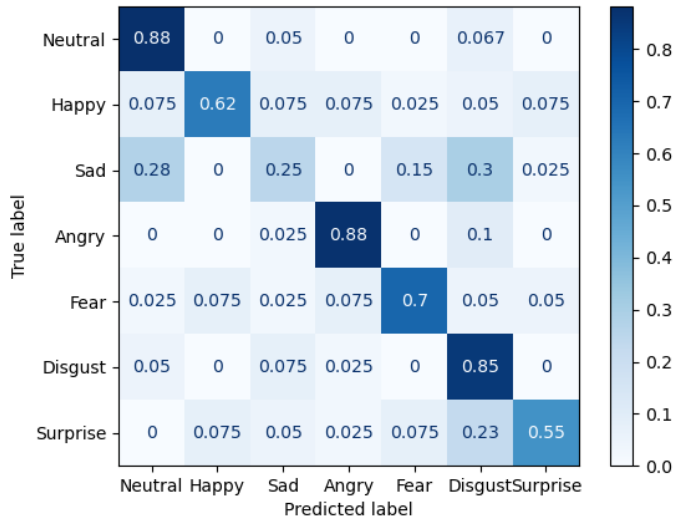


Fig 6. Original TIM-Net confusion matrix

-> sad class가 neutral, fear, disgust로 잘못 분류

Distillation model (CE + KL)

TRAIN	TEST	Fold 0	
		5-CV accuracy	Weighted F1-Score
RAVDESS(7)	RAVDESS(7)	78.67%	80.11%

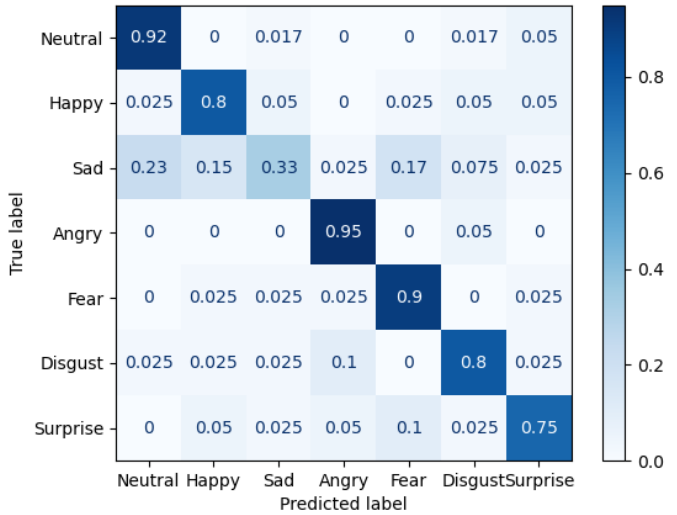


Fig 7. Distillated TIM-Net confusion matrix (CE + KL)

-> distillation을 했을 때 teacher와 다르게 disgust가 잘 예측됨

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Experiments

Distillation model (CE + KL) without soft label

TRAIN	TEST	Fold 0	
		5-CV accuracy	Weighted F1-Score
RAVDESS(7)	RAVDESS(7)	78.00%	78.84%

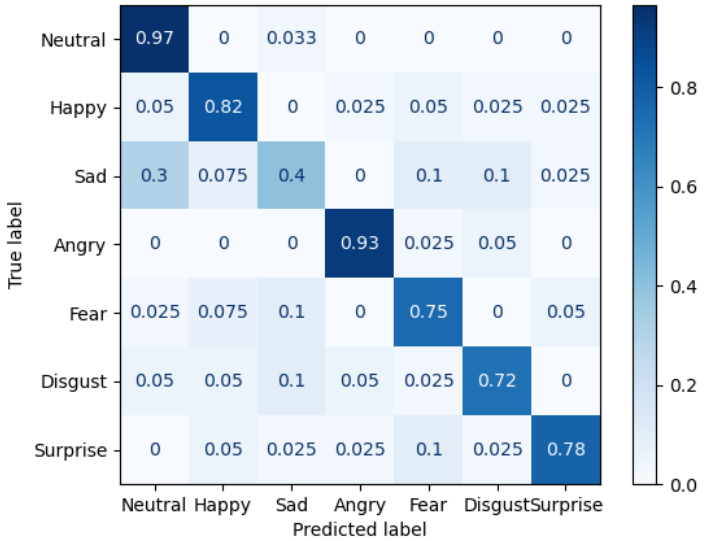


Fig 8. Distillated TIM-Net confusion matrix (CE + KL) w/o soft label

-> class 정확도가 골고루 높지 않음. Neutral과 angry 의 정확도만 확연히 높은 것을 확인할 수 있음

Distillated model (CE + KL + RKD)

TRAIN	TEST	Fold 0	
		5-CV accuracy	Weighted F1-Score
RAVDESS(7)	RAVDESS(7)	81.00%	81.48%

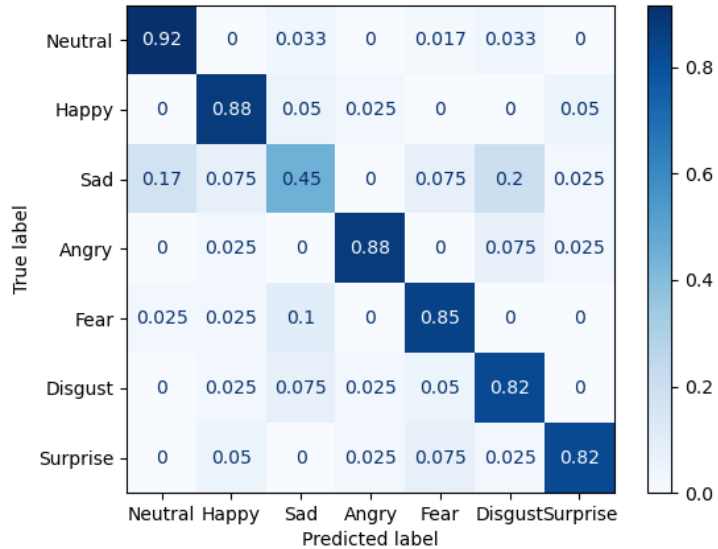


Fig 9. Distillated TIM-Net confusion matrix (CE + KL + RKD)

-> 최종적으로 distillation을 해도 기본 Timnet의 경향성을 따라가는 것을 확인.

WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

- Conclusion
  - 제안된 distillation 방법은 CNN 기반의 경량화 음성 감정 인식 모델에 효과적인 모습을 나타냄
  - 기본적인 logit knowledge 뿐만 아니라 특징들 간의 상대적인 knowledge를 전달함으로써 더욱 향상된 student 모델을 획득할 수 있었음
  - 실험 결과 중 sad class에 대한 정확도가 낮은 경향을 보였으므로 이 점을 보완하는 부분을 향후 연구 내용에 포함할 것임
- Future work
  - RAVDESS 데이터 외에 다른 데이터 추가
  - Sad class의 정확도가 낮으므로 클래스간 구분을 잘 할 수 있게 contrastive learning 추가
  - 다른 distillation 방법들도 시도해 볼 예정

감사합니다

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의
지원을 받아 수행된 연구임(NRF-2023R1A2C2005916).