채널 어텐션을 통한 짧은 발화 대상 화자 확인
Short Utterance Speaker Verification based on Channel Attention

김규진 (Kyujin Kim), 박정식(Jeong-sik Park).
언어공학연구소 한국외국어대학교
(Language Technology Research Institute, Hankuk University of Foreign Studies)
rndlwjs@gmail.com, parkjs@hufs.ac.kr

1 서론

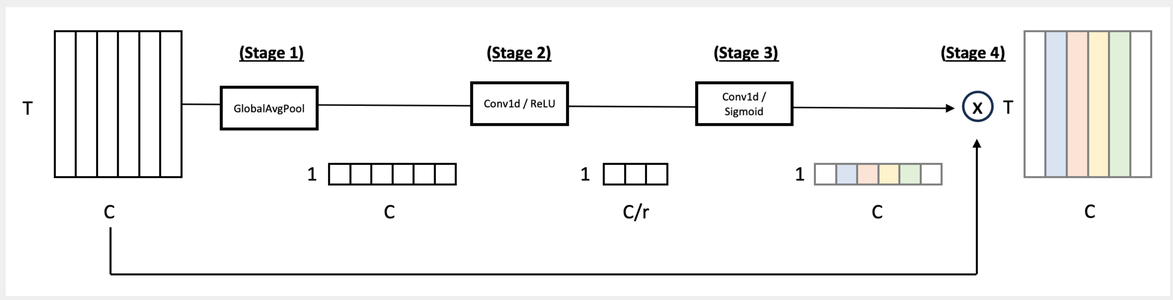
- ECAPA-TDNN은 문장독립형(Text-Independent) 화자 확인 문제에 효율적인 모델
- 짧은 발화 혹은 다양한 언어 등 환경 변이 상황에서 모델 성능이 저하되는 것 확인
- ECAPA-TDNN 내 Res2Net의 구조는 짧은 발화 화자 확인에 유리한 구조를 가짐
  - Res2Net의 Scale 내 채널 어텐션을 적용
- 2D Convolutional 모듈을 통해 프레임 정보를 더 반영하는 특징맵 생성
  - 2D Convolutional 모듈 도입

2 연구배경

- Res2Net은 특징맵을 scale 단위로 자른 후, 단위별로 지역적인 정보를 학습하여 계층적으로 전달하는 구조
  - 이전 scale에서 학습된 특징은 잔차 연결을 통해 다음 특징에 전달되며, 이는 지역적인 특징을 효과적으로 학습하는 구조
  - scale 단계에서 화자 확인에 필요한 특징에 채널 어텐션을 적용하면 더 정확한 특징을 전달할 것이라 기대
- 2D ResNet 기반 특징 추출 방법을 1D TDNN에 적용하는 모듈

3 제안한 방법 - 1

- 채널 어텐션은 채널 특징 간의 중요도를 학습하여, 유의미한 채널을 돋보이게 유도하는 매커니즘
- 채널 어텐션의 4단계 수행방법

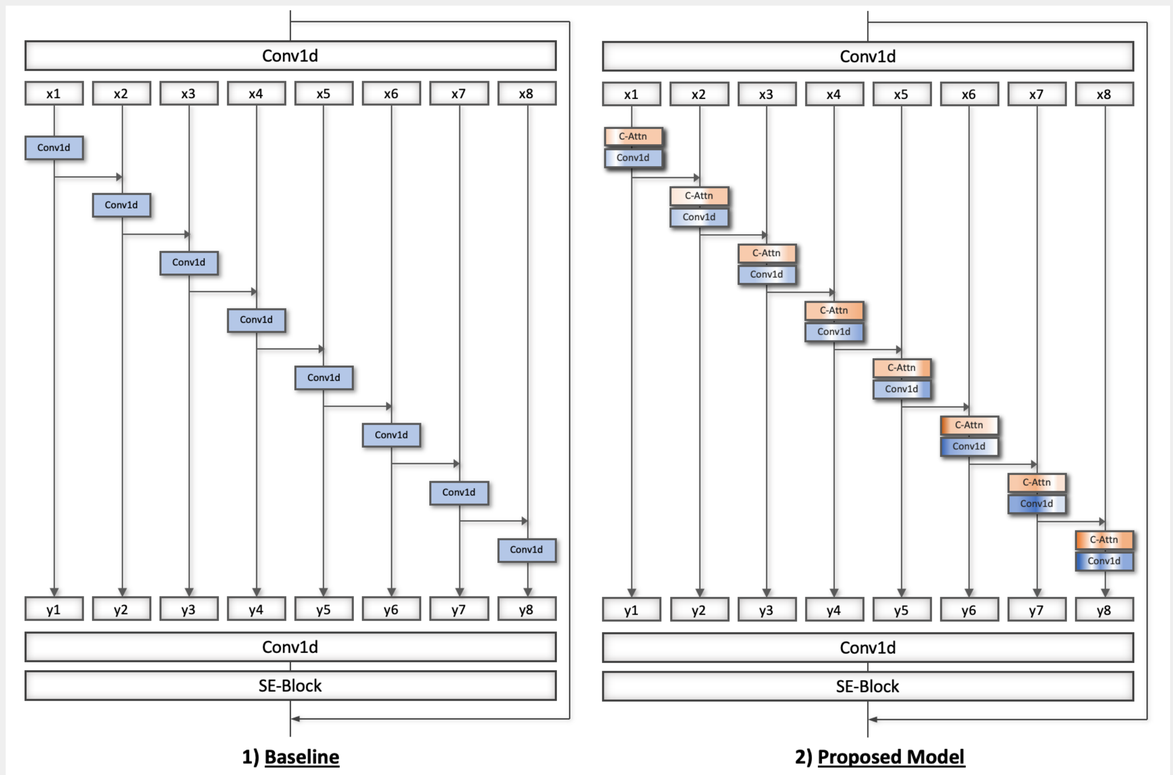


Stage 1) 모든 채널의 시간에 해당되는 특징을 Global Average Pooling을 통해 압축
- 특징맵(C X T)를 (C X 1) 크기로 압축함

Stage 2) 컨볼루션 레이어를 통과하여 Inter-Channel Dependency 학습
- 컨볼루션 레이어를 통해 압축된 채널 간의 관계를 학습
- Reduction Ratio(r)에 따라 출력 채널 차원을 축소할 수 있음
- 출력 채널 차원 축소시 오버피팅 방지 및 연산량을 줄일 수 있음

Stage 3) 다시 한번 채널 정보 간의 연관성을 학습함
- 앞서 축소된 채널을 원본 채널 수에 맞게 출력
- 출력된 채널은 시그모이드 활성화 함수를 통과하여 각 채널의 중요도 점수 계산

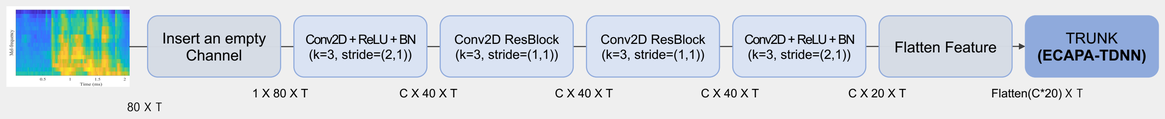
Stage 4)마지막으로 계산된 중요도를 기존 입력값에 곱하여 중요한 특징에 채널의 차이를 표현



1) Baseline　　2) Proposed Model

- Baseline 모델의 Res2Net은 채널 차원을 기준으로 8개의 scale 단위로 자름
- 채널 사이즈는 512차원으로 선정
- 지역적인 특징을 scale 내 학습하기 전, 채널 어텐션을 적용하면 중요도에 따라 나눈 특징맵을 점차적으로 전달할 수 있음
- 기존 Res2Net 후반부에 위치한 SE-Block은 특징을 종합하여 전역적 상호관계에 따른 중요도를 구하는 반면, 본연구는 채널 어텐션을 지역적 학습에 반영하여 전달할 수 있게 설계

4 제안한 방법 - 2



(a) 2D Convolutional Module

- 2D 컨볼루션 도입 이유:
  - 음성 특징의 시간 정보를 유지하면서, 프레임 정보를 자세하게 처리함
  - Trunk 모델의 구조 변경 없이 특징 추출 레이어만 수정 가능
- Empty Channel을 도입하여 프레임과 시간 정보를 모두 채널 특징맵으로 변환
- 커널 사이즈 및 스트라이드를 통해 인접 특징을 반영하는 특징맵 생성
- 사전 연구에 따라 채널 사이즈는 128차원으로 선정

5 실험결과

		Voxceleb1-O (Long)	Google Speech Command (Short)	Common Voice (Short and Cross Lingual)
	Baseline	EER 7.81% minDCF 0.4466%	EER 11.07% minDCF 0.5975%	EER 22.78% minDCF 0.7686%
r = 1	C_attn Model	EER 7.56% minDCF 0.4347%	EER 10.92% minDCF 0.5506%	EER 20.72% minDCF 0.6532%
r = 2	**C_attn Model**	EER 7.46% minDCF 0.4338%	EER 10.36% minDCF 0.4976%	EER 19.50% minDCF 0.7104%
r = 4	C_attn Model	EER 7.63% minDCF 0.4406%	*EER 12.65%* *minDCF 0.6048%*	EER 21.94% minDCF 0.8057%

- Channel Attention을 Scale내 도입한 결과, 전반적인 성능향상을 확인
  -> Channel은 512차원 크기의 특징맵이며, 각 scale은 64차원 크기의 특징맵
  -> r=2인 경우, 가장 높은 성능을 보임
  -> r=4인 경우, Baseline 대비 전반적인 성능향상이 있지만, Google Speech Command 데이터 성능 저하

		Voxceleb1-O (Long)	Google Speech Command (Short)	Common Voice (Short and Cross Lingual)
(1)	Baseline	EER 7.81% minDCF 0.4466%	EER 11.07% minDCF 0.5975%	EER 22.78% minDCF 0.7686%
(1a)	2D CNN + Baseline	EER 7.26% minDCF 0.4559%	EER 10.41% minDCF 0.5118%	EER 16.59% minDCF 0.7263%
(2)	C_attn Model	EER 7.46% minDCF 0.4338%	EER 10.36% minDCF 0.4976%	EER 19.50% minDCF 0.7104%
(2a)	**2D CNN + C_attn Model**	EER 6.97% minDCF 0.4390%	EER 9.59% minDCF 0.3988%	EER 16.36% minDCF 0.7400%

- Baseline 모델 (1), (1a)에 비해, 채널 어텐션을 도입한 구조 (2), (2a)는 성능 개선
- 일반적인 문장으로 구성된 Voxceleb1 데이터의 성능개선율 4.6%, 4.2%
- 1초 길이 이내로 구성된 Google Speech Command 데이터의 성능개선율 6.9%, 8.6%
- 평균 2.4초 내외로 구성된 Common Voice 데이터 성능 개선율 각 16.8%, 1.4%

6 결론

이 같은 실험을 통해, 채널 어텐션을 ECAPA-TDNN의 scale내에 적용함으로써 화자 정보를 학습하는 효율성이 개선되고 이를 통해 짧은 길이의 음성에서 눈에 띄게 성능이 개선됨을 확인하였다. 이는 Res2Net에서 scale 단위로 특징맵을 처리하는 방법은 짧은 발화에서 효과적인 방법임을 나타내며, 일반 화자확인을 개선하는 모듈을 짧은 발화 화자확인에 맞게 구조화하여 개선시킬 수 있다는 점을 보여준다.