한국음향학회지 제37권 제6호 pp. 483~488 (2018)
The Journal of the Acoustical Society of Korea Vol.37, No.6 (2018)
https://doi.org/10.7776/ASK.2018.37.6.483

pISSN : 1225-4428
eISSN : 2287-3775

교사 학생 심층신경망을 활용한 다채널 원거리 화자 인증

Multi channel far field speaker verification using teacher student deep neural networks

정지원,[1] 허희수,[1] 심혜진,[1] 유하진[1†]

(Jee-weon Jung,[1] Hee-Soo Heo,[1] Hye-jin Shim,[1] and Ha-Jin Yu[1†])

[1]서울시립대학교 컴퓨터과학과

(Received September 18, 2018; accepted November 22, 2018)

초 록: 원거리 발성은 화자 인증 시스템의 성능을 하락시키는 주요 요인으로 알려져 있다. 본 논문에서는 교사 학생 학습을 이용하여 원거리 발성에 의한 화자 인증 시스템의 성능 하락을 보상하는 기법을 제안한다. 교사 학생 학습은 미리 학습된 교사 심층신경망의 출력과 학생 신경망의 출력이 같아지도록 학생 신경망을 학습하는 기법이다. 여기서 교사 신경망에는 근거리 발성을, 학생 신경망에는 원거리 발성을 입력한 뒤, 두 신경망의 출력을 동일하게 만드는 과정을 통해 원거리 발성을 보상할 수 있을 것이라고 기대하였다. 하지만 원거리 발성을 보상하는 과정에서, 근거리 발성에 대한 인식률이 저하되는 현상을 실험적으로 발견하였다. 위와 같은 현상을 예방하기 위해 본 논문에서는 교사 심층신경망을 학생 심층신경망의 초깃값으로 사용하는 기법과 학생 심층신경망을 근거리 발성에 대해서도 학습하는 기법을 제안하였다. 모든 실험은 원 음성을 입력 받는 심층신경망을 활용해 수행하였다. 동일한 발성을 각각 4 채널로 근거리와 원거리에서 자체적으로 수집한 문장 종속 데이터셋을 활용하였다. 동일 오류율을 기준으로 근거리 / 원거리 발성에 대한 화자 인증 성능을 평가한 결과 교사 학생 학습을 사용하지 않을 경우 2.55 % / 2.8 %, 기존의 교사 학생 학습을 사용할 경우 9.75 % / 1.8 %, 제안한 기법들을 적용한 경우 2.5 % / 2.7 %의 오류율을 확인하였다.

핵심용어: 교사 학생 학습, 심층신경망, 원거리 화자 인증, 다채널 화자 인증

ABSTRACT: Far field input utterance is one of the major causes of performance degradation of speaker verification systems. In this study, we used teacher student learning framework to compensate for the performance degradation caused by far field utterances. Teacher student learning refers to training the student deep neural network in possible performance degradation condition using the teacher deep neural network trained without such condition. In this study, we use the teacher network trained with near distance utterances to train the student network with far distance utterances. However, through experiments, it was found that performance of near distance utterances were deteriorated. To avoid such phenomenon, we proposed techniques that use trained teacher network as initialization of student network and training the student network using both near and far field utterances. Experiments were conducted using deep neural networks that input raw waveforms of 4-channel utterances recorded in both near and far distance. Results show the equal error rate of near and far-field utterances respectively, 2.55 % / 2.8 % without teacher student learning, 9.75 % / 1.8 % for conventional teacher student learning, and 2.5 % / 2.7 % with proposed techniques.

Keywords: Teacher student learning, Deep neural networks, Far-distance speaker verification, Multi channel speaker verification

PACS numbers: 43.60.Bf, 43.72.Bs

I. 서 론

원거리 화자 인증은 원거리에서 발화된 발성 (원거리 발성)으로 수행되는 화자 인증을 의미한다. 발

†**Corresponding author:** Ha-Jin Yu (hjyu@uos.ac.kr)
School of Cmputer Science, College of Engineering, University of Seoul, 163 Siripdae-ro, Dongdaemun-gu, Seoul 02504, Republic of Korea
(Tel: 82-2-6490-5697, Fax: 82-2-6490-2444)

"이 논문은 2018년도 한국음향학회 음성통신 및 신호처리 학술대회에서 발표하였던 논문임."

성이 원거리에서 입력될 경우 신호 세기의 감소, 잔향 및 잡음의 간섭, 롬바드 현상으로 인해 발성에 포함된 화자 정보가 손실될 수 있다. 이는 화자 인증 시스템의 성능을 저하시키는 주요 요인 중 하나이다.

원거리 발성에 의한 성능 저하를 보상하기 위해 개발된 다양한 기술들이 존재한다. 해당 기술들은 크게 별도의 과정을 통해 원거리 신호를 보상한 뒤 화자 인증을 수행하는 기술들과,[1,2] 원거리 발성을 보상하며 화자 인증을 수행하는 기술들로[3,4] 구분된다.

본 논문에서는 최근 여러 분야에서 최고 성능을 보이는 심층신경망을 이용해 발성의 거리에 상관없이, 입력된 발성을 동일한 화자 특징 공간으로 사상함으로써 원거리 발성에 의한 성능 저하 요소들을 보상하는 연구를 기술한다. 이를 위해 교사 심층신경망이 학생 심층신경망의 학습을 지도하는 교사 학생 학습을 이용하였다.[5] 구체적으로, 먼저 근거리, 원거리 발성을 모두 이용하여 교사 심층신경망을 학습하였다. 이때 교사 심층신경망은 각 발성의 화자 라벨을 사용해 계산한 크로스 엔트로피를 목적함수로 하여 학습하였다. 학습이 완료된 교사 심층신경망은 학생 심층신경망의 학습에 사용되는 유사 정답 라벨(soft label)을 제공하는 용도로 사용되었다. 학생 심층신경망의 출력과 교사 심층신경망의 유사 정답 라벨과의 쿨백-라이블러 발산(Kullback-Leibler divergence)을 목적함수로 이용하여 학생 심층신경망을 학습하였다.

본 논문의 II장에서는 교사 학생 학습의 개념 및 본 논문에서 소개하는 학생 심층신경망의 초기화 기법, 교사 학생 학습에서의 교사 심층신경망 학습 데이터 선택에 대한 내용을 소개한다. III장은 원거리 화자 인증 및 교사 학생 학습에 사용된 심층신경망과 화자 인증 과정을 설명한다. IV장은 제안한 교사 학생 학습 기법들을 이용한 원거리 화자 인증 실험의 설계 및 실험 결과의 분석을 다루며, 마지막으로 V장에서는 결론 및 향후 연구 계획을 보인다.

II. 교사 학생 학습

2.1 기존의 교사 학생 학습

교사 학생 학습은 특정 심층신경망의 연산을 상대적으로 적은 규모의 심층신경망에 임베딩시키는 모델 압축을 위해 처음 제안되었다.[5] 해당 학습 기법은 사전에 학습된 교사 심층신경망이 학생 심층신경망의 학습에 사용되는 유사 정답 라벨을 제공하는 방식을 취한다. 이때, 학생 심층신경망의 목적함수는 다음과 같이 정의된다.

$$KL_{loss} = \sum_{j=1}^{J}\sum_{i=1}^{I} p_T(s_i|x_j)\log\left(\frac{p_T(s_i|x_j)}{p_S(s_i|x_j)}\right), \quad (1)$$

여기서 $p_T(s|x)$와 $p_S(s|x)$는 각각 교사와 학생 심층신경망의 posterior 분포를 나타내며, $p_T(s_i|x_j)$와 $p_S(s_i|x_j)$는 발성 x_j가 입력되었을 때 각각 교사와 학생 심층신경망 출력층의 i번째 노드 값을 가리킨다.

Eq. (1)에서 $p_T(s_i|x_j)\log(p_T(s_i|x_j))$ 항은 학생 심층신경망 학습에 영향을 끼치지 않는다.[5] 따라서 교사 학생 학습에서의 목적함수는 다음과 같이 교사 심층신경망을 통해 생성한 유사 정답 라벨을 이용한 크로스 엔트로피 학습과 동일하다.

$$KL_{loss} = -\sum_{j=1}^{J}\sum_{i=1}^{I} p_T(s_i|x_j)\log(p_S(s_i|x_j)). \quad (2)$$

기존의 원-핫 벡터(one-hot vector)를 이용한 학습에서는 정답 라벨은 1 값을, 기타 노드들은 0 값을 가지게 되지만, 이와 달리 유사 정답 라벨은 정답 라벨을 가리키는 노드의 값이 1보다 작아질 뿐만 아니라 기타 노드들도 값이 0보다 약간 커지게 된다. 이는 심층신경망의 일반화 성능을 높이는 것으로 알려져 있다.[3]

원-핫 벡터를 이용한 심층신경망 학습에는 정답 라벨이 주어진 데이터만을 사용할 수 있다는 한계가 있다. 하지만 교사 학생 학습에서는 교사 심층신경망의 학습이 완료된 이후, 정답 라벨이 없는 데이터를 이용해 학생 심층신경망을 추가로 학습시킬 수 있다. 이러한 특성을 활용해 정답 라벨이 없는 추가 데이터를 활용한 추가 학습을 진행하고, 이를 통해 학생 심층신경망의 성능을 추가로 향상시키는 연구들이 보고되었다.[5]

본 논문에서는 교사 학생 학습을 이용해 원거리 발성 입력에 따른 화자 인증 시스템의 성능 하락을

보상하는 연구를 기술하였다. 이를 위해, 근거리 및 원거리 발성을 이용해 교사 심층신경망을 먼저 학습한다. 교사 학생 학습이 모델 압축에 사용될 경우, 서로 다른 크기의 교사, 학생 심층신경망에 동일한 발성을 입력하게 된다. 이와 달리, 교사 학생 학습을 원거리 발성 보상에 사용할 경우 교사 심층신경망에는 동일 발화에 대한 근거리 발성을, 학생 심층신경망에는 원거리 발성을 입력해 학생 심층신경망을 학습하게 된다. 이때, 교사 심층신경망과 학생 심층신경망은 동일한 구조와 크기를 가지는 모델이며, 학생 심층신경망을 학습하는 과정에서 교사 심층신경망은 더 이상 학습되지 않고 고정된다. 이러한 구조에서 학생 심층신경망의 목적함수는 아래와 같이 정의된다.

$$KL_{loss} = -\sum_{j=1}^{J}\sum_{i=1}^{I} p_T(s_i|x_{j,n})\log(p_S(s_i|x_{j,f})), \qquad (3)$$

여기서 $x_{j,n}$과 $x_{j,f}$는 동일 발성 x_j를 각각 근거리와 원거리에서 녹음한 발성을 가리킨다. 위와 같이 정의한 목적함수가 0으로 수렴할 경우 원거리에서 녹음한 발성에 존재하는 잔향, 잡음 등을 학생 심층신경망이 내부적으로 보상하여 교사 심층신경망에 근거리에서 녹음한 발성이 입력되었을 때와 동일한 출력값을 보이게 된다. Fig. 1은 원거리 보상을 위한 교사 학생 학습 구조 전반을 나타낸다.

2.2 학생 심층신경망의 근거리 발성에 대한 성능 향상

제안한 교사 학생 학습을 이용한 원거리 발성 보상에서는 학생 심층신경망을 학습시킬 때 1. 학생 심층신경망을 임의의 값들로 초기화하며, 2. 교사 심층신경망에는 근거리 발성을, 학생 심층신경망에는 원거리 발성을 입력하는 방식으로 학습을 진행한다. 하지만 이러한 방식을 통해 학습된 학생 신경망에서 추출한 화자 특징은 원거리 발성은 잘 표현하는 반면 근거리 발성을 잘 표현하지 못할 수 있다. IV장의 실험을 통해 위의 현상이 실제로 발생함을 확인하였다. 학생 신경망의 근거리 발성에 대한 성능을 향상시키기 위해 본 논문에서는 세 가지 기법을 제안한다: 1. 교사 심층신경망을 학생 심층신경망의 초기 상태로 사용, 2. 교사 심층신경망 학습에 원거리 발성 사용, 3. 학생 심층신경망 학습에 근거리 발성 사용.

본 논문에서는 학습이 완료된 교사 심층신경망을 학생 심층신경망의 초기값으로 사용하는 것이 임의 초기화 방식에 비해 학생 심층신경망 학습에 비교적 용이할 것이라고 가정하였다. 임의 초기화 방식의 경우, 학생 심층신경망이 화자를 식별하는 연산과 더불어 원거리 발성을 보상하는 연산을 동시에 학습해야 한다. 반면 교사 심층신경망을 학생 심층신경망의 초기값으로 사용하게 되면 이미 화자 식별을 수행할 수 있는 상태이므로, 원거리 발성을 보상하는 연산만 학습하면 된다.

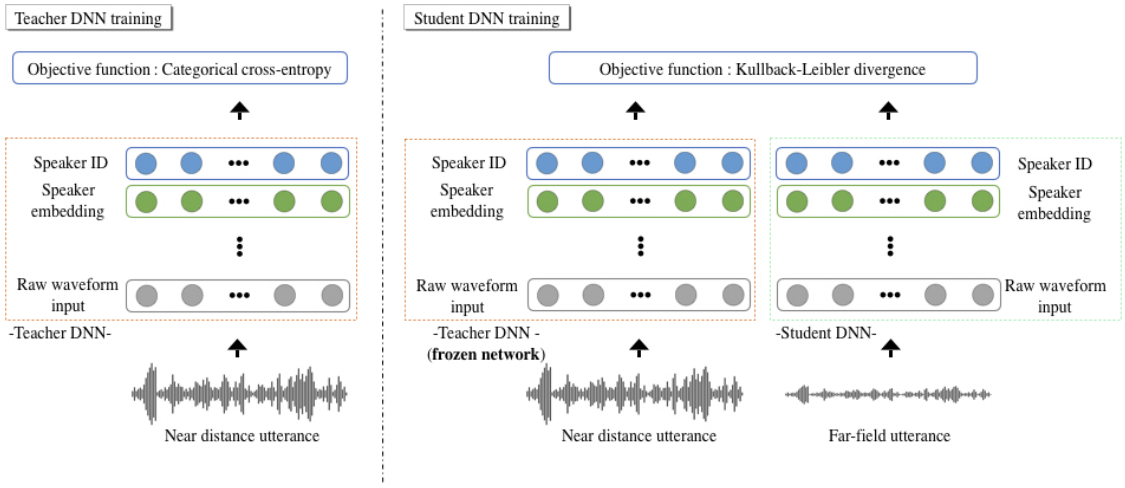

Fig. 1. General pipeline process of teacher student learning framework for far-field speaker verification.

근거리 발성만을 이용해 교사 심층신경망을 학습하게 되면 심층신경망에서 추출한 화자 특징 공간이 근거리의 발성들을 나타내는 것에만 적합할 수 있다. 본 논문에서는 교사 심층신경망 학습 시 근거리와 원거리에서 입력된 발성을 모두 사용함으로써 발성이 입력된 거리에 관계없이 화자를 나타낼 수 있는 화자 특징을 추출할 수 있다고 가정하였다.

기존의 교사 학생 학습에서는 학생 심층신경망이 원거리 발성에 대해서만 학습되기 때문에 근거리 발성에 대한 성능이 보장되지 않는다. 제안하는 교사 심층신경망을 활용한 초기화 기법을 통해 학생 심층신경망의 근거리 발성에 대한 성능이 동일 오류율 기준 10.5 %에서 9.75 %로 비교적 향상됨을 IV장의 실험 결과를 통해 확인하였다. 하지만 실험 결과, 여전히 학생 심층신경망의 근거리 발성에 대한 동일 오류율(9.75 %)과 원거리 발성에 대한 동일 오류율(1.8 %) 간의 차이가 크게 존재하였다. 이에 본 논문에서는 학생 심층신경망을 근거리 발성에 대해서도 학습시키는 대안을 제안한다. 제안하는 방식은 교사와 학생 심층신경망 양쪽에 모두 근거리에서 발화된 발성을 입력하여 학생 심층신경망을 학습한다. 제안한 세 기법에 대한 실험 결과는 IV장에 제시되어 있다.

III. 심층신경망 및 화자 인증 시스템

3.1 원 신호를 이용하는 심층신경망

본 논문에서는 교사 학생 학습을 이용해 원거리에서 입력된 발성을 보상하는 연구를 위해 Jung *et al.*[6]이 제안한 잔차 연결을 적용한 원 신호 기반 합성곱 신경망(Raw Waveform-based Convolutional Neural Network with residual connection, RWCNN-residual)을 사용하였다. 사용된 심층신경망은 기존의 음향 특징 추출 및 기타 전처리 과정을 거치지 않고 원 신호로부터 화자 특징을 추출한다는 특성을 지닌다. 해당 심층신경망은 합성곱 은닉층 위주로 구성되어 있으며, 잔차 연결(residual connection)과 배치 정규화(batch normalization) 기법을 적용하였다.[7,8] Fig. 2의 online phase는 본 논문에서 화자 인증 실험에 사용한 심층신경망의 구조를 보여준다.

3.2 화자 인증 시스템

본 논문에서는 심층신경망의 학습이 완료된 이후, 발성이 입력되었을 때 마지막 은닉층 rectified unit 활성화 함수의 출력값을 화자 특징으로 사용한다. 해당 방식을 이용해 등록 발성과 평가 발성으로부터 각각 화자 특징을 추출한다. 다수의 등록 발성들이 존재하는 경우 화자별로 계산한 화자 특징의 평균값을 화자 모델로 취한다. 화자 모델과 평가 발성의 화자 특징 간의 코사인 유사도를 계산한 뒤, 이를 기준으로 화자 인증을 수행한다. 코사인 유사도는 두 벡터 a와 b가 주어졌을 때 $\frac{a \times b}{\|a\| \times \|b\|}$ 로 정의된다. Fig. 2의 online phase에 전체 화자 인증 시스템의 구성이 소개되어 있다.

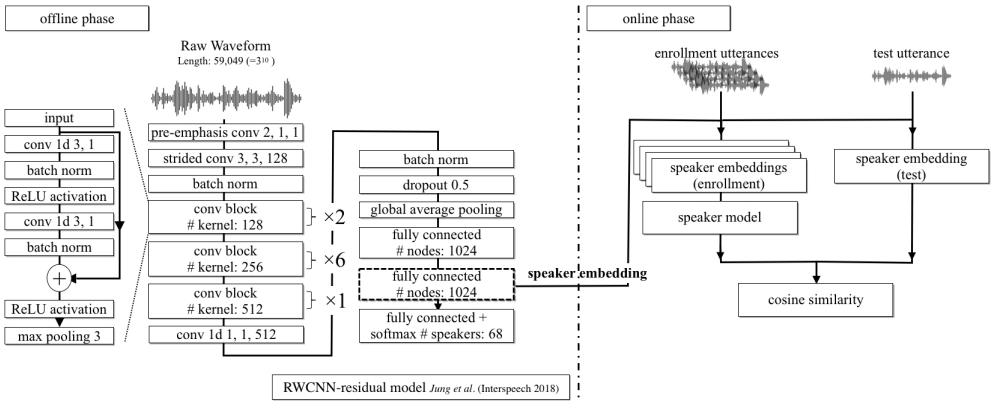

Fig. 2. Illustration of RWCNN-residual model (offline phase, Jung *et al.*, Interspeech 2018[9]) and overall speaker verification pipeline (online phase). Three numbers next to convolutions each refer to the length of kernel, stride size, and the number of kernels.

IV. 실험 설계 및 결과

4.1 데이터셋

본 논문에서는 원거리 환경에서의 화자 인증 실험을 위해 직접 수집한 데이터셋을 사용하였다. 가로 3.4 m, 세로 5 m의 방에서 근거리에 4개, 2.5 m 거리에 4개의 마이크를 동시에 활용해 8 채널로 구성된 음성을 수집하였다. 이 중, 근거리에 위치한 4개 마이크에서 수집한 음성을 근거리 발성, 원거리에 위치한 4개 마이크에서 수집한 음성을 원거리 음성으로 활용하였다. 따라서 근거리 발성과 원거리 발성은 4 채널 음성으로 구성된다. 화자 수는 총 78 명이며 화자별로 약 250개의 발성을 각각 근거리와 원거리에서 수집하였다. 이 중 68 화자의 발성을 심층신경망 학습 및 검증을 위해 사용하였고, 나머지 10 화자의 발성을 이용해 동일 오류율(Equal Error Rate, EER)을 기준으로 화자 인증 성능 평가를 진행하였다.

4.2 심층신경망 설계

본 논문의 실험에 사용된 심층신경망은 약 3.59 s 길이의 16 kHz 원 음성 신호(59,049개의 샘플로 구성)를 입력한다. 심층신경망은 Jung *et al*.[6]에서 사용된 RWCNN-residual 모델을 추가적인 변형 없이 사용하였다. Optimizer는 stochastic gradient descent를 learning rate 0.001, momentum 0.9의 값을 주어 사용하였다. 학습이 완료된 심층신경망의 마지막 은닉층 rectified unit 활성화 함수의 출력을 화자 특징으로 사용하였다.

4.3 결과 분석

본 논문의 실험 결과들은 Table 1에 나타내었다. Table 1의 각 행은 학습 방식에 따른 근거리, 원거리 발성에 대한 화자 인증 성능을 나타낸다. baseline1과 baseline 2는 각각 교사 학생 학습을 적용하지 않은 심층신경망과, 기존의 교사 학생 학습을 적용한 심층신경망을 가리킨다. 두 시스템의 비교를 통해 기존의 교사 학생 학습을 수행할 경우, 원거리 발성에 대한 성능은 향상되지만, 근거리 발성에 대한 성능이 하락함을 확인할 수 있다.

baseline 1과 baseline 1 + w far-field 시스템의 비교를 통해 교사 심층신경망을 학습시킬 때 원거리 발성을 활용하는 기법에 따른 성능 향상을 확인할 수 있었다. baseline 2와 baseline 2 + teacher init 두 시스템의 비교는 교사 심층신경망을 학생 심층신경망의 초깃값으로 활용하는 기법에 따른 성능 향상을 관측할 수 있었다.

Table 1. EER of the baseline and proposed teacher student based systems (near / far field evaluation). 'ts' means teacher student learning, 'teacher init' refers to initializing the student network using learned teacher network, and 'student w near' refers to using near-field utterances for student training as well.

Model	EER (%, near / far)
baseline 1 (w/o t-s)	3.2 / 9.75
baseline 1 + w far	2.55 / 2.8
baseline 2 (w t-s)	10.5 / 2.65
baseline 2 + teacher init	9.75 / **1.8**
baseline 2 + teacher w far+ teacher init + student w near	**2.5** / 2.7

baseline 2와 baseline 2 + teacher init 두 시스템 모두 원거리 발성에 대한 성능은 향상하지만 근거리 발성에 대한 성능이 오히려 감소함을 확인할 수 있었다. 본 논문에서 제안한 기법들을 모두 적용한 결과, 근거리 발성과 원거리 발성에 대한 동일 오류율이 각각 2.5 %, 2.7 %로서 교사 학생 학습에서의 근거리 발성에 대한 성능 저하 현상을 예방할 수 있었다.

V. 결 론

본 논문에서는 교사 학생 학습을 이용하여 다채널 문장 종속 상황에서의 원거리 화자 인증 성능을 향상시키는 방안을 소개하였다. 교사 학생 학습은 미리 학습시킨 교사 심층신경망을 이용하여 학생 심층신경망을 학습시키는 기법으로서 원거리 발성의 보상에 적용될 수 있음을 실험으로 확인하였다. 이때, 교사 학생 학습을 이용해 원거리 발성에 대한 보상을 수행할 경우 학습된 학생 심층신경망의 근거리 발성에 대한 성능이 감소하는 현상을 확인하였다. 따라서, 교사 심층신경망을 활용한 학생 심층신경망 초기화와 교사 학생 학습 시 근거리 발성에 대해서도 학습을 진행하는 기법을 통해 근거리 발성과 원

거리 발성에 대한 성능 차이를 줄일 수 있었다. 제안하는 기법들을 적용한 학생 심층신경망은 근거리와 원거리 평가 세트에 대해 각각 2.5 %, 2.7 %의 동일 오류율을 보였다.

감사의 글

이 논문은 2017년도 서울시립대학교 연구년교수 연구비에 의하여 연구되었음.

References

1. M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications* (Springer Science & Media , Heidelberg, 2013), pp. 39-60.
2. J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection" IEEE signal processing letters, **6,** 1-3 (1999).
3. J. Li, R. Zhao, Z. Chen, C. Liu, X. Xiao, G. Ye, and Y. Gong, "Developing Far-Field Speaker System via teacher student Learning," Proc. ICASSP, 5699-5703 (2018).
4. M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Batch-normalized joint training for DNN-based distant speech recognition," Proc. SLT workshop, 28-34 (2016).
5. J. Li, R. Zhao, J. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," Proc. Interspeech, 1910-1914 (2014).
6. J. Jung, H. Heo, Y. Yang, H. Shim, and H. Yu, "A complete end-to-end speaker verification system using deep neural networks: from raw signals to verification result," Proc. ICASSP, 5349-5353 (2018).
7. H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Identity mappings in deep residual networks," Proc. ECCV, 30-645 (2016).
8. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," Proc. ICML, 448-456 (2015).
9. J. Jung, H. Heo, Y. Yang, H. Shim, and H. Yu, "Avoiding speaker overfitting in End-to-End DNNs using raw waveform for text-independent speaker verification" Proc. Interspeech, 3583-3587 (2018).

저자 약력

▸ **정 지 원 (Jee-weon Jung)**

2017년 2월: 서울시립대학교 컴퓨터과학부 학사
2017년 ~ 현재: 서울시립대학교 컴퓨터과학과 석 · 박사통합과정

▸ **허 희 수 (Hee-Soo Heo)**

2013년 2월: 서울시립대학교 컴퓨터과학부 학사
2013년 ~현재: 서울시립대학교 컴퓨터 과학부 석 · 박사통합과정

▸ **심 혜 진 (Hye-jin Shim)**

2017년 2월: 광운대학교 동북아통상학부 학사
2017년 ~ 현재: 서울시립대학교 컴퓨터과학과 석사과정

▸ **유 하 진 (Ha-Jin Yu)**

1990년 2월: KAIST 전산학과 학사
1992년 2월: KAIST 전산학과 석사
1997년 2월: KAIST 전산학과 박사
1997년 ~ 2000년: LG 전자 전자기술원 선임연구원
2000년 ~ 2002년: SL2(주) 연구소장
2002년 ~ 현재: 서울시립대학교 컴퓨터과학부 교수