WAV에서 MFCC로의 지식 증류을 이용한 음성감정인식

홍윤아[1], 이보경[1], 구본화[1], 고한석[1]
[1] 고려대학교

SPEECH EMOTION RECOGNITION USING KNOWLEDGE DISTILLATION WITH WAV TO MFCC

Yuna HONG[1], Bokyeung LEE[1], Bonhwa KU[1], Hanseok KO[1]
[1] Korea University

중심어: 음성감정인식, 지식 증류

요약

음성 감정인식은 음성 정보를 통해 사용자의 의도 및 현재 상태를 파악할 수 있기 때문에 human-computer interaction에서 중요한 부분을 차지하고 있다. 최근 딥러닝 기반의 고성능 음성 감정인식 모델들은 raw audio signal을 입력으로 하여 높은 정확도를 가지지만 resource 및 연산량 측면에서 상당한 비용이 발생하게 된다. 또한 inference 시간 측면에서 실생활에 활용하기 위해서는 아직 한계가 있다. 반면에 MFCC를 활용하는 경량화 모델들은 연산량 측면에서 우수한 모습을 보이지만 낮은 감정 인식 성능을 나타내고 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 wav2mfcc knowledge distillation을 제안한다. 제안한 방법은 wav를 입력으로 하는 teacher 모델의 풍부한 지식을 MFCC를 입력으로 하는 student 모델에게 전달하여 경량화 모델의 성능을 향상시키고자 한다.

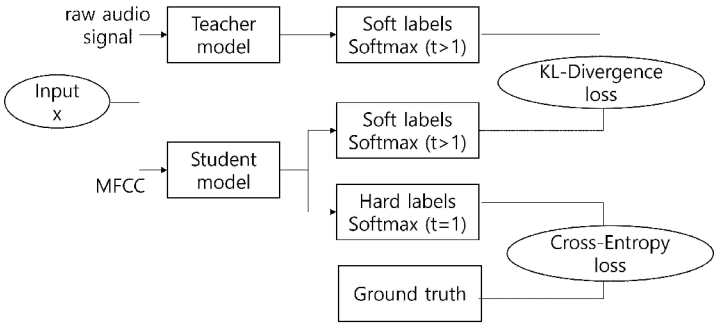

그림1. overview of wav2mfcc knowledge distillation

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

참고문헌

1. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
2. Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33, 12449-12460 (2020).

고한석 hsko@korea.ac.kr