자기 지도 학습 모델의 다계층 특징을 활용한 짧은 발화 화자인식 연구

명지수, 한상욱, 신종원*
광주과학기술원

myoungjs@gm.gist.ac.kr, swhan9873@gm.gist.ac.kr, *jwshin@gist.ac.kr

Short-Utterance Speaker Verification with Multi-Level Features from Self-Supervised Model

Jisoo Myoung, Sangwook Han, Jong Won Shin*
Gwangju Institute of Science and Technology (GIST)

요약

본 논문은 화자인식 모델인 ECAPA-TDNN 에 국지적 조건화 정보로 다해상도 정보를 사용하는 기존의 방법을 확장하여 자기 지도 학습된 모델로부터 나온 여러 은닉 표현들의 융합된 특징 정보를 얻어 추가 활용해 짧은 발화 화자인식에서의 성능 저하를 완화하는 방안을 제시하였다. 트랜스포머 층에서 나온 은닉 표현들을 네 가지 조합으로 가중합하여 다양한 다계층 특징 정보를 추출하였고, 제안된 시스템을 화자인식의 벤치마크 데이터셋인 VoxCeleb1 으로 학습 및 평가하여 성능 지표 EER (Equal Error Rate)을 기준으로 전체 길이를 포함한 짧은 길이 발화에 대하여 성능 향상을 이루었다.

Ⅰ. 서 론

화자인식(Speaker Verification)은 등록이 되어있는 목소리와 시스템에 접근한 사람의 목소리의 동일 여부를 판별하는 과정이다. 최근 몇 년간 딥러닝 기반의 x-vector 시스템이 뛰어난 성능을 보여왔으며, 이들 시스템은 입력된 음성 데이터를 프레임 단위로 특징을 추출한 후, 시간 축 통계 풀링(Statistical Pooling)을 거쳐 발화 단위 임베딩(Embedding)을 생성한다.

그러나 기존 시스템들은 상당히 긴 길이의 테스트 발화에 대해 성능을 평가하고 있으며, 테스트 발화가 짧아질수록 성능이 크게 저하되는 결과를 보였다. 이에 짧은 발화에서의 성능 저하를 완화하기 위한 연구가 진행되어 왔다[1-3]. 특히 [1]에서는 여러 시간 해상도 정보를 통합한 특징 정보를 출력하는 Multi-resolution encoder (MRE)를 제안하였고, ECAPA-TDNN 의 각 SE-Res2Block 에 앞 단에 어댑터(Adapter) 모듈을 위치시켜 MRE 의 출력을 국지적 조건화(local conditioning) 정보로 주입하였다. 이는 ECAPA-TDNN 의 입력인 로그 멜-필터뱅크(log mel-filterbank)가 갖는 한계점들을 보완하며 기존 ECAPA-TDNN 대비 2 초 이하 길이의 발화에서 성능 향상을 이루었다.

본 논문에서는 대규모 음성 데이터로 자기 지도 학습(self-supervised learning, SSL) 모델의 여러 트랜스포머(Transformer) 계층에서 나온 은닉 표현(hidden representation)들을 가중합을 통해 네 가지로 융합(fusion)하여 추가 정보로 활용하였다. 해당 융합 특징 정보들을 MRE 의 다해상도 특징 정보와 함께 국지적 조건화 정보로 주입하였고, VoxCeleb1 데이터셋에 대하여 성능 지표 EER (Equal Error Rate)로 성능을 평가하여 2 초 길이 발화에서 기존 시스템 대비 약 32%의 성능 향상을 확인하였다.

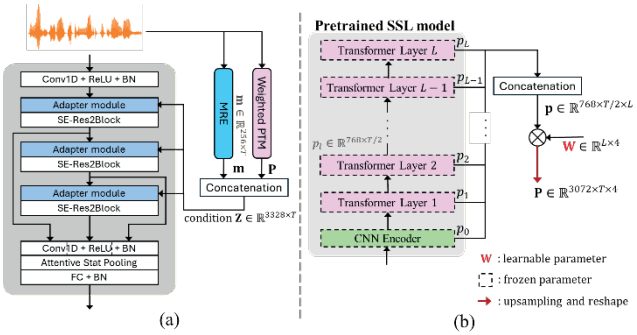

그림 1. (a) 제안된 시스템, (b) 사전 학습된 자기 지도 학습 모델(pre-trained self-supervised model, PTM)의 은닉 표현들의 융합 과정.

Ⅱ. 본 론

그림 1(a)는 제안된 시스템의 전체적인 구성을 보여준다. 기존 MRE 는 50, 100, 200, 400 의 프레임 길이를 갖는 single resolution encoder 들로 구성이 되어있다. MRE 는 내부에서 single resolution encoder 로 얻은 서로 다른 시간 해상도의 특징 정보들을 새로운 축으로 concatenation 하여 다해상도 특징 정보를 생성한다. 그림 1(b)는 자기 지도 학습 모델의 각 트랜스포머 계층에서 나온 은닉 표현 $p_1, p_2, \dots, p_L \in \mathbb{R}^{768\times T/2}$ 들의 융합 과정을 보여준다. 은닉 표현들을 새로운 레이어 L축으로 concatenation 한 후 학습 가능한 파라미터 $\mathbf{W} \in \mathbb{R}^{L\times 4}$ 와 행렬 곱셈을 수행하여 계층 방향으로 가중합된 다계층 특징 정보들 $\mathbf{p}'$을 생성한다.

$$\mathbf{p} = \text{concatenation}([p_1, p_2, \dots, p_{L-1}]) \in \mathbb{R}^{768\times T/2\times L} \quad (1)$$

$$\mathbf{p}' = \mathbf{p}\cdot\mathbf{W} \in \mathbb{R}^{768\times T/2\times 4},\ \mathbf{W} = [\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3] \quad (2)$$

Systems	Condition		EER (%)					
	MRE	PTM	Full	5s	2s	1.5s	1s	0.5s
ECAPA	✗	✗	2.59	2.79	4.42	5.50	7.75	15.36
+ MRE	✓	✗	2.66	2.88	4.21	5.14	7.31	14.25
Proposed	✓	✓	1.60	1.72	2.86	3.81	5.97	13.99

표 1. VoxCeleb1-O 에 대한 베이스라인과 제안된 시스템의 EER (%) 성능.
0.5~5s 길이는 테스트 발화의 중간 구간을 잘라 사용하였음.

이때 자기 지도 학습 모델의 은닉 표현들은 시간 축 해상도가 20ms 로 시스템의 입력의 해상도인 10ms 와 다르기 때문에 업샘플링(Up-sampling)을 통해 동일하게 맞추어 주었다. 이후 네 가지의 다계층 특징 정보들로 구성된 $\mathbf{p}'$의 차원을 재구성(reshape)하고 feature 축으로 concatenation 해주어 융합 특징 정보 $\mathbf{P}$를 생성한다. 최종적으로 국지적 조건화 정보 $\mathbf{z}$는 융합 특징 정보 $\mathbf{P}$를 MRE 의 다해상도 특징 정보 $\mathbf{m} \in \mathbb{R}^{256\times T}$ 와 feature 축으로 concatenation 하여 구해진다.

$$\mathbf{P} = \text{reshape}\big(\text{upsampling}(\mathbf{p}')\big) \in \mathbb{R}^{3072\times T} \quad (3)$$

$$\mathbf{z} = \text{concatenation}([\mathbf{m}, \mathbf{P}]) \in \mathbb{R}^{3328\times T} \quad (4)$$

생성된 $\mathbf{z}$는 ECAPA-TDNN 의 어댑터 모듈에 입력으로 들어가 ECAPA-TDNN 의 은닉 표현을 국지 조건화 한다.

Ⅲ. 실 험

제안된 시스템에 사용된 사전 학습 모델은 13 개의 트랜스포머 계층으로 구성되며 768 의 은닉 차원을 갖는 WavLM Base+ [4]를 사용하였으며, 학습 과정에서 자기 지도 학습 모델의 파라미터는 고정하였다. 실험에 사용된 ECAPA-TDNN 의 채널 크기는 1024 이며 모든 시스템은 VoxCeleb1 데이터셋의 development set 과 test set 으로 학습 및 평가하였다. 이때 기존 시스템과의 공정한 비교를 위하여 ECAPA-TDNN 과 MRE 를 재구현하였다.

학습 과정에서 시스템의 입력은 기존 논문과 동일하게 각 발화에서 2 초 구간을 무작위로 잘라 프레임 길이 25ms, 프레임 이동 간격이 10ms 인 80 차원 로그 멜-필터뱅크로 형태를 변환하여 사용하였다. 데이터 증강을 위하여 simulated room impulse response (RIR)와 MUSAN noise 데이터셋을 사용하였으며, 이후 speed perturbation 을 적용하였다. 음성 활동 감지(Voice activity detection, VAD)는 적용하지 않았으며 목적함수 (objective function)는 AAM-Softmax 를 0.2 의 magin 과 30 의 scale 로 사용하였다. Adam Optimizer 가 사용되었으며 학습률(learning rate)은 0.001 로 초기화하였고 매 에폭마다 0.97 배로 감소시키면서 총 100 에폭 동안 학습을 진행하였다.

표 1 은 ECAPA-TDNN 과 이에 MRE 의 다해상도 특징 정보를 사용하여 국지 조건화한 베이스라인과 제안된 시스템의 성능을 제시한다. 전체 길이(full)를 제외한 짧은 발화에 대한 성능 평가는 테스트 발화의 중간 구간을 기준으로 0.5s~5s 길이를 사용하였다. 실험 결과 제안된 시스템이 베이스라인 대비 모든 발화 길이에서 성능 향상을 이루는 것을 확인하였다. 특히 2 초, 1.5 초 1 초 길이의 테스트 발화가 주어졌을 때 EER 기준 각각 32.07, 25.88, 18.33% 가량 성능 향상을 보였으며, 이를 통해 MRE 와 로그 멜-필터뱅크에서 얻을 수 없던 특징 정보를 제안된 방법으로 보충하여 짧은 발화 시나리오에서 효과적으로 활용하였음 입증하였다. 그림 2 는 학습 후의 가중치 $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^L$의 값을 나타낸다. 색상이 어두울수록 융합 특징 정보를 생성 시에 해당 계층의 은닉 표현을 강하게 반영하였음을 의미한다. 그림 2 를

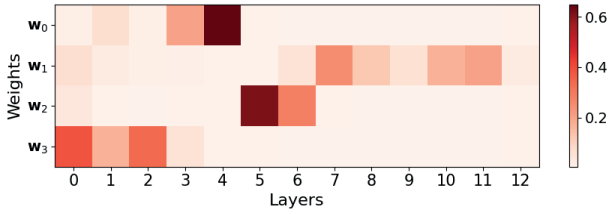

그림 2. WavLM 의 은닉 표현 별 가중치.

통하여 각 가중치가 특정 위치의 계층들을 강하게 반영한 것을 확인할 수 있으며, 서로 다른 위치에 집중한 것으로 보아 자기 지도 학습 모델의 다양한 정보를 활용하였음을 추측할 수 있다.

Ⅳ. 결 론

본 논문에서는 ECAPA-TDNN 에 MRE 의 다해상도 특징 정보에 자기 지도 학습 모델의 은닉 표현들의 융합된 네 가지 다계층 특징 정보를 추가하여 기존 시스템에서 활용되지 못했던 풍부한 음성 정보를 활용하는 방안을 제안하였다. 제안된 시스템은 매우 짧은 0.5 초를 포함한 모든 길이의 발화에 대하여 기존 시스템 대비 성능 향상을 이루었으며 특히 2 초 길이에 대하여 32.07%의 성능 개선을 확인하여 제안된 방법의 효과를 입증하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및
정보통신기획평가원의 대학 ICT 연구센터지원사업의
연구결과로 수행되었음.
(IITP-2025-RS-2021-II211835)

참 고 문 헌

[1] S. Han, Y. Ahn, K. Kang and J. W. Shin, “Short-segment speaker verification using ECAPA-TDNN with multi-resolution encoder,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[2] S.-B. Kim, C.-Y. Lim, J. Heo, J.-H. Kim, et al., “MR-RawNet: Speaker verification system with multiple temporal resolutions for variable duration utterances using raw waveforms,” in *Proc. Interspeech*, 2024.

[3] T. Liu, R. K. Das, K. A. Lee, and H. Li, “MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[4] S. Chen, C. Wang, Z. Chen, Yu Wu, et al., “WavLM: Large-scale self-supervised pretraining for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Proc.*, 2022.