한국음향학회지 제39권 제5호 pp. 461~467 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.461

pISSN : 1225-4428
eISSN : 2287-3775

화자 구분 시스템의 관심 화자 추출을 위한 i-vector 유사도 기반의 음성 분할 기법

I-vector similarity based speech segmentation for interested speaker to speaker diarization system

배아라,[1] 윤기무,[2] 정재희,[1] 정보경,[1] 김우일[1†]

(Ara Bae,[1] Ki-mu Yoon,[2] Jaehee Jung,[1] Bokyung Chung,[1] and Wooil Kim[1†])

[1]인천대학교 컴퓨터공학부, [2]미디어젠

(Received June 29, 2020; accepted August 6, 2020)

초 록: 잡음이 많고 여러 사람이 있는 공간에서 음성인식의 성능은 깨끗한 환경보다 저하될 수밖에 없다. 이러한 문제점을 해결하기 위해 본 논문에서는 여러 신호가 섞인 혼합 음성에서 관심 있는 화자의 음성만 추출한다. 중첩된 구간에서도 효과적으로 분리해내기 위해 VoiceFilter 모델을 사용하였으며, VoiceFilter 모델은 여러 화자의 발화로 이루어진 음성과 관심 있는 화자의 발화로만 이루어진 참조 음성이 입력으로 필요하다. 따라서 본 논문에서는 Probabilistic Linear Discriminant Analysis(PLDA) 유사도 점수로 군집화하여 혼합 음성만으로도 참조 음성을 대체해 사용하였다. 군집화로 생성한 음성에서 추출한 화자 특징과 혼합 음성을 VoiceFilter 모델에 넣어 관심 있는 화자의 음성만 분리함으로써 혼합 음성만으로 화자 구분 시스템을 구축하였다. 2명의 화자로 이루어진 전화 상담 데이터로 화자 구분 시스템의 성능을 평가하였으며, 분리 전 상담사(Rx)와 고객(Tx)의 음성 Source to Distortion Ratio(SDR)은 각각 5.22 dB와 −5.22 dB에서 분리 후 각각 11.26 dB와 8.53 dB로 향상된 성능을 보였다.
핵심용어: 화자 구분, 군집화, 화자 임베딩, 음성 세그멘테이션

ABSTRACT: In noisy and multi-speaker environments, the performance of speech recognition is unavoidably lower than in a clean environment. To improve speech recognition, in this paper, the signal of the speaker of interest is extracted from the mixed speech signals with multiple speakers. The VoiceFilter model is used to effectively separate overlapped speech signals. In this work, clustering by Probabilistic Linear Discriminant Analysis (PLDA) similarity score was employed to detect the speech signal of the interested speaker, which is used as the reference speaker to VoiceFilter-based separation. Therefore, by utilizing the speaker feature extracted from the detected speech by the proposed clustering method, this paper propose a speaker diarization system using only the mixed speech without an explicit reference speaker signal. We use phone-dataset consisting of two speakers to evaluate the performance of the speaker diarization system. Source to Distortion Ratio (SDR) of the operator (Rx) speech and customer speech (Tx) are 5.22 dB and −5.22 dB respectively before separation, and the results of the proposed separation system show 11.26 dB and 8.53 dB respectively.
Keywords: Speaker diarization, Clustering, Speaker embedding, Speech segmentation

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

화자 구분이란 다수의 화자가 섞여 있는 음성에서 사람마다 고유한 음성 특징을 이용해 정보를 분석하여 각 화자의 신원에 대응되는 음성 조각으로 분할하는 방법이다. 기존의 화자 구분 시스템은 음성 구간을 검출한 뒤 각 음성 구간에서 화자의 정보를 담고 있는 특징을 추출해 화자별 음성으로 그룹화하여 분할하였으며, 그룹화를 위한 대표적인 군집화 방법으로 K-means, Agglomerative Hierarchical Clustering (AHC) 그리고 스펙트럼 군집화가 있다.[1-5] 최근 이러한 시스템 없이 구글에서 혼합된 음성과 관심 있는 화자의 음성으로만 이루어진 참조 음성을 신경망의 입력으로 넣어 음성을 분리하는 VoiceFilter 모델을 제안하였으나 혼합 음성만 존재하는 경우 사용할 수 없다는 문제점이 있다.[6] 본 논문에서는 기존의 화자 구분 시스템으로 화자별 음성을 생성함으로써 관심 있는 화자의 음성만 존재하는 참조 음성을 대체하여 VoiceFilter와 접목해 이러한 문제점을 해결하였다. 화자별 음성을 생성하는 데 필요한 군집화 기법 중 K-means와 스펙트럼 군집화를 비교하고, VoiceFilter에 사용되는 화자 특징으로 i-vector와 임베딩을 사용하여 성능을 측정하였다. 2장에서는 화자 구분 시스템을 위한 2가지 모델을 소개하고, 3장에서는 군집화를 이용하여 화자별 음성을 생성하는 모델을 제안한다. 4장은 실험에 사용된 데이터와 각 방법에 대해 성능을 비교하여 정리하였다.

II. 화자 구분 시스템

전화 상담 데이터는 수신자(상담사, Rx)와 송신자(고객, Tx) 2명의 화자로 이루어져 있다. 본 논문에서는 상담사와 고객의 음성이 섞여 있는 혼합 음성에서 두 음성을 분리해내는 것을 목표로 한다. 두 명의 화자가 존재하는 음성에서 독립적으로 발화하는 구간뿐만 아니라 중첩되는 구간에서도 관심 있는 화자(Person Of Interest, POI)의 음성만 뽑아내기 위해 VoiceFilter 모델을 사용하였다. VoiceFilter 모델은 Fig. 1과 같이 여러 명의 발화로 이루어진 혼합 음성과 POI의 화자 정보를 추출하기 위해 사용할 참조 음성

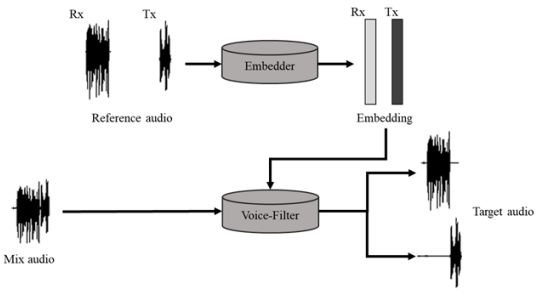

Fig. 1. Framework of proposed speaker diariazation system.

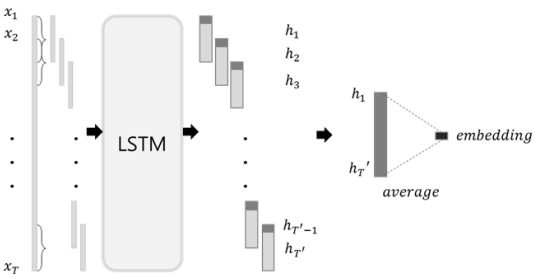

Fig. 2. Extractor of LSTM based speaker embedding.

이 입력으로 필요하다. 2.1절에서는 화자 특징 추출기를 설명하고, 2.2절에서는 VoiceFilter 모델을 이용하여 POI의 음성만 분리하는 시스템에 관해 설명한다.

2.1 화자 정보 추출기(Embedder)

I-vector는 화자 정보 특징 중 널리 사용되고 있는 특징으로, 최근 딥러닝이 발전하며 이를 대체하려는 임베딩에 관한 연구들이 활발하게 진행되었다.[7,8] 본 논문에서는 i-vector와 임베딩을 화자 특징으로 사용하여 성능을 비교하였다. I-vector 추출을 위해 1024개의 요소로 이루어진 Gaussian mixture Model-Universal Background Model(GMM-UBM)과 100차원의 Total Variability Matrix(TVM) 모델을 구축하였다. 39차원 Mel Frequency Cepstral Coefficient(MFCC) 특징을 입력으로 사용하였다.[9,10] 임베딩 추출기는 3개의 층으로 이루어진 Long Short Term Memory(LSTM)를 Generalized End-To-End(GE2E) 손실로 훈련하여 구축하였다.[8,11] 50 ms의 창을 20 ms씩 이동하면서 추출한 log-mel filterbank를 LSTM의 입력으로 사용하였다. Fig. 2와 같이 한 발화에 대해 80프레임 크기의 창을 50 % 겹치게 옆으로 이동하며 256차원의 임베딩을 생성한

뒤 각 창의 마지막 부분만 모아 평균 낸 것을 최종 화자 임베딩으로 사용하였다.

2.2 VoiceFilter 기반 음성 분리 시스템

VoiceFilter는 Convolutional Neural Network(CNN)와 LSTM, 완전 연결 Deep Neural Network(DNN)로 구성된다. VoiceFilter는 참조 음성으로부터 화자 정보를 추출하여 혼합 음성 내에 POI의 음성만 걸러낼 수 있는 마스크를 생성한다. 생성된 마스크는 혼합 음성과 컨볼루션 연산을 통해 POI의 음성을 분리해낸다. 본 논문에서 사용한 전화 상담 데이터에서 두 채널의 음성을 더한 것을 혼합 음성, 각 채널을 목표로 하는 음성과 참조 음성으로 VoiceFilter의 입력에 사용하였다. 목표로 하는 음성과 마스크와 연산하여 얻은 음성의 손실을 계산하여 VoiceFilter를 훈련하였다. 그러나 실제 사용환경에서는 깨끗한 환경에서 발화한 POI의 음성으로만 이루어진 데이터가 없는 경우가 많다. 또한, VoiceFilter는 참조 음성이 없는 경우 음성을 분리해낼 수 없으며, 참조 음성에 성능이 의존적인 문제점이 존재한다. 따라서 본 논문에서는 혼합된 음성에서 최대한 POI의 음성만 추출하여 참조 음성을 대체할 수 있는 시스템을 구축하였다. 이는 POI의 음성으로만 이루어진 이미 분리되어있는 데이터를 참조 음성으로 사용하였을 때와 유사한 성능을 보였다.

III. 군집화 기반 음성 분할

혼합 음성에서 POI의 음성만 추출하기 위해 혼합 음성 내에 존재하는 화자의 수만큼 군집화하여 화자별 음성을 생성하였다. 음성이 아닌 구간은 화자의 특성을 추출하는데 불필요한 요소이므로 음성만 가져오는 선행 작업이 필요하다. 신호의 짧은 구간마다 검출한 에너지는 비음성 구간과 음성 구간을 구별하는데 사용할 수 있다. Fig. 3(a)와 같이 구간별 에너지를 추출 후 정렬하여 25 %에 해당하는 값을 음성/비음성을 구분하는 문턱 값으로 설정하였다.[12] 문턱 값보다 큰 구간이 30 프레임 이상 지속 되면 해당 구간 모두 음성으로 판단하였으며, 그 이하인 구간은 모두 비음성으로 판단하였다. Fig. 3(b)는 3 min 길이의 데이터의 0 s ~ 5 s 구간에서 음성으로 검출된 것을 보여준다.

2.1장에서 사용한 i-vector 추출기와 같은 구조의 모델을 사용하여 각 음성 구간에서 i-vector를 추출한 뒤, 추출된 i-vector 간 Probabilistic Linear Discriminant Analysis (PLDA) 기반의 유사도 점수를 계산하여 음성에 존재하는 화자의 수만큼 군집화 기법을 이용해 그룹화하였다.[5,13] 본 논문에서는 K-means와 스펙트럼 군집화 방법을 비교하였으며, Fig. 4는 두 군집화 방법으로 음성 구간이 그룹화된 결과이다. 스펙트럼 군집화의 가우시안 블러 σ는 1, 행 단위 임계값은 0.95로 두었다. Fig. 4에서 K-means가 스펙트럼 군집화보다 잘 그룹화되는 것을 확인하였다. K-means 군집화 후 그룹의 중심으로부터 같은 거리에 떨어져 있는 음성을 2명 이상의 화자 발화가 겹쳐있는 중첩 구간으로 두었으며, 0 ~ 1 사이의 값을 갖는 가중치를 이용하여 중첩 구간의 폭을 조정 후 사용하였다. Table 1과 Fig. 5는 가중치에 따른 화자별 Diarization

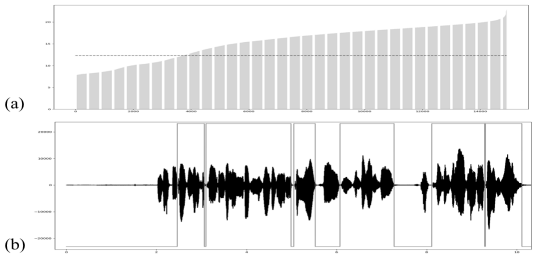

Fig. 3. (a) Threshold of sorted energies in detected speech durations and (b) voice activity detected position of mixed speech.

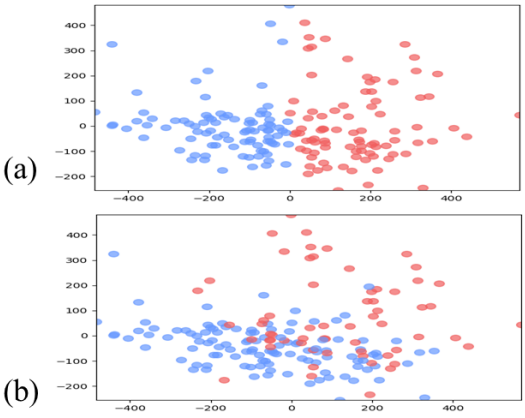

Fig. 4. (Color available online) Clustering of segmented speech by (a) K-mean algorithm and (b) spectral clustering.

Table 1. DER [%] value for each speaker in the mixed speech that changes by weight (α).

α	spk 0	spk 1	AVG
0.1	32.37	46.92	39.65
0.2	32.37	46.92	39.65
0.3	32.36	46.88	39.62
0.4	32.31	46.73	39.52
0.5	32.10	46.74	39.42
0.6	31.91	46.29	39.10
0.7	32.01	44.79	38.40
0.8	**32.11**	**43.46**	**37.79**
0.9	33.13	42.94	38.04
1.0	33.79	42.60	38.20

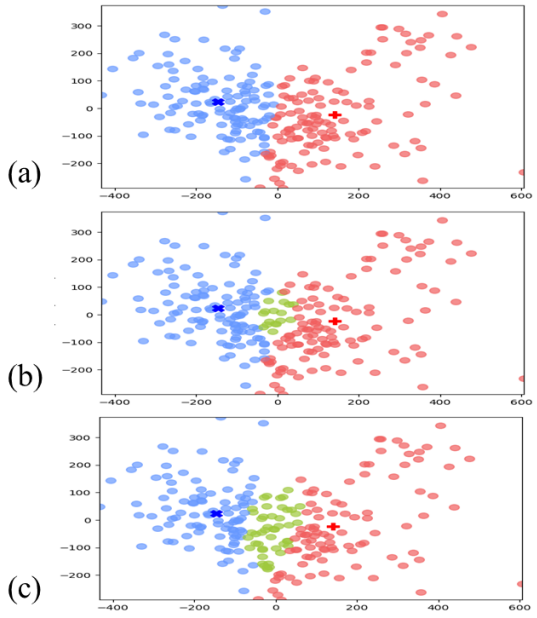

Fig. 5. (Color available online) Visualization of changed overlapping speech by weight.

Error Rate(DER) 성능과 중첩 구간의 폭 변화를 시각화한 것이다. 가중치를 0.8로 두었을 때 성능이 가장 좋았으며, 이때 중첩 구간은 제외하고 각 그룹의 음성을 모아 화자별 음성을 생성하여 화자 구분 시스템의 참조 음성으로 활용하였다.

IV. 실험 및 결과

I-vector 추출에 필요한 UBM과 TVM 모델 훈련을 위해 잡음이 없는 깨끗한 환경에서 996명의 화자가 49~50회씩 발화한 한국어 발화 데이터를 사용하였다. 약 3 s~4 s 서로 다른 문장을 발화하였으며 8 kHz로 샘플링되었다. 두 가지 화자 정보 추출기와 화자 구분 시스템은 8 kHz로 샘플링된 2명의 화자로 이루어진 평균적으로 5분 길이의 4784개 전화 상담 데이터를 사용하였다. 훈련 데이터와 같은 환경의 48개 전화 상담 데이터로 화자 구분 시스템 성능 평가를 시행하였다. 본 논문에서 제안한 화자 구분 시스템으로 추출한 POI 음성의 스펙트로그램은 Fig. 6(c)이며, 혼합 음성 (a)과 모델에서 얻은 마스크(b)로 연산한 결과이다. 또한, 실제 분리되어있는 음성 스펙트로그램 (d)과 오차는 (e)와 같다.

모델의 성능지표로 화자 정보 추출기는 POI가 아닌 것을 POI로 잘못 판단한 확률(False Acceptance Rate, FAR)과 POI인 것을 POI가 아니라고 잘못 판단할 확률(False Rejection Rate, FRR)이 같아지는 비율인 Equal Error Rate(EER)을, 화자 구분 시스템은 혼합 음성에서 POI의 신호와 POI가 아닌 신호의 에너지 비율인 SDR을 성능지표로 이용하였다.

4.1 Equal Error Rate

Table 2는 i-vector, 임베딩의 성능 결과표이다. EER은 화자 인식에서 화자 정보 특징이 화자의 특성을 잘 표현하는지 평가한다. i-vector와 LSTM으로 추출한 임베딩의 성능을 비교하였을 때 잡음이 없는 환경에서 발화한 데이터는 i-vector의 EER이 상대적으로 53.13 % 낮았고, 잡음이 있는 전화 대화 데이터에서 임베딩의 EER이 상대적으로 5.36 % 낮았다.

4.2 Source to Distortion Ratio[14]

신호 대 왜곡 비율(Source to Distortion Ratio, SDR)은 추정한 신호와 목표 신호로 측정할 수 있으며, 값이 클수록 왜곡이 적어 좋은 성능을 보인다. Table 3는 음성 분리 전과 i-vector, 임베딩을 화자 특징으로 사용하였을 때의 성능 결과표이다. 전화 데이터에서 수신자(상담사, Rx)와 송신자(고객, Tx)의 음성 분리하지 않은 혼합 음성의 SDR은 각각 5.22 dB, −5.22 dB 성능을 보였다. LDA를 적용한 i-vector를 화자 정보로 사용하여 훈련한 VoiceFilter로 분리된 Rx 음성의 SDR은 11.01 dB, Tx 음성의 SDR은 8.54 dB을 보였으

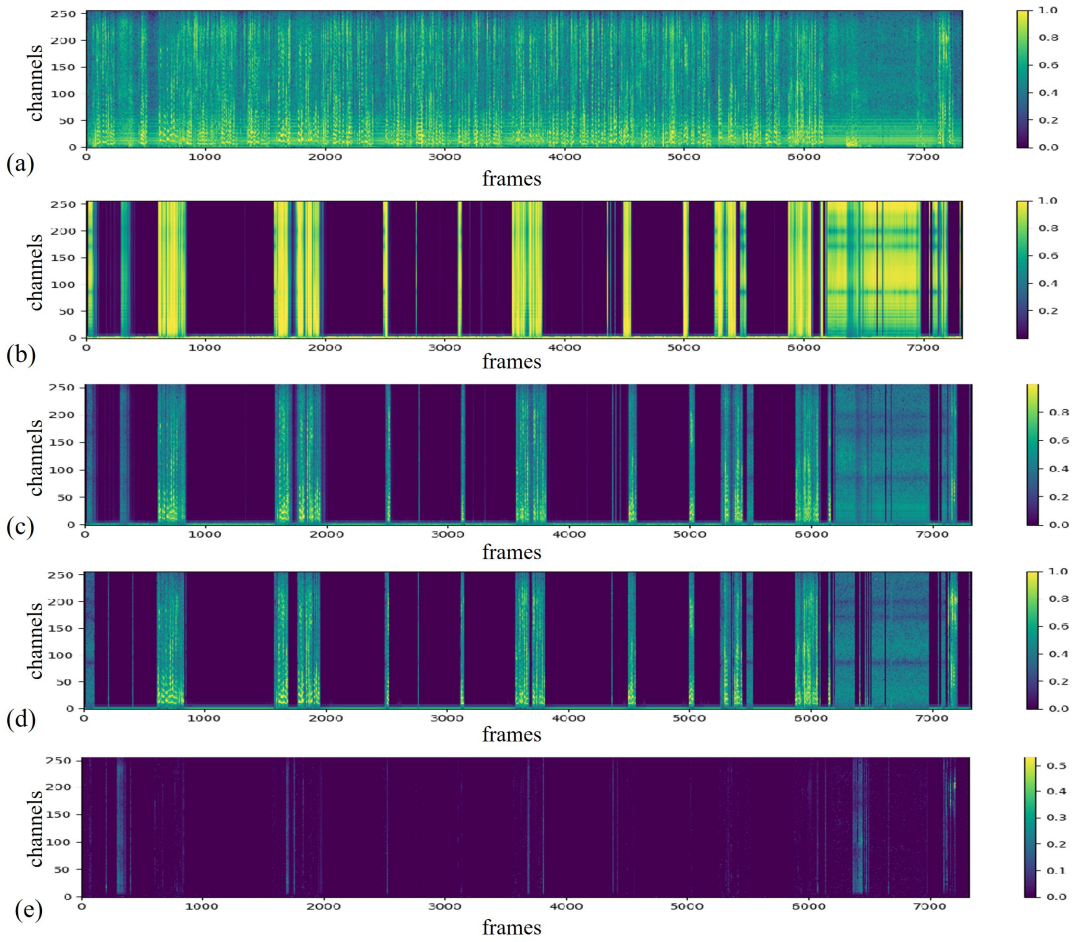

Fig. 6. (Color available online) Results of spectrogram from a VoiceFilter model. In order (a) mixed speech, (b) mask created by VoiceFilter, (c) speech of interested speaker obtained by operation of mixed speech and mask, (d) speech only of actual interested speaker and (e) loss between (c) and (d).

Table 2. EER [%] results of i-vector and embedding system.

EER [%]	i-vector	Embedding
Korean	2.32	4.95
Telephone conversations	13.74	13.04

Table 3. SDR [dB] results of No VoiceFilter and VoiceFilter with i-vector/embedding.

SDR [dB]	Rx Mean	Rx Standard deviation	Tx Mean	Tx Standard deviation
No VoiceFilter	5.22	5.67	-5.22	5.66
VoiceFilter + i-vector (+LDA)	11.01	2.15	8.54	1.64
VoiceFilter + Embedding	**11.26**	**2.08**	**8.86**	**2.23**

며 임베딩을 사용하여 훈련한 모델의 경우 Rx와 Tx 음성의 SDR이 각각 11.26 dB, 8.86 dB을 보였다.

V. 결 론

다양한 잡음 환경과 다수의 화자가 있는 환경에서 음성인식의 성능은 저하된다. 관심 있는 화자의 음성만 추출하여 인식할 경우 성능이 향상될 수 있다. 본 논문에서는 분할된 음성 구간에서 화자 정보를 뽑고 화자별로 군집화하여 혼합 음성 안에 존재하는 화자별 음성에서 추출한 화자 특징과 VoiceFilter 모델을 접목해 독립 발화한 구간과 중첩된 구간 모든 음성에서 관심 있는 화자의 음성만 추출하였다.

임베딩을 이용한 VoiceFilter가 i-vector를 사용하였

을 때보다 상대적으로 Rx 음성은 2.22 %, Tx 음성은 3.61 % 성능이 향상되었다. i-vector보다 임베딩이 잡음이 많은 환경에서 더 좋은 성능을 보였다. 이로써 화자 특징 추출기 성능이 VoiceFilter 성능에 영향을 주는 것을 확인하였다. 따라서 화자 특징 추출기 그리고 입력으로 사용되는 참조 음성이 분리기 성능에 핵심적인 요소라고도 볼 수 있다.

본 논문에서는 POI의 음성만 존재하는 깨끗한 발화 대신 군집화 기법을 통해 혼합 음성으로부터 POI의 음성만 추출해 참조 음성을 대체함으로써 혼합 음성만 존재하는 때도 화자별 음성을 효과적으로 분리하였다.

감사의 글

본 논문은 인천대학교 2016년 자체연구비 지원에 의하여 연구되었음.

References

1. G. Sell and D. Garcia-Romero, “Speaker diarization with plda i-vector scoring and unsupervised calibration,” Proc. of the IEEE Spoken Language Technology Workshop, 413-417 (2014).
2. G. Sell and D. Garcia-Romero, “Diarization resegmentation in the factor analysis subspace”, Proc. ICASSP. 4794-4798 (2015).
3. D. Dimitriadis and P. Fousek, “Developing on-line speaker diarization system,” Proc. Interspeech, 2739-2743 (2017).
4. Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with LSTM,” Proc. ICASSP. 5239-5243 (2018).
5. Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, “LSTM based similarity measurement with spectral clustering for speaker diarization,” Proc. Interspeech, Graz, 366-370 (2019).
6. Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, “VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking,” arXiv: 1810.04826 (2018).
7. 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).
8. G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” Proc. IEEE ICASSP. 5115-5119 (2016).
9. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Processing, **10**, 19-41 (2000).
10. 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).
11. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” arXiv preprint rXiv:1710.10467 (2017).
12. W. Kim and J. H. L. Hansen, “Advanced parallel combined Gaussian mixture model based feature compensation integrated with iterative channel estimation,” Speech Communication, **73**, 81-93 (2015).
13. S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” Proc. IEEE 11th ICCV. 1-8 (2007).
14. E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” IEEE Trans. on Audio, Speech, and Lang. Processing, **14**, 1462-1469 (2006).

저자 약력

▸ 배 아 라 (Ara Bae)

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

▸ 윤 기 무(Ki-mu Yoon)

2018년 2월 : 인천대학교 수학과 학사
2020년 2월 : 인천대학교 컴퓨터공학과 석사
2020년 3월 ~ 현재 : 미디어젠 연구원

▸ 정 재 희 (Jaehee Jung)

2017년 3월 ~ 현재 : 인천대학교 컴퓨터공학부 학사과정

▸ 정 보 경 (Bokyung Chung)

2017년 3월 ~ 현재 : 인천대학교 컴퓨터공학부 학사과정

▸ 김 우 일 (Wooil Kim)

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