한국음향학회지 제39권 제5호 pp. 447~453 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.447

pISSN : 1225-4428
eISSN : 2287-3775

화자 인식을 위한 적대학습 기반 음성 분리 프레임워크에 대한 연구

A study on speech disentanglement framework based on adversarial learning for speaker recognition

권유환,[1] 정수환,[1] 강홍구[1†]

(Yoohwan Kwon,[1] Soo-Whan Chung,[1] and Hong-Goo Kang[1†])

[1]연세대학교 전기전자공학부

(Received July 31, 2020; accepted September 16, 2020)

초 록: 본 논문은 딥러닝 기법을 활용하여 음성신호로부터 효율적인 화자 벡터를 추출하는 시스템을 제안한다. 음성신호에는 발화내용, 감정, 배경잡음 등과 같이 화자의 특징과는 관련이 없는 정보들이 포함되어 있다는 점에 착안하여 제안 방법에서는 추출된 화자 벡터에 화자의 특징과 관련된 정보는 가능한 많이 포함되고, 그렇지 않은 비화자 정보는 최소화될 수 있도록 학습을 진행한다. 특히, 오토-인코더 구조의 부호화 기가 두 개의 임베딩 벡터를 추정하도록 하고, 효과적인 손실 함수 조건을 두어 각 임베딩이 화자 및 비화자 특징만 각각 포함할 수 있도록 하는 효과적인 화자 정보 분리(disentanglement)방법을 제안한다. 또한, 화자 정보를 유지하는데 도움이 되는 생성적 적대 신경망(Generative Adversarial Network, GAN)에서 활용되는 판별기 구조를 도입함으로써, 디코더의 성능을 향상시킴으로써 화자 인식 성능을 보다 향상시킨다. 제안된 방법에 대한 적절성과 효율성은 벤치마크 데이터로 사용되고 있는 Voxceleb1에 대한 동일오류율(Equal Error Rate, EER) 개선 실험을 통하여 규명하였다.
핵심용어: 화자 임베딩, 정보 분리, 다중작업, 판별기

ABSTRACT: In this paper, we propose a system to extract effective speaker representations from a speech signal using a deep learning method. Based on the fact that speech signal contains identity unrelated information such as text content, emotion, background noise, and so on, we perform a training such that the extracted features only represent speaker-related information but do not represent speaker-unrelated information. Specifically, we propose an auto-encoder based disentanglement method that outputs both speaker-related and speaker-unrelated embeddings using effective loss functions. To further improve the reconstruction performance in the decoding process, we also introduce a discriminator popularly used in Generative Adversarial Network (GAN) structure. Since improving the decoding capability is helpful for preserving speaker information and disentanglement, it results in the improvement of speaker verification performance. Experimental results demonstrate the effectiveness of our proposed method by improving Equal Error Rate (EER) on benchmark dataset, Voxceleb1.
Keywords: Speaker embedding, Disentanglement, Multi-task, Discriminator

PACS numbers: 43.71.Bp, 43.72.Fx

I. 서 론

화자 인식은 주어진 음성신호에서 발화자의 목소리 특징 정보를 추출하여 인식하는 기술이다. 특히, 최근에는 사람과 기기 간에 효과적으로 정보를 교환하고 제어하기 위하여 사용자를 파악하고 사용자 맞춤형 서비스를 제공하기 위한 핵심 기반 기술로서 화자 인식기술의 중요성이 대두되고 있다. 최근 다

†**Corresponding author:** Hong-Goo Kang (hgkang@yonsei.ac.kr)
School of Electrical and Electric Engineering, YONSEI University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
(Tel: 82-2-2123-4534)

양한 딥러닝 알고리즘이 개발되어 화자인식의 성능이 크게 개선되고 있어 그 중요성과 유용성이 더욱 커지고 있다. 딥러닝을 활용한 화자인식 시스템은 신경망을 이용하여 동일한 화자가 발화한 음성신호 전반에서 공통적으로 나타나지만 다른 화자와는 다른 특징 벡터를 추출하고, 이를 기반으로 인식 및 분류하는 시스템이다. 대표적인 딥러닝 기반 화자 인식 기술 방법에는 d-vector,[1] x-vector[2] 등이 있다. 이러한 방법들은 시간적 모델링과 화자에 관한 손실함수를 기반으로 다양한 네트워크 구조를 통해 저차원에서 표현되는 임베딩 벡터를 추출하는 것으로서 다양한 환경에서 강인한 화자인식 성능을 보였다.

본 논문에서는 분리 방법을 통한 효과적인 화자 정보 모델링 기법에 대하여 제안한다. 기존의 인코더를 이용한 화자 모델링 기법은 매우 효과적이나 음성 신호 내에 존재하는 화자 정보와 무관한 정보를 제거하는데 목적을 두지 않는다. 이로 인하여 순수한 화자 정보만을 추출하기 어려우며 이는 인식 성능의 하락을 초래한다. 최근 제안된 음성의 정보 분리 기반의 인코더-디코더 방법[3]은 화자의 특징을 모델링하되 화자와 관련 없는 정보를 잔여 인코더로 학습한다. 각각 모델링된 정보를 디코더를 이용하여 입력 음성으로 복원하여 분리로 인한 정보의 유실을 방지하며, 이를 통해 화자 정보 모델링의 학습 효율성을 강화한다. 하지만, 이 방법 또한 분리된 정보의 유실을 완전히 방지하지 못하여 화자인식 및 디코더의 성능을 저하시킨다. 본 논문에서는 이러한 분리 기반 방법의 문제점을 해결하기 위하여 생성적 적대 신경망(Generative Adversarial Network, GAN)을 이용하고 이를 통해 기존 방법의 디코더에서의 복원 성능의 한계점을 보완할 수 있도록 학습 기준을 새롭게 제안한다. 특히 다중작업 기반의 판별기를 이용하여 화자 분리 및 인식 성능의 향상을 기대한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 시스템 구현에 있어서 관련 정보를 설명하고, 3장에서는 제안하는 시스템의 구조 및 학습방법에 대해 설명한다. 4장에서는 실험 내용 및 결과를 설명하고 이를 바탕으로 5장에서는 연구의 결론 및 효과에 대하여 정리한다.

II. 배경 지식

2.1 화자 인식

딥러닝 기반 화자인식 시스템은 학습된 모델로부터 화자 임베딩 벡터를 추출하고 이를 이용하여 화자를 분류 혹은 검증한다. 화자 임베딩은 음성 신호에 적합한 여러 가지 신경망 구조를 활용하여 구현되었다. 예를 들어, 심층 신경망(Deep Neural Network, DNN)에 기반한 d-vector, 시간 지연 신경망(Time-Delay Neural Network, TDNN)에 기반한 x-vector, 그리고 합성곱 신경망(Convolution Neural Network, CNN)에 기반한 등의 방법들이 제시되었고 이들은 기존의 통계적 모델링 방법들에 비하여 뛰어난 성능을 보여주었다.[4,5]

최근 Tai *et al*.[3]은 정보 분리 방법을 적용하여 더욱 뛰어난 성능을 보였다. 이 방법은 두 개의 인코더와 하나의 디코더를 사용하였으며, 두 개의 인코더는 각각 화자 정보를 추출하는 화자 인코더와 그 외의 정보들을 모델링하는 잔여 인코더로 구성된다. 디코더는 앞선 두 인코더의 출력을 입력으로 사용하여 입력된 음성 신호를 복원하는 역할을 한다. 디코더의 역할은 인코더를 통해 분리된 정보들의 정보량이 유실되지 않도록 방지하며 이를 통해 화자의 특징을 제외한 정보들이 잔여인코더에 나타날 수 있도록 한다.

이를 학습하기 위한 학습 기준은 여러 가지 손실함수의 조합으로 나타내어진다. 첫째, 입력된 화자의 레이블을 학습하도록 하는 교차 엔트로피 오차를 이용하여 화자 인코더를 학습한다. 이 때, 화자 인코더 후에 분류기를 추가하여 화자의 레이블을 학습함으로써 화자 임베딩이 효과적으로 표현되도록 한다. 이렇게 학습된 분류기는 잔여 인코더를 학습하는데 사용되는데, 이 때 적대 학습 방법을 이용하여 잔여 인코더의 임베딩이 어떤 화자의 레이블도 추정하지 못하도록 학습한다. 마지막으로, 디코더는 전체 시스템의 입력과 디코더의 출력이 유사해지도록 평균 제곱근 오차(Mean Squared Error, MSE)를 최소화하도록 학습한다.

이러한 방식으로 신경망을 학습함으로써 화자 레이블을 추정하도록 학습하는 인코더 방식에 비해 화자 정보를 추정할 뿐만 아니라 화자 외의 정보는 추

출하지 않음으로 향상된 화자 인식 성능을 보인다.

2.2 Generative Adversarial Network(GAN)

생성적 적대 신경망[6]은 게임이론에 바탕을 둔 방법이다. GAN은 생성기와 판별기, 두 개의 연속된 네트워크로 이루어진다. 생성기는 목표로 하는 데이터를 생성하며 판별기는 생성된 데이터의 진실 유무를 판별한다. 그러므로 학습할 때에는, Minimax 게임에서 차용된 손실함수 $V(G,D)$를 이용하여 생성기는 정교한 데이터를 생성하도록 하며, 판별기는 생성된 데이터의 진실 유무를 정확하게 구별할 수 있도록 학습한다.

$$\min_G \max_D V(G,D) = E_{x\sim p_{data(x)}}[\log D(x)] + E_{z\sim p_{z(z)}}[\log(1-D(G(z))]. \quad (1)$$

이 때, z는 확률변수를 가리키며, x는 입력 데이터, $G(z)$는 생성기로 생성된 데이터를 나타낸다. GAN은 주로 생성기반의 학습에 이용되는 방법이지만 최근에는 이와 같이 분류 기반의 연구[7-9]에도 사용한다.

III. 본 론

3.1 기본 시스템 구조

본 연구에서는 Fig. 1에서의 구조와 같이 Tai *et al.*[3]에서 사용된 구조를 기반 시스템으로 사용한다. 이 구조는 앞서 설명한 내용과 같이 정보 분리 방법을 활용하여 더욱 효과적인 화자 임베딩 벡터를 추출하는 방법이다.

기존의 방법은 화자 인코더와 잔여 인코더를 이용하여 분리된 임베딩 벡터들로부터 디코더를 통해 입력 데이터를 복원한다. 하지만, 시간 및 주파수 축에 존재하는 입력 데이터의 정보들을 일차원의 임베딩으로 압축하는 과정에서 불가피하게 일부의 정보가 유실될 수 있다. 예를 들어, 입력 신호에 배경 잡음이나 잔향이 존재할 때에는 잔여 인코더에서 이를 모두 임베딩하기 어렵다. 이와 같은 결과는 디코더가 데이터를 복원하는데 치명적인 영향을 주며 결과적으로 원본 데이터를 추론하여 화자 인식의 성능을 높이고자 하는 방법을 방해한다.

제안하는 방법에서는 데이터의 복원 성능에 한계가 있다는 점을 보완하고 분리된 데이터의 정보량을 유지한다는 점에 착안하여 디코더와 판별기를 이용하여 학습하도록 한다. GAN에서 사용된 바와 같이 디코더가 음성을 생성하도록 하고, 판별기는 생성된 음성이 실제 음성신호의 특성과 유사해지도록 디코더를 학습한다. 제안하는 방법은 총 네 가지의 손실함수에 의해 복합적으로 학습되며, 사용된 각 손실함수와 의미는 다음과 같다.

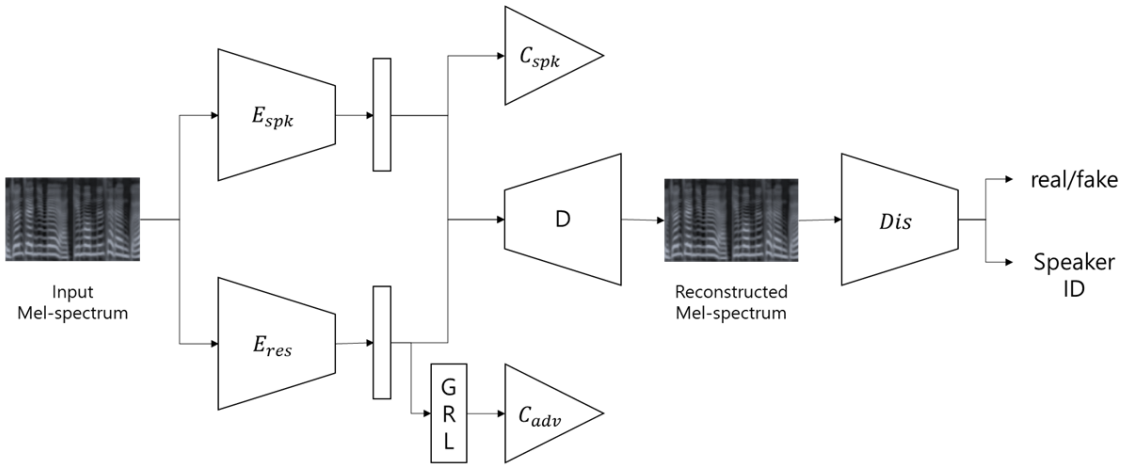

Fig. 1. Overview structure of the proposed system. E_{spk} means speaker encoder, E_{res} means residual encoder which encodes speaker unrelated information. D indicates decoder network and Dis indicates discriminator which determines that input is real or fake and speaker id.

3.2 화자 정보 인코더

화자 정보 인코더의 목적은 더욱 정확한 화자 임베딩을 얻는데 있다. 화자 정보 인코더의 출력인 화자 임베딩은 화자 분류기 C_{spk}와 함께 기존의 화자 분류 구조로 학습한다. 본 논문에서는 아래 식과 같이 화자 분류 구조로 교차 엔트로피 오류를 사용하여 화자 정보 인코더를 학습한다.

$$L_{spk} = -\sum_{i=1}^{N} k_i \log(C_{spk}(E_{spk}(x))). \quad (2)$$

이때, N는 총 화자 수, k_i는 화자 레이블, E_{spk}는 화자 인코더를 가리킨다. 학습된 화자 정보 인코더는 입력 멜-스펙트럼(mel-spectrum)에서 화자 정보를 추출하여 다른 화자들과 구별될 수 있는 벡터를 출력한다.

3.3 잔여 정보 인코더

잔여 정보 인코더는 음성 신호 내 화자 외의 정보를 나타내는 인코더이며, 이는 적대적 분류기(adversarial classifier)를 이용하여 학습한다. 분류기는 화자를 분류하도록 학습하되 잔여 정보 인코더는 학습된 분류기를 통해 화자를 구분할 수 없도록 한다.

$$L_{res,c} = -\sum_{i=1}^{N} k_i \log(C_{adv}(E_{res}(x))). \quad (3)$$

$$L_{res,enc} = \frac{1}{N}\sum_{i=1}^{N} \log(C_{adv}(E_{res}(x))). \quad (4)$$

이와 같이 학습하기 위해서 손실 함수는 각각 Eq. (3)과 Eq. (4)로 표현 할 수 있는데, E_{res}는 잔여 정보 인코더, C_{adv}는 적대적 분류기를 나타낸다. Eq. (3)에서는 화자 인코더와 마찬가지로 교차 엔트로피 오류를 통해 분류기 만을 학습하고 잔여 인코더는 업데이트하지 않도록 한다. Eq. (4)에서는 잔여 인코더와 분류기를 통해 구한 사후 확률(posteriori probability)이 균일 분포를 이루어 어떤 화자 레이블도 특정할 수 없도록 학습을 한다. 이 때에는 분류기는 업데이트하지 않고 잔여 인코더만을 학습한다.

이를 통해 잔여 인코더는 화자를 특정하지 못하는 임베딩을 생성하며 이를 통해 잔여 임베딩은 화자를 제외한 나머지 정보를 표현하게 된다.

3.4 디코더

3.1과 3.2에서 설명한 두 인코더의 임베딩 벡터들은 각각 화자와 잔여 정보를 내포하는 임베딩 내에 이상적으로 입력 멜-스펙트럼의 정보량을 모두 유지해야한다. 이를 위해 디코더는 앞의 두 임베딩을 사용하여 입력 멜-스펙트럼을 재구성한다. 디코더는 복원된 멜-스펙트럼과 입력 멜-스펙트럼의 차이를 Eq. (5)와 같이 평균 제곱오차를 사용하여 측정하고 최소화하도록 학습한다.

$$L_{mse} = \| x - D(E_{spk}(x) \oplus E_{res}(x)) \|^2. \quad (5)$$

이때 D는 디코더, $\oplus$는 두 임베딩 벡터의 결합(concatenation)을 의미한다.

3.5 판별기(Discriminator)

제안하는 방법에서는 L_{mse}뿐만 아니라 Eq. (6)의 수식과 같이 추가적으로 GAN에서 사용되는 판별기(*Dis*) 방법을 차용하여 두 개의 인코더와 디코더를 공동 학습한다.

$$L_D = \min_{D,E_{spk},E_{res}} \max_{Dis} E[\log Dis(x)] + E[\log(1 - Dis(\hat{x})]. \quad (6)$$

이 때, $\hat{x}$은 디코더를 통해 복원된 멜-스펙트럼이다. 또한, References [7], [8] 등에서 사용된 것과 같이 판별기를 다중 작업 판별기로 구성하여 Eq. (7)과 같이 복원된 스펙트럼에서 화자를 구분하는 작업을 병행하여 학습한다.

$$L_C = -\sum_{i=1}^{N} k_i \log(Dis(\hat{x})). \quad (7)$$

이때, k_i는 화자 레이블을 나타낸다. 이와 같이 다중 작업 판별기는 복원된 데이터의 품질향상과 더불어 화자 인식 기능을 수행하므로 화자 임베딩 학습에

대한 효과를 높이는 방향으로 네트워크를 학습하게 된다.

IV. 실험 및 결과

본 장에서는 제안하는 화자 인식 시스템의 성능을 확인하기 위한 실험의 내용을 설명하고 그 결과에 대해 분석한다. 실험의 내용은 Reference [3]의 시스템과 비교할 수 있도록 동일한 실험 구성에서 진행되었다.

4.1 학습 데이터셋

본 논문에서 제안된 시스템을 학습하기 위해 사용한 데이터셋은 Voxceleb2[9]로 5,994명의 화자와 유튜브에서 추출한 1,000,000개 이상의 음성 샘플로 구성된 데이터셋이다. 검증 데이터셋은 여러 논문에서 기준지표 데이터셋으로 사용되는 Voxceleb1[10] test를 사용하여 실험을 진행하였다. 각 음성 샘플은 3초로 분절하여 고정된 길이로 학습을 진행하였고, 단시간 푸리에 변환을 위하여 프레임 길이는 25 ms로 정하고 매 10 ms 간격으로 푸리에 변환을 하였다. 최종적으로 입력 특징 벡터로서 64차의 로그 스케일의 멜 스펙트럼을 추출하여 사용하였다.

4.2 네트워크 구조 및 검증 방법

실험에 사용한 네트워크의 상세 구조는 인코더는 Table 1과 같이 ResNet34[11] 구조를 사용하였고, 디코더는 DCGAN[12]에서 사용한 바와 같이 9개의 전치 합성곱 신경망(transposed CNN)와 3개의 완전 연결 층을 사용하였다. 판별기는 8개의 합성곱 신경망 구조를 사용하고 다중업무를 위해 별도의 완전 연결층 두 개를 사용하였다. 각 인코더에서 풀링 방법으로는 Self-Attentive Pooling(SAP)[13] 방법을 사용하였다. 학습을 위한 옵티마이저로는 Adam을 사용하였으며 이의 학습율은 0.001로 설정하였다. 실험의 유용성 및 성능은 동일 오류율(Equal Error Rate, EER)을 사용하여 검증하였다. 동일오류율은 화자 인증에 사용되는 검증 방법으로 오인식율(False Acceptance Rate, FAR)과 오거부율(False Rejection Rate, FRR)이 같아지는 때의 비율을 의미한다.

Table 1. Details of Encoder architecture (Thin Res Net34).

Model	Encoder (Thin ResNet34)
Conv1 Pool	$[7\times7, 16]$ $[3\times3]$, MaxPool
layer1	$\begin{bmatrix}3\times3, 16\\3\times3, 16\end{bmatrix}\times3$
layer2	$\begin{bmatrix}3\times3, 32\\3\times3, 32\end{bmatrix}\times4$
layer3	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times6$
layer4	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times3$
Pooling	SAP[11]
FC1	$[128\times512]$

4.3 실험 내용 및 결과

Table 2는 인코더 기반의 Xie *et al*.[14] 본 논문의 기반 방법인 Tai *et al*.[3] 그리고 본 논문의 화자 인식 시스템 실험에 대한 동일오류율 결과를 정리한 것 이다. 비교군으로 설정한 방법들은 제안한 시스템과 같은 인코더 구조(ResNet34)와 풀링 방법을 가지고 있고, 동일한 학습 방법을 사용하여 학습하고 성능을 측정하였으므로 제안 방법에 대한 유용성을 공정하게 검증 가능하다. 기반 논문인 Tai *et al*.은 화자 정보 분리를 활용한 방법이 기존의 인코더 방식보다 EER이 상대적 개선 지표 측면에서 12.1 % 더 우수하다는 것을 보였고, 제안된 시스템은 Tai *et al*.보다 9.1 % 더 우수하였다. 이를 통해 본 논문에서 제안하는 방법이 화자 정보와 잔여 정보를 분리하는데 더욱 유용하다는 것을 확인하였다.

Table 3은 본 논문에서 제안하는 시스템의 ablation study 결과를 나타내고 있다. 본 논문은 기존의 오토인코더 구조에서 GAN 구조를 적용하여 정보 분리 효과를 강화하고 있고, GAN에서 판별기가 정보 분리에 미치는 영향에 대해 분석하고 있다. 판별기가 없는 기존 방법의 경우 동일오류율성능이 4.41 %이며, 판별기를 도입하여 GAN의 학습방법과 같이 전체 시스템을 학습할 경우는 동일 오류율이 4.29 %로 성능의 향상이 있음을 확인하였다. 디코더를 이용한 화자 정보 분리 방법은 입력 멜-스펙트럼을 복원하

Table 2. EER results on Voxceleb1 testset. S: Softmax, AM: Angular Margin Softmax. : all experiments are reimplemented.

Model	Network	Loss	EER (%)
Xie *et al.*[12]	E_{spk}	S	5.02
Tai *et al.*[3]	E_{spk}+E_{spk}+D	S	4.41
		AM	3.12
Proposed	E_{spk}+E_{spk}+D+Dis	S	4.01
		AM	2.98

Table 3. Ablation study of proposed system.

Model	Losses			EER (%)
	L_{mse}	L_D	L_C	
Tai *et al.*	√			4.41
Proposed	√	√		4.29
	√	√	√	4.01

는 성능이 정보의 보존에 중요한 역할을 하므로 분류기를 통한 복원 성능 향상이 전체 시스템에서 화자 인식 성능의 향상을 가져왔다.

또한, 다중 작업 판별기로 기존의 분류기의 손실함수 L_D와 다중작업으로 복원된 데이터에서 화자를 분류하는 손실함수 L_C로 학습하여 동일 오류율을 측정하였다. 다중작업 분류기는 디코더의 성능과 더불어 복원되는 멜-스펙트럼이 화자의 정보를 잘 포함하도록 학습한다. 즉, 비록 정보량의 유실이 존재하더라도 복원된 멜 스펙트럼에는 화자의 정보는 유지되며 이는 결과적으로 화자 인코더의 성능을 향상시킬 수 있다.

제안된 시스템의 화자 임베딩의 우수성은 t-SNE[15] 시각화를 통해 확인할 수 있다. Fig. 2는 ablation study에서 사용한 방법들의 화자 임베딩의 분포를 나타내고 있다. 임베딩 공간에서 같은 화자 특성을 지닌 화자 임베딩은 군집하여 각 화자별로 군집하는 모습을 확인할 수 있고, 화자 임베딩 간의 거리가 먼 음성 신호들은 화자의 특성이 매우 다름을 의미한다. 그러므로 임베딩 공간에서 군집화가 이루어진 정도를 보고 발화별 변동성을 알 수 있다. Fig. 2에서 확인할 수 있는 바와 같이 기존 방법에 비해 제안된 방법에서 더욱 군집화한 임베딩 분포를 보여주므로 다른 발화 내용 및 녹음 환경에도 강인한 화자 특징 정보를 추출함을 알 수 있다.

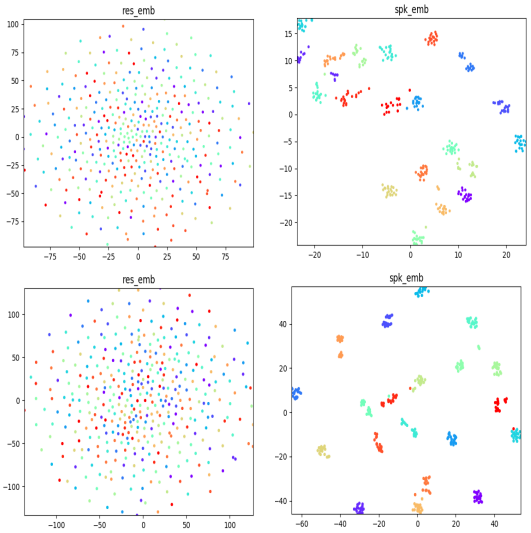

Fig. 2. (Color available online) t-SNE plots of proposed method: samples are extracted from 10 speakers on test set. There are speaker embedding plots on the right side and residual embeddings on the left side. Upper plots are from baseline,[3] and bottom pictures are from proposed system.

V. 결 론

본 논문에서는 화자 인식 성능을 개선하기 위해 정보 분리를 통한 화자 벡터 추출 방법을 제안하였다. 제안 방법은 기존의 인코더 형식의 구조와 다르게 오토-인코더 형식의 프레임워크를 도입하여 화자와 비 화자 정보를 분리하고, 다중 태스크 판별기를 도입함으로써 효과적으로 분류 가능한 화자 벡터를 추출하였다. 실험 결과를 통하여 제안 방법이 기존의 방식들에 비해 향상된 성능을 보임을 확인하였으며, 제안된 각 네트워크가 화자인식 성능에 미치는 결과를 확인하였다.

References

1. 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).
2. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” Proc. IEEE ICASSP. 5329-5333 (2018).
3. T. Jianwei, J. Xiaoqi, H. Qingjia, Z. Weijuan, and Z.

Shengzhi, “SEF-ALDR: A speaker embedding framework via adversarial learning based disentangled representation,” arXiv preprint arXiv:1912.02608 (2020).
4. C. Li, M. Xiaokong, J. Bing, L. Xiangang, Z. Xuewei, L. Xiao, C. Ying, K. Ajay, and Z. Zhenyao, “Deep speaker: an end-to-end neural speaker embedding system,” arXiv preprint arXiv:1705.02304 650 (2017).
5. I. Kim, K. Kim, J. Kim, and C. Choi, “Deep speaker representation using orthogonal decomposition and recombination for speaker verification,” Proc. IEEE ICASSP. 6126-6130 (2019).
6. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in NIPS. 2672-2680 (2014).
7. W. Ding and L. He, “MTGAN: Speaker verification through multitasking triplet generative adversarial networks,” arXiv preprint arXiv: 1803.09059 (2018).
8. Y. Liu, Z. Wang, H. Jin, and I. Wassell, “Multi-task adversarial network for disentangled feature learning.” Proc. IEEE CVPR. 3743-3751 (2018).
9. J. S. Chung, N. Arsha, and A. Zisserman, “Voxceleb2: deep speaker recognition,” arXiv preprint arXiv:1806.05622 (2018).
10. N. Arsha, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” arXiv preprint arXiv:1706.08612 (2017).
11. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE CVPR. 770-778 (2016).
12. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv: 1511.06434 (2015).
13. W. Cai, J. Chen, and M. Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” arXiv preprint arXiv: 1804.05160 (2018).
14. W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” Proc. IEEE ICASSP. 5791-5795 (2019).
15. L. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Machine Learning Research, **9**, 2579-2605 (2008).

저자 약력

▸ 권 유 환 (Yoohwan Kwon)

2017년 2월 : 서울시립대 전기전자컴퓨터공학과 학사
2019년 3월 ~ 현재 : 연세대 전기전자공학과 석사 과정

▸ 정 수 환 (Soo-Whan Chung)

2016년 2월 : 연세대 전기전자공학과 학사
2016년 3월 ~ 현재 : 연세대 전기전자공학과 통합과정

▸ 강 홍 구 (Hong-Goo Kang)

1989년 2월 : 연세대 전기전자공학과 학사
1991년 2월 : 연세대 전기전자공학과 석사
1995년 8월 : 연세대 전기전자공학과 박사
1996년 4월 : AT&T Lab. Senior Technical Staff Member
2002년 9월 ~ 현재 : 연세대 전기전자공학과 교수