한국음향학회지 제40권 제5호 pp. 496~502 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.496

pISSN : 1225-4428
eISSN : 2287-3775

문장 독립 화자 검증을 위한 그룹기반 화자 임베딩

Group-based speaker embeddings for text-independent speaker verification

정영문,[1†] 엄영식,[1] 이영현,[1] 김회린[1]

(Youngmoon Jung,[1†] Youngsik Eom,[1] Yeonghyeon Lee,[1] and Hoirin Kim[1])

[1]KAIST 전기및전자공학부

(Received July 16, 2021; accepted August 23, 2021)

초 록: 딥러닝 기반의 심층 화자 임베딩 방식은 최근 문장 독립 화자 검증 연구에 널리 사용되고 있으며, 기존의 i-vector 방식에 비해 더 좋은 성능을 보이고 있다. 본 연구에서는 심층 화자 임베딩 방식을 발전시키기 위하여, 화자의 그룹 정보를 도입한 그룹기반 화자 임베딩을 제안한다. 훈련 데이터 내에 존재하는 전체 화자들을 정해진 개수의 그룹으로 비지도 클러스터링 하며, 고정된 길이의 그룹 임베딩 벡터가 각각의 그룹을 대표한다. 그룹 결정 네트워크가 각 그룹에 대응되는 그룹 가중치를 출력하며, 이를 이용한 그룹 임베딩 벡터들의 가중 합을 통해 집합 그룹 임베딩을 추출한다. 최종적으로 집합 그룹 임베딩을 심층 화자 임베딩에 더해주어 그룹기반 화자 임베딩을 생성한다. 이러한 방식을 통해 그룹 정보를 심층 화자 임베딩에 도입함으로써, 화자 임베딩이 나타낼 수 있는 전체 화자의 검색 공간을 줄일 수 있고, 이를 통해 화자 임베딩은 많은 수의 화자를 유연하게 표현할 수 있다. VoxCeleb1 데이터베이스를 이용하여 본 연구에서 제안하는 방식이 기존의 방식을 개선시킨다는 것을 확인하였다.

핵심용어: 심층 화자 임베딩, 그룹기반 화자 임베딩, 문장 독립 화자 검증, 클러스터링

ABSTRACT: Recently, deep speaker embedding approach has been widely used in text-independent speaker verification, which shows better performance than the traditional i-vector approach. In this work, to improve the deep speaker embedding approach, we propose a novel method called group-based speaker embedding which incorporates group information. We cluster all speakers of the training data into a predefined number of groups in an unsupervised manner, so that a fixed-length group embedding represents the corresponding group. A Group Decision Network (GDN) produces a group weight, and an aggregated group embedding is generated from the weighted sum of the group embeddings and the group weights. Finally, we generate a group-based embedding by adding the aggregated group embedding to the deep speaker embedding. In this way, a speaker embedding can reduce the search space of the speaker identity by incorporating group information, and thereby can flexibly represent a significant number of speakers. We conducted experiments using the VoxCeleb1 database to show that our proposed approach can improve the previous approaches.

Keywords: Deep speaker embedding, Group-based speaker embedding, Text-independent speaker verification, Clustering

PACS numbers: 43.71.Bp, 43.72.Fx

†**Corresponding author:** Youngmoon Jung (dudans@kaist.ac.kr)
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
(Tel: 82-42-350-7617, Fax: 82-42-350-7619)

I. 서 론

화자 검증은 사용자가 발성한 음성신호를 이용하여 사용자가 해당 시스템에 허가된 사람인지 아닌지 판별하는 태스크이다. 크게 개발, 등록, 그리고 테스트 단계로 나뉘는데, 개발 단계에서는 화자 검증 모델을 훈련하고, 등록 단계에서는 훈련된 화자 검증 모델을 이용하여 등록 화자의 화자 임베딩을 생성한다. 마지막으로 테스트 단계에서는 입력 발화로부터 생성된 화자 임베딩과 등록 화자 임베딩 간의 유사도를 계산하며, 이를 통해 테스트 화자가 등록된 화자가 맞는지 판별한다. 화자 검증은 크게 문장 종속 화자 검증과 문장 독립 화자 검증으로 나뉜다. 문장 종속 화자 검증은 사용자가 발성을 할 때 사전에 정의된 고정된 문장만을 발성해야 하며, 문장 독립 화자 검증에서는 이와 다르게 사용자가 자유롭게 발성할 수 있다. 문장 독립 화자 검증은 음성적 변이성으로 인해 상대적으로 더 어려운 태스크이지만, 사용자의 편의성은 더 높다.[1]

전통적으로 *i*-vector[2]/PLDA[3] 기법이 문장 독립 화자 검증 태스크에 널리 사용되었다. 이 방식은 긴 등록/테스트 발화(보통 10 s 이상)에 대해서는 잘 동작하지만, 짧은 등록/테스트 발화(보통 10 s 이하)에 대해서는 성능 하락을 보인다.[4] 딥러닝의 발전과 더불어 심층 화자 임베딩 방식이 도입되었으며, *i*-vector/PLDA 기법에 비해 짧은 등록/테스트 발화에 대해서 좋은 성능을 보이고 있다.[5-7]

심층 화자 임베딩 방식은 화자 검증모델로부터 추출된 심층 화자 특징에 전역 풀링 기법을 적용하여 화자 임베딩을 생성하는 방식이다. 모델에는 TDNN,[8] VGG,[9] 그리고 ResNet[10]과 같은 Convolutional Neural Network(CNN)이 주로 이용된다. 전역 풀링 기법으로는 Global Average Pooling(GAP)[11], Self-Attentive Pooling (SAP),[12] 그리고 Spatial Pyramid Encoding(SPE)[13] 방식 등이 이용된다. 일반적으로 네트워크는 훈련 데이터 셋의 모든 화자들을 분류하도록 훈련된다. 이 때 softmax loss,[14] angular softmax(A- Softmax) loss,[15] 그리고 additive margin softmax loss[16] 등의 비용 함수가 이용된다.

본 연구에서는 이러한 심층 화자 임베딩 방식을 발전시키기 위하여, 화자의 그룹 정보를 도입한 그룹기반 화자 임베딩을 새롭게 제안한다. 심층 화자 특징을 이용하여 각 화자 그룹을 대표하는 그룹 임베딩을 생성하며, 그룹 결정 네트워크를 통해 각 그룹에 대응하는 그룹 가중치를 생성한다. 이 둘의 가중 합을 통해 그룹 정보를 취합하며, 이렇게 생성된 벡터를 심층 화자 임베딩에 더함으로써 최종적으로 그룹기반 화자 임베딩을 생성한다. 이러한 방식을 통해 그룹 정보를 심층 화자 임베딩에 도입함으로써, 화자 임베딩이 나타낼 수 있는 전체 화자의 검색 공간을 효과적으로 줄일 수 있다. 따라서 그룹기반 화자 임베딩은 고정된 길이의 벡터로도 많은 수의 화자를 유연하게 표현할 수 있다.

본 논문의 구성은 다음과 같다. II장에서는 그룹기반 화자 임베딩을 학습하기 위한 전체적인 방식에 대해 설명하고, III장에서는 제안한 기법을 이용한 실험 및 결과를 다루며, 마지막으로 IV장에서는 결론 및 향후 연구 계획을 서술한다.

II. 그룹기반 화자 임베딩

2.1 전체적인 구조

본 연구에서 제안하는 시스템의 전체적인 구조는 Fig. 1과 같다. 이 구조는 얼굴 인식 태스크에서 제안된 GroupFace 연구[17]에 착안하여 설계되었다. GroupFace 방식은 여러 그룹인지 표현들을 이용하여 얼굴 임베딩의 성능을 향상시키는 방식으로, 얼굴 검증 및 얼굴 식별 태스크에서 최고 성능을 달성하였다. 본 연구에서 제안하는 방식은 화자의 그룹 정보를 화자 임베딩에 도입하여 화자 검증 성능을 향상시킨다. 사람들의 음성은 개인의 고유한 특성을 갖는데, 그와 동시에 사람들마다 공통적으로 갖는 특성도 존재한다. 그러한 특성을 바탕으로 사람들을 여러 그룹으로 나누어 생각할 수 있으며, 본 연구에서는 그룹기반 화자 임베딩을 통해 그러한 특성을 나타낸다. 기존의 심층 화자 임베딩에 화자의 그룹 정보를 추가함으로써, 화자 임베딩이 표현할 수 있는 화자의 후보군의 수를 줄일 수 있게 된다. 그림을 바탕으로 본 연구에서 제안하는 방식을 설명하면 다음과 같다.

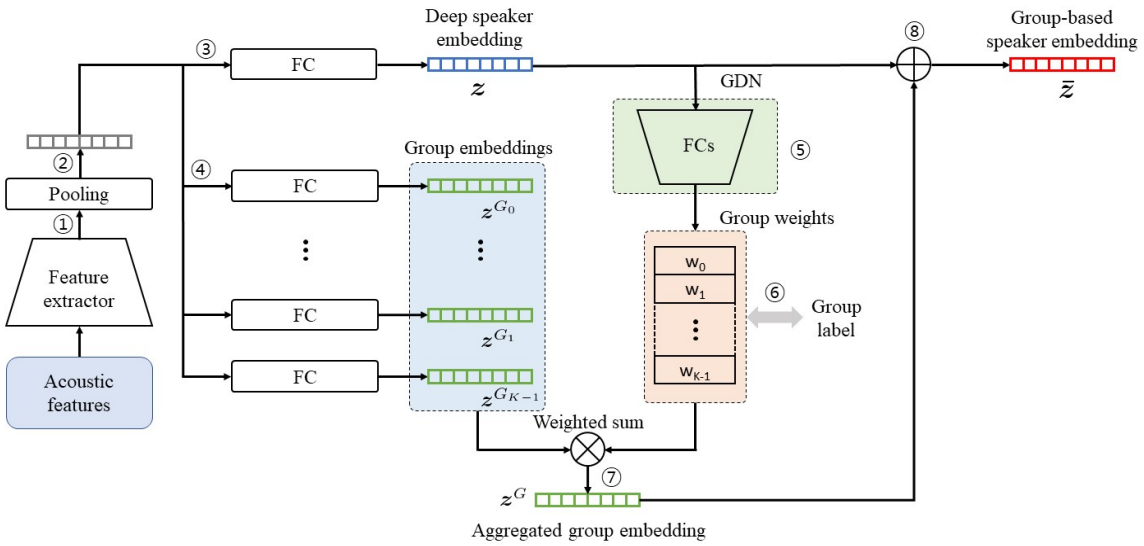

Fig. 1. (Color available online) Overall structure of the proposed method. FC : fully-connected layer, GDN : group decision network.

우선, 입력 음성의 음성 특징 벡터열이 ResNet 기반의 특징 추출기의 입력으로 이용된다. 그 결과 심층 화자 특징이 생성되는데(과정 ①), 이를 128차원의 고정된 길이의 벡터로 변환하기 위하여 GAP 기반의 전역 풀링을 적용한다(과정 ②). 이렇게 생성된 벡터는 Fully-Connected (FC) 레이어를 통과하여 심층 화자 임베딩인 $z \in \mathbb{R}^{128}$가 생성된다(과정 ③). 한편, GAP 결과로 생성된 벡터는 K개의 독립적인 FC 레이어들을 통과하며, 그 결과 K개의 그룹 임베딩 벡터들($z^{G_k} \in \mathbb{R}^{128}$ for $\mathrm{k}=0,1,\cdots,\mathrm{K}-1$)이 생성된다(과정 ④). Deep speaker embedding z는 그룹 결정 네트워크의 입력으로 이용되며, 그 결과 각 그룹 임베딩 벡터들에 대응되는 그룹 가중치(w_k for $\mathrm{k}=0,1,\cdots,\mathrm{K}-1$)들이 생성된다(과정 ⑤). 그룹 임베딩 벡터들과 그룹 가중치들의 가중 합을 통해 집합 그룹 임베딩($z^G \in \mathbb{R}^{128}$)을 생성한다.

$$\boldsymbol{z}^G = \sum_{k=0}^{K-1} w_k \boldsymbol{z}^{G_k}. \tag{1}$$

최종적으로, z^G를 z에 더해 줌으로써 그룹기반 화자 임베딩($\bar{z} \in \mathbb{R}^{128}$)을 생성한다.

$$\bar{\boldsymbol{z}} = \boldsymbol{z}^G + \boldsymbol{z}. \tag{2}$$

전체 모델은 훈련 데이터셋에 존재하는 전체 화자를 분류하도록 훈련된다. 따라서 가장 마지막 출력 레이어는 전체 화자수와 동일한 개수의 node를 갖는 FC 레이어이며, $\bar{z}$를 입력으로 한다. 그 결과에 softmax 함수를 취하여 화자를 예측하며, 교차 엔트로피 비용함수를 통해 훈련된다.

2.2 그룹 결정 네트워크

그룹 결정 네트워크는 그룹 임베딩 벡터 $\boldsymbol{z}^{G_k}$에 대응되는 그룹 가중치인 w_k를 출력하며, 3개의 FC 레이어와 1개의 sigmoid 레이어로 구성된다. FC 레이어는 128개의 node로 구성되며, 마지막의 sigmoid 레이어를 통해 K개의 그룹(G_k for $\mathrm{k}=0,1,\cdots,\mathrm{K}-1$)에 대응하는 확률을 출력한다. 이 확률들이 결국 각 그룹에 대응되는 가중치를 의미한다.

$$w_k = p(G_k \mid \mathrm{x}). \tag{3}$$

GroupFace 연구에서 그룹 가중치를 계산할 때 softmax 함수를 이용한 것과 다르게 본 연구에서는 sigmoid 함수를 이용한다. Softmax 함수를 이용하면 결과 확률 값에 sum-to-one 제약 조건이 존재하는 반면에, sigmoid의 경우는 그러한 제약 조건이 없다. GroupFace 연구에서는 각 그룹이 사람의 얼굴을 표현하기에 적

합한 특성들에 각각 대응된다. 예를 들어, 그룹 1은 '머리카락이 금색'인 그룹을 나타내며, 그룹 2는 '검은색 콧수염'을 가진 그룹을 나타낸다고 가정해보자. 이때, 금발이면서 검은색 콧수염을 가진 사람은 그룹 1과 그룹 2에서 모두 높은 가중치를 가질 것이다. 본 연구에서 제안하는 방식 역시 이와 마찬가지로 동작하며, 각 그룹은 서로 독립적인 특성을 나타내므로, 개개인의 음성이 여러 그룹의 특성들을 동시에 포함할 수 있다. 따라서 그룹 가중치에 sum-to-one 제약 조건이 없는 sigmoid 함수를 이용하는 것이 그룹 정보를 나타내는데 있어 더 자연스럽다. 그룹들이 이러한 방식으로 학습되는 이유는 self-distributed labeling 방식을 통해 네트워크가 학습되기 때문이며, 이는 다음 문단에서 설명될 예정이다. III장에서 그룹 가중치 계산에 softmax 함수를 이용할 때 보다 sigmoid 함수를 이용할 때 더 좋은 성능을 보임을 확인한다.

그룹 결정 네트워크는 GroupFace 연구와 마찬가지로 그룹 레이블을 이용하여 훈련이 이루어진다. 그룹 레이블을 생성할 때, self-distributed labeling 방식을 이용한다. 이는 K개의 그룹에 데이터 샘플들을 최대한 균일하게 할당하기 위한 것으로, 그룹 레이블이 최대한 균일 분포에 따르도록 한다. 이를 통해 특정 소수의 그룹만이 큰 가중치를 갖는 현상을 방지할 수 있다. 생성된 그룹 레이블과 그룹 결정 네트워크의 출력 사이의 비용 함수를 교차 엔트로피 함수로 정의하며, 이를 L_{GDN}라고 한다. $\bar{\boldsymbol{z}}$를 이용하여 훈련 데이터셋 내의 전체 화자를 분류할 때 사용하는 교차 엔트로피 비용 함수를 L_{cl}이라 하면, 최종 비용 함수는 다음과 같이 정의된다.

$$L = L_{cl} + \lambda L_{GDN}, \tag{4}$$

여기서 λ는 두 비용 함수 사이의 균형을 맞추기 위한 비용 함수 가중치이며, 0.1로 지정하였다. 최종 비용 함수인 L을 이용하여 전체 모델이 훈련된다. 이를 통해 모델은 화자 검증에 유용한 그룹 임베딩을 학습할 수 있으며, 최종적으로는 그룹 정보를 포함한 화자 임베딩을 생성할 수 있다.

III. 실험 및 결과

3.1 데이터셋

본 연구에서는 화자검증을 위한 대표적인 벤치마크 데이터셋 중 하나인 VoxCeleb1[11] 데이터셋을 이용하였다. VoxCeleb1 데이터셋은 문장 독립 화자 검증을 위한 대규모의 데이터셋으로, 1250명의 화자가 발성하였다. 데이터셋 내의 발화들은 유튜브 비디오로부터 추출된 것으로, 실생활 잡음에 의해 오염되었다. 데이터셋은 development set과 test set으로 나뉘며, 두 세트 간에 겹치는 화자가 존재하지 않는다.

3.2 실험 구성

본 연구에서 사용되는 모든 심층 화자 임베딩 방식에서는 64차원의 log Mel-filterbank 특징 벡터를 입력으로 이용하였으며, 별도의 Voice Activity Detection (VAD)를 이용하지 않았다. 모델의 훈련 단계에서는 랜덤으로 선택된 3초의 고정된 길이를 갖는 음성으로부터 추출된 특징 벡터열이 입력으로 이용되며, 등록 및 테스트 단계에서는 주어진 전체 음성이 입력으로 이용된다.

ResNet 기반 특징 추출기의 구조 및 전체적인 훈련 방식은 Reference [7]과 동일하다. 베이스라인 모델은 Reference [7]의 2D-Res34-Fbank64를 선택하였는데, 이는 34개 레이어의 ResNet 기반 특징 추출기로 구성된다. 그리고 GAP가 이용되며, 훈련 시 교차 엔트로피 비용 함수가 이용된다. 화자 검증 태스크의 성능 평가에는 동일 오류율(Equal Error Rate, EER)이 이용된다.

3.3 결과

먼저 본 연구에서 제안하는 방식의 효용성을 보이기 위하여 ablation study를 수행하였다(Table 1). (a)는 설정된 그룹의 개수에 따른 성능 비교를 보여준다. 베이스라인이 4.63 %의 EER을 보이는데, 그룹 개수에 상관없이 제안하는 방식이 더 좋은 성능을 보인다는 것을 확인할 수 있다. 그리고 그룹의 개수가 64개일 때 가장 좋은 성능을 보이며, 오히려 128개로 증가했을 때는 성능이 감소한다는 것을 확인할 수 있

Table 1. Ablation results of the proposed method (EER %). Grs : Groups, Naive : naive labeling, SDL : self-distributed labeling, Add. : addition, Concat. : concatenation.

(a) Number of Groups

System	EER (%)
Baseline	4.63
32 Grs	3.96
64 Grs	**3.88**
128 Grs	4.18

(b) Learning for GDN

System	EER (%)
Baseline	4.63
w/o Loss	4.22
Naive	3.97
SDL	**3.88**

(c) Add. Vs. Concat.

System	EER (%)
Baseline	4.63
Add.	**3.88**
Concat.	4.14

(d) Last activation for GDN

System	EER (%)
Baseline	4.63
Sigmoid	**3.88**
Softmax	5.47

다. (b)는 그룹 결정 네트워크(Group Decision Network, GDN)의 학습에 사용되는 비용함수에 따른 성능 비교를 보여준다. 두 번째 줄의 "w/o Loss"는 그룹 결정 네트워크에 별도의 비용함수 없이 L_{cl}만을 이용하여 훈련을 한 경우를 의미한다. 즉, Eq. (4)에서 λ가 0으로 설정된 상황이다. 세번째 줄의 Naive는 Reference [7]의 naive labeling을 의미하며, 네번째 줄의 SDL은 self-distributed labeling을 의미한다. SDL을 사용하였을 때 가장 좋은 성능을 보임을 확인할 수 있다. (c)의 Add.는 $\bar{z}$를 구할 때 Eq. (2)에서와 같이 z^G와 z 합으로 구하는 경우를 뜻한다. Concat.은 z^G와 z의 연결(concatenation)을 통해 $\bar{z}$를 구하는 경우를 나타내는데, 이를 사용하였을 때보다 Add.를 사용했을 때 더 좋은 성능을 보임을 확인할 수 있다. (d)는 그룹 결정 네트워크의 마지막 레이어의 activation 함수를 sigmoid로 설정한 경우와 softmax로 설정한 경우의 성능 비교를 보여준다. Sigmoid를 이용함으로써 각각의 그룹의 가중치가 0에서 1사이의 확률 값을 갖도록 할 때(즉, sum-to-one 제약 조건이 없도록 할 때) 더 좋은 성능을 나타낸다는 것을 알 수 있다.

Table 2는 본 연구에서 제안한 방식과 다른 연구에서 제안된 방식들 간의 성능 비교를 보여준다. 첫번째 줄은 전통적인 *i*-vector/ PLDA 방식의 성능을 보여주며, 두번째 줄부터 네번째 줄까지는 다른 연구에서 제안된 심층 화자 임베딩 방식들의 성능을 보여준다. 이 경우, 괄호안의 첫번째 항은 훈련에 사용한 비용 함수를 나타내며, 두번째 항은 사용된 전역 풀링 기법을 나타낸다. 예를 들어 Proposed(SM+GAP)는 제안한 방식에서 softmax 비용 함수를 이용하며, GAP 기반의 전역 풀링을 사용했음을 나타낸다. Proposed (SM+GAP)는 3.88 %의 EER을 보이며, 이는 Reference [18]에서 제안된 TDNN 기반의 방식을 제외한 나머지 방식들보다 더 좋은 성능을 보인다. Reference [18]의 TDNN 기반 방식은 데이터 증강 기법이 적용되었으므로, 본 연구에서 제안한 방식보다 더 많은 훈련 데이터를 이용하여 훈련된 것이다. 그럼에도 불구하고 Proposed(SM+GAP)는 이와 비슷한 성능을 보인다. 제안한 방식에서 A-Softmax 비용 함수를 이용한 경우, 3.70 %의 EER을 달성하여 가장 높은 성능을 보임을 알 수 있다. A-Softmax 비용 함수[15]는 화자 임베딩 벡터 사이의 각도에 마진을 둠으로써, 화자 임베딩 벡터 사이의 각도를 넓혀주는 역할을 한다. 따라서 기존의 softmax 비용 함수를 이용할 때에 비해 화자 변별력이 증가하게 된다.

Table 2. Performance comparison with state-of-the-art systems in terms of EER (%). ASM : A-Softmax, SAP : self-attentive pooling, SPE : spatial pyramid encoding, SM : softmax, ASP : attentive statistics pooling.

System	EER (%)
i-vector / PLDA[11]	8.8
ResNet34 (ASM+SAP)[12]	4.40
ResNet34 (ASM+SPE)[13]	4.03
TDNN (SM+ASP)[18]	3.85
Proposed (SM+GAP)	3.88
Proposed (ASM+GAP)	**3.70**

IV. 결 론

본 연구에서는 문장 독립 화자 검증 분야에서 가장 많이 이용되고 있는 심층 화자 임베딩 방식을 발전시키기 위해, 그룹기반 화자 임베딩 방식을 제안하였다. 기존의 심층 화자 임베딩에 화자의 그룹 정보를 담은 그룹 임베딩 벡터를 도입함으로써, 화자 임베딩이 나타낼 수 있는 전체 화자의 검색 공간을 줄이고, 이를 통해 기존의 심층 화자 임베딩 방식을 향상시킨다. Ablation study를 통해 제안한 방식의 효

용성을 확인하였으며, 추가로 A-Softmax 기반의 비용 함수를 이용함으로써 기존에 제안된 방식들보다 더 높은 성능을 달성하였다. 향후 연구에서는 단순한 GAP 방식이 아닌 더욱 발전된 형태의 전역 풀링 방식을 적용하여 추가적인 성능 향상을 이끌어내는 연구를 수행할 계획이다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A2C1014044).

References

1. J. H. L. Hansen and T. Hasan, “Speaker recognition by machines and humans: A tutorial review,” IEEE Signal Processing Magazine, **32**, 74-99 (2015).
2. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans on. Audio, Speech, and Lang. Process. **19**, 788-798 (2011).
3. S. Ioffe, “Probabilistic linear discriminant analysis,” Proc. ECCV. 531-542 (2006).
4. A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “I-vector based speaker recognition on short utterances,” Proc. Interspeech, 2341-2344 (2011).
5. A. Hajavi and A. Etemad, “A deep neural network for short-segment speaker recognition,” Proc. Interspeech, 2878-2882 (2019).
6. Y. Jung, S. M. Kye, Y. Choi, M. Jung, and H. Kim, “Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances,” Proc. Interspeech, 1501-1505 (2020).
7. Y. Jung, Y. Choi, H. Lim, and H. Kim, “A unified deep learning framework for short-duration speaker verification in adverse environments,” IEEE Access, **8**, 175448-175466 (2020).
8. V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” Proc. Interspeech, 3214-3218 (2015.)
9. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Proc. ICLR. 1-14 (2015).
10. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE CVPR. 770-778 (2016).
11. A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A largescale speaker identification dataset,” Proc. Interspeech, 2616-2620 (2017).
12. W. Cai, J. Chen, and M. Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” Proc. Odyssey, 74-81 (2018).
13. Y. Jung, Y. Kim, H. Lim, Y. Choi, and H. Kim, “Spatial pyramid encoding with convex length normalization for text-independent speaker verification,” Proc. Interspeech, 4030-4034 (2019).
14. E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” Proc. IEEE ICASSP. 4052-4056 (2014).
15. Z. Huang, S. Wang, and K. Yu, “Angular softmax for short-duration text-independent speaker verification,” Proc. Interspeech, 3623-3627 (2018).
16. Y. Liu, L. He, and J. Liu, “Large margin softmax loss for speaker verification,” Proc. Interspeech, 2873-2877 (2019).
17. Y. Kim, W. Park, M-C. Roh, and J. Shin, “Groupface: learning latent groups and constructing group-based representations for face recognition,” Proc. IEEE CVPR. 5621-5630 (2020).
18. K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” Proc. Interspeech, 2252-2256 (2018).

저자 약력

▸ 정 영 문 (Youngmoon Jung)

2016년 2월 : 서강대학교 전자공학부학사
2016년 3월 ~ 현재 : KAIST 전기및전자공학부 석박사 통합과정

▸ 엄 영 식 (Youngsik Eom)

2021년 2월 : 성균관대 전자전기공학부 학사
2021년 3월 ~ 현재 : KAIST 전기및전자공학부 석사 과정

▸ 이 영 현 (Yeonghyeon Lee)

2021년 2월 : 울산과학기술원 전자공학과 학사
2021년 3월 ~ 현재 : KAIST 전기및전자공학부 석사 과정

▸ 김 회 린 (Hoirin Kim)

1984년 2월 : 한양대학교 전자공학과학사
1987년 2월 : KAIST 전기및전자공학부 석사
1992년 2월 : KAIST 전기및전자공학부 박사
1987년 10월 ~ 1999년 12월 : ETRI 선임연구원
2000년 1월 ~ 현재 : KAIST 전기및전자공학부 교수