한국음향학회지 제39권 제2호 pp. 137~142 (2020)
The Journal of the Acoustical Society of Korea Vol. 39, No. 2 (2020)
https://doi.org/10.7776/ASK.2020.39.2.137

pISSN : 1225-4428
eISSN : 2287-3775

Attention-long short term memory 기반의 화자 임베딩과 I-vector를 결합한 원거리 및 잡음 환경에서의 화자 검증 알고리즘

Speaker verification system combining attention-long short term memory based speaker embedding and I-vector in far-field and noisy environments

배아라[1], 김우일[1†]

(Ara Bae[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부

(Received January 21, 2020; revised February 27, 2020; accepted March 20, 2020)

초 록: 문장 종속 짧은 발화에서 문장 독립 긴 발화까지 다양한 환경에서 I-vector 특징에 기반을 둔 많은 연구가 수행되었다. 본 논문에서는 원거리 잡음 환경에서 녹음한 데이터에서 Probabilistic Linear Discriminant Analysis (PLDA)를 적용한 I-vector와 주의 집중 기법을 접목한 Long Short Term Memory(LSTM) 기반의 화자 임베딩을 추출하여 결합한 화자 검증 알고리즘을 소개한다. LSTM 모델의 Equal Error Rate(EER)이 15.52 %, Attention-LSTM 모델이 8.46 %로 7.06 % 성능이 향상되었다. 이로써 본 논문에서 제안한 기법이 임베딩을 휴리스틱 하게 정의하여 사용하는 기존 추출방법의 문제점을 해결할 수 있는 것을 확인하였다. PLDA를 적용한 I-vector의 EER이 6.18 %로 결합 전 가장 좋은 성능을 보였다. Attention-LSTM 기반 임베딩과 결합하였을 때 EER이 2.57 %로 기존보다 3.61 % 감소하여 상대적으로 58.41 % 성능이 향상되었다.

핵심용어: 화자 검증, 화자 임베딩, 주의 집중 기법, 원거리 잡음 환경

ABSTRACT: Many studies based on I-vector have been conducted in a variety of environments, from text-dependent short-utterance to text-independent long-utterance. In this paper, we propose a speaker verification system employing a combination of I-vector with Probabilistic Linear Discriminant Analysis (PLDA) and speaker embedding of Long Short Term Memory (LSTM) with attention mechanism in far-field and noisy environments. The LSTM model's Equal Error Rate (EER) is 15.52 % and the Attention-LSTM model is 8.46 %, improving by 7.06 %. We show that the proposed method solves the problem of the existing extraction process which defines embedding as a heuristic. The EER of the I-vector/PLDA without combining is 6.18 % that shows the best performance. And combined with attention-LSTM based embedding is 2.57 % that is 3.61 % less than the baseline system, and which improves performance by 58.41 %.

Keywords: Speaker verification, Speaker embedding, Attention mechanism, Far-field and noisy environments

PACS numbers: 43.72.Bs, 43.72.Ne

†**Corresponding author:** Wooil Kim (wikim@inu.ac.kr)
Department of Computer Science and Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon 22012, Republic of Korea
(Tel: 82-32-835-8459, Fax: 82-32-835-0780)

I. 서 론

음성에는 사람마다 고유한 특성이 존재하는데 이를 정보 분석을 통해 발화자를 식별하는 것을 화자 인식이라고 한다. 이러한 음성에 담긴 화자 특성을 이용하여 지문이나 홍채와 같이 생체보안기술로써 활용하기 위해 잡음이나 잔향이 있는 환경에서도 강인한 화자 인식시스템을 구축하는 것이 중요하다. 화자와 잡음 및 잔향 등의 채널 정보를 분리하기 위해 Joint Factor Analysis(JFA) 방법이 소개되었다.[1] 그러나 화자마다 다양한 채널 정보가 필요할 뿐만 아니라 분리되면서 중요한 화자 정보가 손실될 수 있는 문제점이 있다. 이러한 문제점을 보완하고자 화자와 채널 정보 등의 변이성을 모두 하나의 하위공간에 화자의 신원을 표현하는 I-vector가 등장하였다.[2,3] 또한, 심층 신경망이 등장하여 많은 분야에서 이를 사용한 모델들이 증가하였다. 화자 인식 분야에서는 I-vector를 대체하는 다양한 특징이 소개되었다. 대표적으로 완전 연결 Deep Neural Network(DNN)의 마지막 은닉층을 특징으로 사용하는 d-vector와 Time-Delay Neural Network(TDNN)의 마지막 은닉층을 통계적으로 풀링하여 특징으로 사용하는 x-vector가 있다.[4-8] 신경망 모델은 입력으로 Mel-Frequency Cepstral Coefficients (MFCC), mel-filter bank와 같은 음향특징을 사용하고, 화자 ID를 출력하도록 훈련된다. 최근 기계번역에서 고정된 크기의 벡터로 입력 문자열을 압축할 때 발생하는 정보 손실을 해결하기 위해 주의 집중 기법이 처음으로 도입되었다.[9] 번역되는 단어와 연관이 높은 입력문장의 단어에 집중해 번역하는 것이 주의 집중 기법의 접근 방법이다. 입력문장을 인코더에 넣어 디코더를 통해 새로운 문장으로 번역할 때 인코더의 은닉층 상태와 디코더의 은닉층 상태의 유사도를 계산한다. 이렇게 계산된 유사도를 인코더의 은닉층 상태와 가중 합하여 입력문장의 문맥을 가지면서 연관이 높은 단어 즉, 가중치가 높은 단어에 집중해서 컨텍스트 벡터를 생성한다. 기존 인코더의 은닉층 상태 대신 컨텍스트 벡터를 사용하면 입력문장의 가중치가 높은 단어에 따라 새로운 단어가 생성된다. 본 논문에서는 음향특징에서 발화수준 특징을 갖는 화자 임베딩을 생성할 때 화자의 정체성을 갖는 프레임에 높은 가중치를 두어 주의 집중 기법을 접목하였다.

II. I-vector/PLDA 기반 화자 인식시스템

현재까지 화자 인식에서 I-vector 특징은 널리 사용되고 있다. Gaussian Mixture Model-Universal Background Model(GMM-UBM)의 평균을 슈퍼벡터로 두어 충분통계량을 계산한 뒤 Total Variability(TV) 매트릭스를 학습한다. 이렇게 학습된 TV 공간에서 I-vector를 추출한다.[3] 본 논문에서는 MATLAB 환경에서 MFCC 특징으로 I-vector를 추출하였다. 화자 독립적인 요소를 분리하지 않은 공간에서 I-vector를 추출하였기 때문에 화자의 정보 외 다른 정보를 제거하기 위해 다양한 기법들이 소개되었다.[10,11] 본 논문에서는 대표적인 기법으로 많이 사용되는 LDA(Linear Discriminant Analysis)와 PLDA(Probabilistic Linear Discriminant Analysis)를 적용하여 성능을 비교하였다. LDA는 클

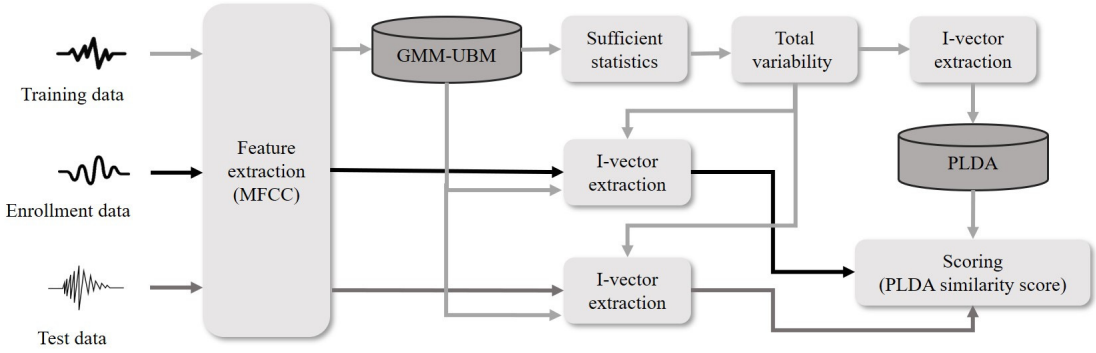

Fig. 1. I-vector system framework.

래스 정보를 사용해 클래스 내의 분산은 작게 하고, 클래스 간의 분산은 크게 하는 축을 찾아 데이터를 사영하여 서로 다른 클래스들을 분리한다. 등록된 화자(클래스 1)와 등록되지 않은 화자(클래스 2)를 분리하는데 LDA를 적용하였다. PLDA는 거리 기반으로 유사도를 측정하는 코사인 유사도 점수와 달리 확률모델 기반으로 서로 같은 화자 벡터로 구성된 확률과 다른 화자 벡터로 구성된 확률로 유사도를 측정한다.[4] 즉, LDA 적용 후 등록 I-vector와 테스트 I-vector가 같은 클래스에 있을 확률을 계산하여 유사도를 측정하였다. Fig. 1은 I-vector 시스템의 구조를 나타낸다.

III. Attention-LSTM 기반 화자 인식시스템

심층 신경망 기법에서는 Figs. 2, 3과 같이 사용자가 fully-connected DNN의 마지막 은닉층 혹은 마지막 프레임의 Long Short Term Memory(LSTM)출력 등 휴리스틱 하게 임베딩으로 사용할 부분을 정의하여 사용하였다.[4-8]

기존의 심층 신경망과 달리 LSTM에 주의 집중 기법을 접목한 Fig. 4 시스템에서 추출한 화자 임베딩은 화자의 특성과 높은 연관성이 있는 프레임에 높은 가중치를 주는 학습을 통해 추출함으로써 임의로 정의하던 문제점을 해결하였다. Eq. (1)에서 v, W, b는 학습되는 파라미터이며 각 프레임 t에서 LSTM의 출력 h와 연산하여 s를 계산한다. 비선형 활성화 함수를 거친 상수 s에 softmax를 사용해 Eq. (2)와 같이 가중치 α를 구할 수 있다.[9,12] v, W, b 파라미터는 LSTM 모델 학습에 사용하는 loss로 같이 훈련된다.

$$s_t = v_t^T \tanh(Wh_t + b). \tag{1}$$

$$\alpha_t = \frac{\exp(s_t)}{\sum_{t=1}^{T} \exp(s_t)}. \tag{2}$$

기존의 LSTM 모델에서는 h_T를 화자 임베딩으로 사용하였지만, attention을 접목한 LSTM 모델은 가중치 α와 LSTM의 출력 h를 가중 합하여 Eq. (3)과 같이 화자 임베딩을 추출하였다.

$$e = \sum_{t=1}^{T} \alpha_t h_t. \tag{3}$$

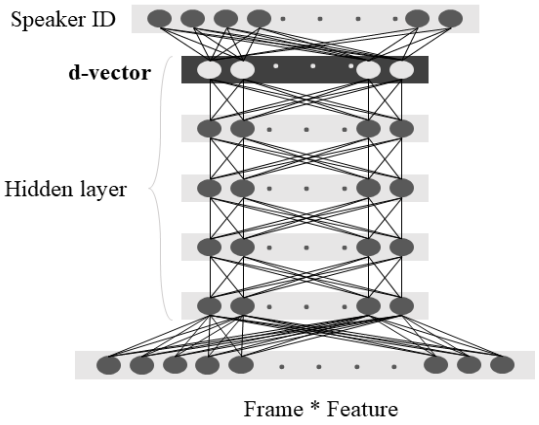

Fig. 2. Extracting process of DNN model based d-vector.

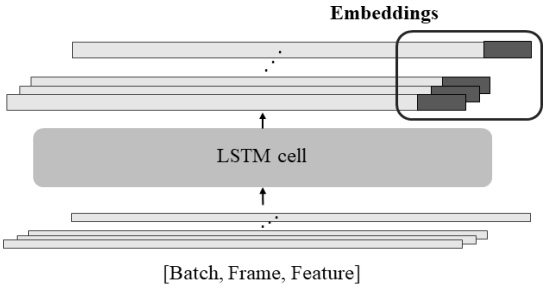

Fig. 3. Extracting process of LSTM model based speaker embedding.

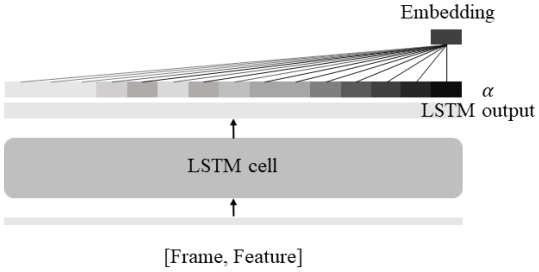

Fig. 4. Extracting process of attention-LSTM model based speaker embedding.

IV. 실험 및 결과

본 논문에서는 1000명의 화자가 깨끗한 환경에서

4음절의 짧은 단어를 3번씩 발화한 음성데이터를 에어컨 작동 조건의 실내 잡음 환경 조성을 위해 에어컨 동작 상태(off, cool, strong, middle, weak)별 바람 소리 잡음을 약 5 dB ~ 15 dB SNR 수준으로 재생하여 재녹음하였다. 증강된 데이터로 UBM과 TV 매트릭스, LDA, PLDA 모델, 신경망 훈련에 사용하였다.

테스트 데이터는 훈련 데이터와 같은 단어를 91명의 화자가 약 10번씩 발화하였다. 먼 거리 잡음 환경을 위해 깨끗한 환경에서 녹음된 음성을 마이크로부터 1 m, 3 m, 5 m 거리에서 정면 및 우측 위치에 재생하고, 훈련 데이터와 다른 에어컨을 사용한 바람 소리 잡음(wind)과 사람들의 웅성거리는 잡음(pub) 그리고 두 가지가 섞인 잡음(wNp)을 이용해 SNR 10 dB 수준으로 재녹음하였다.

4.1 I-vector 모델 아키텍처

MFCC 39차원 특징으로 512개의 component를 갖는 UBM 모델을 훈련하여 400차원의 TV 매트릭스를 생성하였다. 훈련 화자 1000명의 깨끗한 발화와 5가지의 잡음 환경을 포함한 총 18000개 발화의 클래스 정보로 LDA 매트릭스를 계산하였다. UBM과 TV 매트릭스로 I-vector를 추출한 뒤 LDA와 행렬 연산을 통해 350차원으로 축소하였다. LDA 적용 전과 후 I-vector를 코사인 유사도로 점수를 계산하여 Equal Error Rate(EER)로 성능을 비교하였다. LDA 적용 후 I-vector의 코사인 유사도와 PLDA로 점수를 계산하여 같은 지표로 성능을 비교하였다. LDA 적용 후 PLDA로 유사도 점수 계산하였을 때 가장 좋은 성능을 보인 것을 확인하였다.

4.2 Attention-LSTM 모델 아키텍처

Tensorflow 환경에서 log-mel-filter bank 40차원 특징으로 Attention-LSTM을 거쳐 임베딩을 생성하였다. 1000명 중 6명의 화자를 랜덤하게 선택한 후 18개의 발화를 사용하였으며, 각 발화의 100프레임을 선택하여 batch의 입력으로 사용하였다. 256개 노드를 갖는 은닉층과 128개 노드를 갖는 사영층으로 이루어진 레이어를 3개 쌓은 LSTM 셀의 출력과 attention 가중치를 가중 합하여 임베딩을 계산하였다. Batch

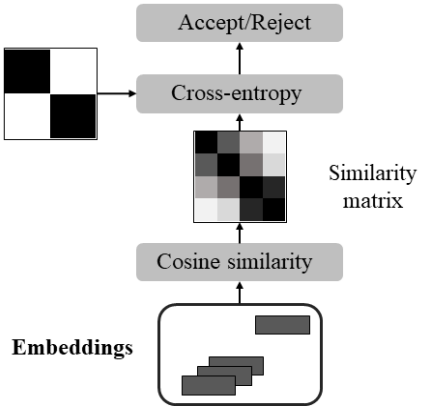

Fig. 5. Attention-LSTM model training framework.

Table 1. Attention-LSTM model architecture.

Layer	Output
Hidden 1	[100, 108, 256]
Projection	[100, 108, 128]
Hidden 2	[100, 108, 256]
Projection	[100, 108, 128]
Hidden 3	[100, 108, 256]
Projection (= LSTM output)	[100, 108, 128]
Attention (= Embedding)	[108, 128]
Similarity	[108, 108]

내에 각 화자의 임베딩으로 유사도 행렬을 계산하여 손실을 구하는 GE2E loss 방법으로 화자 구분을 학습하였다. 본 논문에서는 정답 행렬(같은 화자 발화는 1, 다른 화자 발화는 0 값을 갖는 행렬)과 cross-entropy를 사용하여 손실을 계산하였다.[13,14] Fig. 5는 Attention-LSTM 모델의 훈련 과정을 나타낸다. Table 1은 6명 화자의 18개 발화를 사용하였을 때 LSTM cell과 임베딩 및 유사도 행렬의 구조이다.

4.3 유사도 점수 기반 결합 시스템

400차원의 I-vector와 128차원의 임베딩을 결합하기 위해 유사도 점수를 구한 뒤 가중치를 통해 새로운 유사도 점수를 계산하여 사용하였다.

$$score_c = (1-\alpha) * score_i + \alpha * score_e. \quad (0 \le \alpha \le 1) \tag{4}$$

4.4 실험 결과

1 m, 3 m, 5 m의 거리에서 깨끗한 환경(off), pub 잡음(pub), 바람 잡음(wind), 그리고 pub과 바람이 섞인 잡음(wNp) 환경에서 마이크의 정면과 우측에서 녹음된 테스트 데이터로 성능 평가를 진행하였다.

Table 2. LSTM, Attention-LSTM, I-vector/PLDA and I-vector/PLDA combined with attention-LSTM based speaker verification performance evaluation (EER [%]).

EER [%]	LSTM	Attention-LSTM	I-vector/PLDA	Proposed method
off	14.58	8.21	6.37	2.93
pub	16.45	8.43	5.81	2.20
wind	15.76	8.72	6.00	2.55
wNp	15.30	8.49	6.61	2.60
AVG	15.52	8.46	6.18	2.57

Table 2는 LSTM 및 Attention-LSTM으로 추출한 화자 임베딩과 PLDA를 적용한 I-vector 성능을 각 잡음에 대해 거리와 위치 결과들을 평균 내 작성하였다. LSTM과 Attention-LSTM 시스템의 EER이 각각 15.52 %, 8.46 %이며, I-vector/PLDA 시스템의 EER이 6.18 %로 결합하기 전 가장 좋은 성능을 보였다. Attention-LSTM 기반 임베딩과 I-vector/PLDA를 결합하였을 때 EER이 3.61 % 감소한 2.57 %로 상대적으로 약 58.41 % 성능 향상을 보였다. Fig. 6은 Table 2의 EER 성능 그래프이다.

V. 결 론

본 논문에서는 주의 집중 기법과 GE2E loss를 활용한 LSTM 기반의 화자 임베딩과 I-vector/PLDA의 점수 결합을 제안하였다. LSTM 모델의 임베딩 보다 주

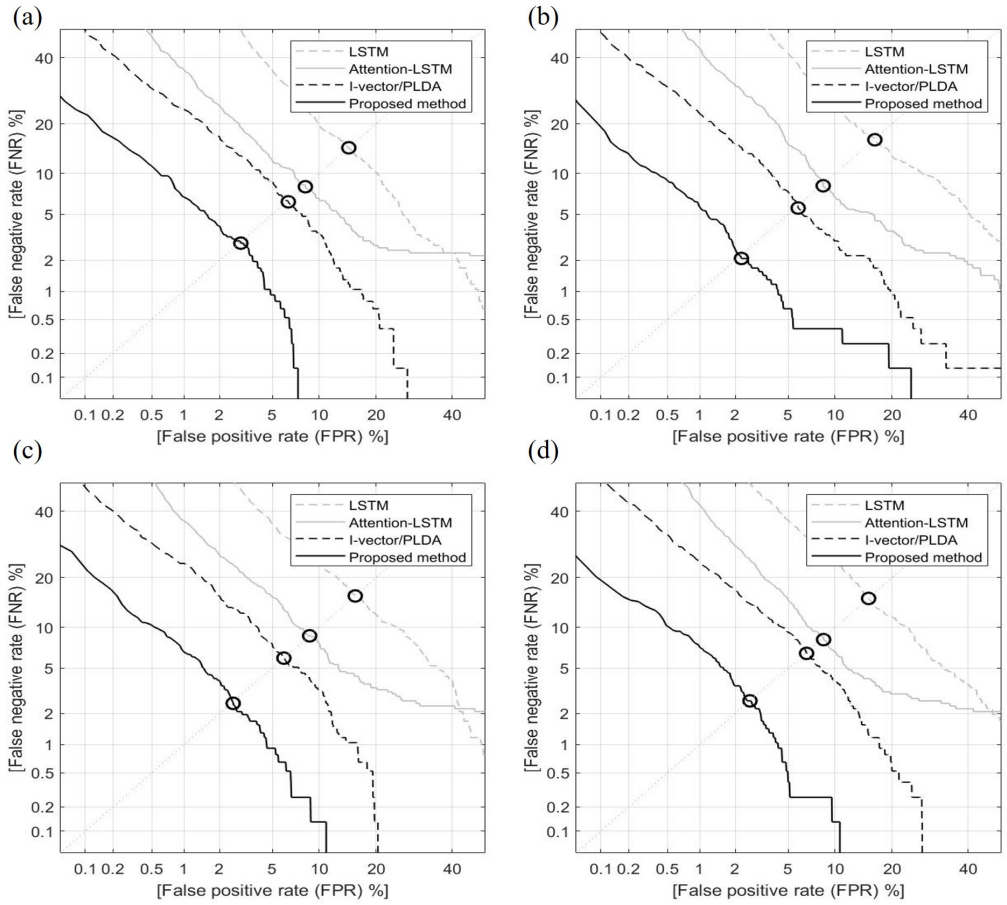

Fig. 6. The plot of Attention-LSTM, I-vector/PLDA and combined system based EER performance in a clean condition (a) off and noisy conditions (b) pub, (c) wind, (d) wNp.

의 집중 기법을 접목한 임베딩의 성능이 더 우수하였다. 이로써 임베딩을 사용자가 임의로 정의하여 사용하는 LSTM 모델의 문제점을 임베딩에 영향을 미치는 프레임의 가중치를 학습함으로써 해결할 수 있는 것을 확인하였다. 또한, I-vector/PLDA 시스템과 결합하였을 때 EER 2.57 %로 가장 좋은 성능을 보였다.

감사의 글

본 논문은 인천대학교 2015년 자체연구비 지원에 의하여 연구되었음.

References

1. P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, “Joint factor analysis versus eigenchannes in speaker recognition,” IEEE Trans on. Audio, Speech, and Language Processing, **15**, 2072-2084 (2007).
2. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Processing, **10**, 19-41 (2000).
3. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans on. Audio, Speech, and Language Processing, **19**, 788-798 (2011).
4. E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” Proc. ICASSP. 4080-4084 (2014).
5. 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).
6. Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, “Deep feature for text-dependent speaker verification,” Speech Communication, **73**, 1-13 (2015).
7. D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” Proc. Interspeech, 999-1003 (2017).
8. G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-toend text-dependent speaker verification,” Proc. IEEE ICASSP. 5115-5119 (2016).
9. D. Bahdanau, K. Cho, and Y. Bengio. “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473 (2014).
10. S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” Proc. IEEE 11th ICCV. 1-8 (2007).
11. B. Fauve, N. Evans, and J. Mason, “Improving the performance of text-independent short duration SVM- and GMM based speaker verification,” Proc. Odyssey, Stellenbosch, 18 (2008).
12. F. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, “Attention-based models for text-dependent speaker verification,” arXiv preprint arXiv:1710.10470 (2017).
13. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” arXiv preprint rXiv:1710.10467 (2017).
14. Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with lstm,” Proc. ICASSP. 5239-5243 (2018).

저자 약력

▸ 배 아 라(Ara Bae)

2019년 2월 : 인천대학교 컴퓨터공학부 학사
2019년 3월 ~ 현재 : 인천대학교 컴퓨터공학과 석사과정

▸ 김 우 일(Wooil Kim)

1996년 2월, 1998년 8월, 2003년 8월 : 고려대학교 전자공학과 학/석/박사
2012년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 조교수, 부교수