한국음향학회지 제40권 제2호 pp. 148~154 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.2 (2021)
https://doi.org/10.7776/ASK.2021.40.2.148

pISSN : 1225-4428
eISSN : 2287-3775

문장 독립 화자 인증을 위한 세그멘트 단위 혼합 계층 심층신경망

Segment unit shuffling layer in deep neural networks for text-independent speaker verification

허정우,[1] 심혜진,[1] 김주호,[1] 유하진[1†]

(Jungwoo Heo,[1] Hye-jin Shim,[1] Ju-ho Kim,[1] and Ha-Jin Yu[1†])

[1]서울시립대학교 컴퓨터과학부

(Received February 15, 2021; accepted March 15, 2021)

초 록: 문장 독립 화자 인증 연구에서는 일반화 성능 향상을 위해 문장 정보와 독립적인 화자 특징을 추출하는 것이 필수적이다. 그렇지만 심층 신경망은 학습 데이터에 의존적이므로, 동일한 시계열 정보를 반복 학습할 경우, 화자 정보를 학습하는 대신 문장 정보에 과적합 될 수 있다. 본 논문에서는 이러한 과적합을 방지하기 위해 시간 축으로 입력층 혹은 은닉층을 분할 및 무작위 재배열하여 시계열 정보의 순서를 뒤섞는 세그멘트 단위 혼합 계층을 제안한다. 세그멘트 단위 혼합 계층은 입력층 뿐만 아니라 은닉층에도 적용이 가능하므로, 입력층에서의 일반화 기법에 비해 효과적이라 알려진 은닉층에서의 일반화 기법으로 활용이 가능하며, 기존의 데이터 증강 방법과 동시에 적용할 수도 있다. 뿐만 아니라, 세그멘트의 단위 크기를 조절하여 혼합의 정도를 조절할 수도 있다. 본 논문에서는 제안한 방법을 적용하여 문장 독립 화자 인증 성능이 개선됨을 확인하였다.
핵심용어: 문장 독립 화자 인증, 심층 신경망, 화자 특징, 혼합 일반화

ABSTRACT: Text-Independent speaker verification needs to extract text-independent speaker embedding to improve generalization performance. However, deep neural networks that depend on training data have the potential to overfit text information instead of learning the speaker information when repeatedly learning from the identical time series. In this paper, to prevent the overfitting, we propose a segment unit shuffling layer that divides and rearranges the input layer or a hidden layer along the time axis, thus mixes the time series information. Since the segment unit shuffling layer can be applied not only to the input layer but also to the hidden layers, it can be used as generalization technique in the hidden layer, which is known to be effective compared to the generalization technique in the input layer, and can be applied simultaneously with data augmentation. In addition, the degree of distortion can be adjusted by adjusting the unit size of the segment. We observe that the performance of text-independent speaker verification is improved compared to the baseline when the proposed segment unit shuffling layer is applied.
Keywords: Text-independent speaker verification, Deep neural network, Speaker embedding, Shuffling generalization

PACS numbers: 43.60.Bf, 43.72.Fx

I. 서 론

화자 인증은 사전에 등록된 발성과 새롭게 입력된 발성 간의 화자 일치 여부를 검증하는 과제이며, 최근에는 딥러닝의 발전으로 심층 신경망을 활용하는 방향으로 활발히 연구되고 있다.[1-5] 문장 독립 화자

†**Corresponding author:** Ha-Jin Yu (hjyu@uos.ac.kr)
School of Computer 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)

인증은 화자 인증의 하위 분류 중 하나로, 발성 문장이 고정되지 않고 자유로운 환경에서 수행된다. 심층 신경망을 활용한 문장 독립 화자 인증 연구에서는 문장 정보와 독립적인 화자 특징을 추출하는 것이 필수적이다.

심층 신경망은 학습 데이터에 의존적이므로, 학습 데이터 과적합으로 인한 평가 성능 저하 문제가 발생할 수 있다. 따라서, 심층 신경망의 일반화 성능 향상을 위한 기법들이 연구되고 있다. 드롭 아웃[6]과 배치 정규화[7]는 다양한 분야에서 적용 가능한 일반화 기법으로, 심층 신경망의 은닉층에 적용할 수 있다. 데이터 증강은 원본 데이터에 왜곡을 적용하여 한정된 학습 데이터의 양을 증폭시켜 일반화 성능을 향상시키는 기법이다. 오디오 도메인에서 적용할 수 있는 데이터 증강 기법으로는 warping,[8] masking,[8] shuffling and mixing[9] 등이 존재한다.

문장 독립 화자 인증에서는 동일한 시계열 정보를 반복 학습할 경우, 심층 신경망이 화자 정보가 풍부한 부분에 집중하도록 학습하는 대신, 문장 전체에 과적합 될 수 있다. 본 논문에서는 이러한 과적합을 방지하여 일반화 성능을 향상시키기 위해 세그멘트 단위 혼합 계층을 제안한다. 제안한 방법은 시간 축으로 나열된 입력층 이전 음향 특징의 프레임들 혹은 은닉층의 노드들을 사전 정의한 크기에 따라 여러 개의 부분 집합인 세그멘트로 분할한 후, 세그멘트 간의 순서를 뒤섞는 방식을 통해 동일한 시계열 정보에 대한 반복 학습을 제한하는 계층을 추가하는 것이다. 세그멘트 단위 혼합 계층은 입력층 뿐만 아니라 은닉층에도 적용이 가능하므로, 입력층에서의 일반화 기법에 비해 효과적으로 알려진 은닉층에서의 일반화 기법[10]으로 활용이 가능하며, 데이터 증강과 동시에 적용할 수 있다는 장점이 있다. 추가로 세그멘트의 단위 크기를 조절하여 혼합의 정도를 조절할 수 있다.

본 논문의 II장에서는 심층 신경망을 활용한 화자 인증 연구와 신경망의 일반화 성능 개선을 목적으로 수행된 연구들을 소개한다. III장에서는 제안하는 기법인 세그멘트 단위 혼합계층에 대하여 설명한다. IV장에서는 실험에 사용한 데이터셋, 베이스라인에 대해 소개하고, 수행한 실험의 결과를 분석한다. 마지막으로, V장에서는 본 논문의 결론과 향후 연구 방향을 기술한다.

II. 관련 연구

2.1 심층 신경망 기반 화자 특징 추출기

화자 특징 추출기는 원시 파형 혹은 음향 특징으로부터 발성의 화자를 구분할 수 있는 특징 벡터인 화자 특징을 추출하는 시스템이다. 최근 심층 신경망의 발전에 따라 심층 신경망 기반 화자 특징 추출기를 사용하는 연구가 활발히 진행중이다.[1-4] 심층 신경망 기반 화자 특징 추출기를 학습시키는 방법에는 d-vector[4], x-vector[11] 시스템 등이 있다. d-vector 시스템은 심층 신경망을 Categorical Cross Entropy(CCE)와 같은 손실함수를 사용하여 사전에 등록된 여러 화자 중의 한 사람을 선택하는 화자 식별을 학습한 뒤, 학습이 완료되면 출력층을 제거한 마지막 은닉층의 출력값을 화자 특징으로 사용하는 방식이다.

2.2 심층 신경망의 일반화 성능 향상 기법

학습 데이터에 의존적인 심층 신경망을 활용한 연구에서는 심층 신경망이 학습 데이터에 과적합 함에 따라 평가 성능이 저하되는 문제가 발생할 수 있다. 이를 방지하기 위해서는 신경망의 과적합을 방지하는 일반화 기법을 적용할 필요가 있다. 심층 신경망의 은닉층에 적용할 수 있는 일반화 기법으로는 드롭 아웃,[6] 배치 정규화[7]등이 있다. 데이터 증강은 원본 데이터를 변형하여 한정된 학습 데이터의 양을 증폭시키는 기법이다. 이는 심층 신경망이 원본 데이터에 과적합 하는 것을 방지하는 효과가 있으며, 다양한 분야에서 효과가 검증된 일반화 기법이다.[8,9,12-14] 오디오 도메인에서 적용할 수 있는 데이터 증강 기법에는 warping,[8] masking,[8] shuffling and mixing[9] 등이 있다.

III. 제안한 세그멘트 단위 혼합 계층

화자 특징 추출기가 동일한 시계열 정보를 반복 학습할 경우, 화자 정보가 풍부한 부분에 집중하도록 학습하는 대신, 문장 정보에 과적합 할 가능성이

있다. 따라서 본 논문에서는 이러한 과적합을 방지하여 일반화 성능을 향상시키는 것을 목표로 세그멘트 단위 혼합 계층을 제안한다. 제안한 기법을 적용할 경우, 동일한 시계열 정보에 대한 반복 학습을 제한하여 과적합을 방지하고 일반화 성능을 향상시키는 데 도움이 될 수 있다.

제안한 세그멘트 단위 혼합 계층의 동작 과정은 Fig. 1과 같다. 세그멘트 단위 혼합 계층에 입력된 시계열 데이터는 분할 과정에서 사전에 정의한 세그멘트 단위 크기로 분할된다. 이후 혼합 과정에서 분할된 세그멘트 단위로 무작위 재배열 되어, 시간 축으로 혼합된 데이터로 출력된다. 예를 들어 총 97개의 프레임을 가진 데이터가 10프레임 단위로 혼합하는 세그멘트 단위 혼합 계층에 입력된 경우, 분할 과정에는 10프레임 집합 9개와 7프레임의 나머지로 분할되고, 혼합 과정에서는 크기가 같은 9개의 집합에 대한 무작위 재배열이 수행된다. 이후 무작위 재배열 과정에서 배제되었던 마지막 7프레임을 무작위 재배열된 데이터의 마지막 프레임 이후에 연결하여 출력한다.

정보를 과도하게 혼합할 경우, 화자 특징을 추출하는데 필요한 정보가 손상되어 화자 특징 추출기의 학습이 정상적으로 진행되지 않을 수 있다. 본 논문에서 제안한 세그멘트 단위 혼합 계층에서는 세그멘트의 단위 크기를 조절하여 혼합의 정도를 조절할 수 있다.

제안한 방법은 화자 특징 추출기의 입력층 혹은 은닉층에 적용할 수 있으므로, 데이터 증강 기법과 비교하여 다음과 같은 두 가지 장점이 있다. 첫 번째로, 은닉층에서 일반화 기법을 수행할 수 있다. Reference [10] 연구에 따르면, 은닉층에서의 일반화 기법이 입력층에서의 일반화 기법에 비해 효과적일 수 있다. 따라서 세그멘트 단위 혼합 계층은 데이터 증강 기법에 비해 효과적인 일반화 기법이 될 수 있다. 두 번째로, 데이터 증강과 동시에 수행할 수 있으며, 해당 계층을 반복 적용할 수 있다. 입력층에서 수행되는 데이터 증강 기법은 화자 특징 추출기의 입력층 이전에 한 번만 적용할 수 있다. 반면 은닉층에서 수행되는 세그멘트 단위 혼합 계층은 드롭 아웃과 같이 화자 특징 추출기의 입력층 이전 뿐만 아니라, 신경망 내부 여러 위치에서 적용할 수 있으며, 은닉층에서 수행되므로 데이터 증강과 동시에 적용하는 것 또한 가능하다.

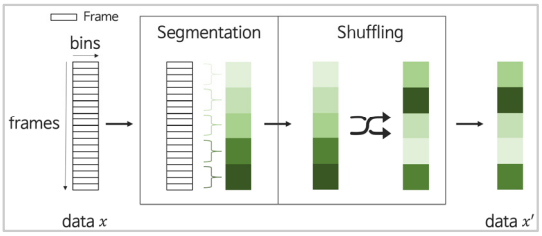


Fig. 1. (Color available online) The process of the segment shuffling layer. It divides the input into small segments of predefined size and shuffles the order of the segments randomly.

IV. 실험 및 결과

4.1 데이터셋

본 논문의 실험에서는 VoxCeleb1 데이터셋을 사용하였다. VoxCeleb1 데이터셋은 전 세계 유명인의 발성을 유튜브에서 수집하여 제작되었으며, 다양한 국적, 연령대, 억양의 발성으로 구성되어 있다. VoxCeleb1 데이터셋은 공식적으로 검증 세트를 제외한 학습 세트와 평가 세트를 제공하고 있다. 신경망의 학습에는 1,211명의 화자로부터 수집한 148,642개의 발성을 사용하였고, 평가에는 학습 과정에 포함되지 않은 40명의 화자로부터 수집한 4,874개의 발성을 사용하였다.

4.2 베이스라인 및 실험 구성

본 논문에서 사용한 베이스라인은 SE ResNet[15] 기반 심층 신경망으로, 구조는 Fig. 2와 같다. Fig. 2의 출력층은 화자 식별층을, 이전의 전 결합층은 d-vector 시스템의 마지막 은닉층을 나타낸다. 베이스라인의 학습은 화자 식별층의 출력값을 활용하여 화자 식별을 학습하는 방식으로 진행하였고, 베이스라인의 평가는 화자 식별층을 제거한 마지막 은닉층의 출력값을 활용하여 화자 검증을 수행하는 방식으로 진행하였다. 베이스라인의 입력은 발성 단위 Z 정규화를 적용한 Log-Melfilterbank 음향 특징을 사용하였으며, 학습에는 이를 무작위로 300프레임 길이로 자른 데

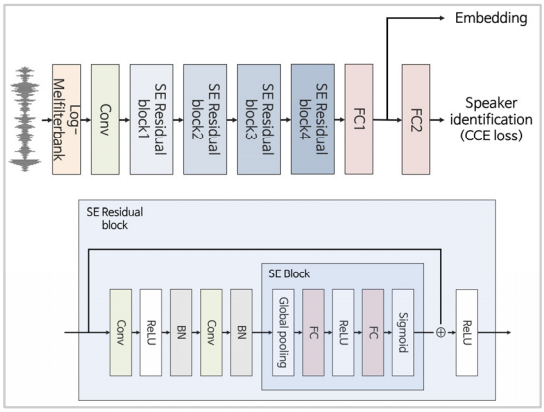


Fig. 2. (Color available online) The baseline structure.

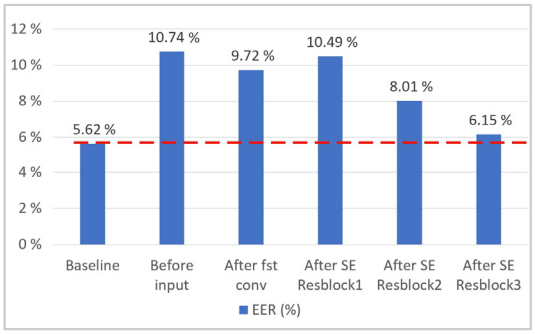


Fig. 3. (Color available online) The variation of the equal error rates according to the position of the shuffling layer. The segment size of shuffling layer is one, and the shuffling layer is activated only in development phase.

Table 1. The variation of the equal error rates according to the shuffling layer's segment size and shuffling layer's position. The columns are segment size of the shuffling layer, and the rows are the position of the shuffling layers. The bold values indicate the superior performances over the baseline and the underlined values indicate the best performance in each rows.

Shuffling layer position	Segment size (Before input: number of frames Hidden layer: number of nodes)				
	5	10	30	50	60
Before input	6.50 %	6.10 %	**5.56 %**	5.65 %	**5.44 %**
After fst conv	6.24 %	5.70 %	5.65 %	**5.58 %**	5.65 %
After res block1	5.85 %	5.64 %	**5.60 %**	5.76 %	5.64 %
After res block2	**5.56 %**	**5.37 %**	**5.42 %**	**5.58 %**	5.64 %
After res block3	**5.45 %**	**5.26 %**	**5.44 %**		

이터를, 평가에는 데이터 전체를 사용하였다. 신경망은 Adam 최적화 알고리즘을 사용하여 학습되었고, 이때 손실 함수는 CCE를 사용하였다. 심층 신경망의 각 층 사이에는 배치 정규화가 적용되었고, 마지막 잔차 블록 이후 Attention pooling을 수행하였다. 300프레임의 음향 특징이 베이스라인의 입력으로 사용되며, 잔차 블록1 이후까지 은닉층의 노드 수는 300, 잔차 블록2 이후는 150, 잔차 블록3 이후는 75이다.

4.3 실험 설계 및 결과

본 논문에서는 세그멘트 단위 혼합의 필요성을 확인한 실험, 제안한 세그멘트 단위 혼합 계층을 활용하여 문장 독립 화자 인증의 성능을 개선한 실험, 제안한 기법을 데이터 증강 기법으로 사용한 경우와 비교한 실험을 수행하였다.

먼저 Fig. 3은 세그멘트 단위 혼합의 필요성, 즉 정보의 혼합의 정도를 조절할 필요성을 확인하기 위해 수행한 실험 결과이다. 해당 실험에서는 세그멘트 단위의 크기가 1인 혼합 계층을 화자 특징 추출기의 입력층 이전, 첫 합성곱 이후, 잔차 블록 사이 중 한 곳에 삽입한 뒤, 삽입 위치에 따른 성능 변화를 확인하였다. Fig. 3의 가로축은 혼합 계층을 삽입한 위치를, 세로축은 해당 위치에 혼합 계층을 삽입했을 때의 동일 오류율을, 그리고 붉은 점선은 베이스라인의 동일 오류율을 나타낸다. 실험 결과, 모든 실험에서 베이스라인 대비 화자 인증 성능이 저하되었음을 확인하였다. 이는 정보를 과도하게 혼합함에 따라 화자 특징을 추출하는데 필요한 정보가 손상되어 화자 특징 추출기의 학습이 정상적으로 진행되지 않음에 따른 결과로 분석할 수 있다. 따라서 본 논문에서는 개별 프레임 혹은 노드 단위 혼합이 아닌 세그멘트 단위 혼합을 이용해 실험을 수행하였다.

Table 1은 제안한 기법인 세그멘트 단위 혼합 계층을 활용하여 문장 독립 화자 인증의 성능을 개선한 실험의 결과이다. 해당 실험에서는 화자 특징 추출기의 입력층 이전, 첫 합성곱 이후, 잔차 블록 사이 중 한 곳에 세그멘트 단위 혼합 계층을 삽입한 뒤, 세그

멘트 단위의 크기에 따른 성능 변화를 측정하였다. 삽입된 세그멘트 단위 혼합 계층은 학습과 평가 과정 모두에서 활성화 되었다. Table 1에서 굵게 표시된 값은 베이스라인의 동일오류율 5.62 %에 비해 우수한 성능을 보인 실험을, 밑줄 표시된 값은 동일 위치에 삽입된 세그멘트 단위 혼합 계층 중 가장 우수한 성능을 보인 실험을 나타낸다. 잔차 블록3 이후 값이 표시되지 않은 칸은 시계열의 길이가 짧아 해당 조건에서 실험을 수행하지 않았음을 나타낸다. 실험 결과, 제안한 세그멘트 단위 혼합 계층을 활용하여 화자 특징 추출기의 성능을 개선할 수 있음을 확인하였다. 구체적으로 모든 위치에서 베이스라인 대비 우수한 성능을 보인 실험이 한 개 이상 존재하였으며, 잔차 블록3 이후 10프레임 단위로 혼합하는 세그멘트 단위 혼합 계층을 삽입하였을 때, 동일 오류율이 5.26 %로 가장 우수한 성능을 보였다. 잔차 블록2와 3이후 혼합 계층을 삽입한 실험에서는, 하나의 경우를 제외하고 모든 실험에서 베이스라인 대비 성능이 개선되었다. 또한 Table 1에 보인 결과에서는 출력층에 가까울수록 작은 단위로 혼합하는 것이 효과적인 것을 확인할 수 있다. 이는 출력층에 가까워질수록 좀 더 구체적인 태스크를 학습하는 딥러닝의 특성 때문인 것으로 보인다. 즉, 출력층에 가까울수록 세부 정보인 문장이 학습될 가능성이 있기 때문에 이를 혼합 계층이 완화해주는 역할을 한 것으로 해석할 수 있다.

Fig. 4의 파란 선은 베이스라인의 학습 과정 CCE Loss를, 보라 선은 제안한 세그멘트 단위 혼합 계층이 적용된 신경망의 학습 과정 CCE Loss를 나타낸다. 아래 그래프는 위 그래프의 y축 범위를 0 ~ 1로 조절한 결과를 나타낸다. Fig. 4를 통해 세그멘트 단위 혼합 계층을 적용할 경우, 베이스라인 대비 CCE Loss가 증가하였으나 대체로 유사한 동향을 나타내고 있음을 확인할 수 있다. 이는 혼합을 수행함에 따라 과제의 난이도가 증가하였으나, 그럼에도 화자 정보가 크게 손상되지 않은 결과로 분석된다.

Table 2는 시간 축으로 데이터를 혼합하는 방식을 제안한 기법과 같이 독립적인 계층으로 사용한 경우와 데이터 증강 기법으로 사용한 경우를 비교하기 위해 수행한 실험의 결과이다. 해당 실험에서는 Fig. 3과 같이 입력층 이전에 발성을 단일 프레임 단위로 혼합할 경우 화자 특징을 추출하는데 필요한 정보가 과도하게 손상될 수 있을 것으로 보고, Table 1의 입력층 이전에 세그멘트 단위 혼합 계층을 삽입한 실험 중 가장 우수한 성능을 보였던 실험을 모방하여 60프레임 단위로 혼합하여 데이터 증강을 수행하였다. Table 2의 “Frame-level data augmentation”은 Fig. 5와 같이 입력 음향 특징을 반으로 분할한 뒤, 절반의 원본 데이터와 절반의 60프레임 단위로 혼합된 데이터를 프레임 축으로 연결하여 입력으로 사용한 실험의 결과이다. 그리고 “Channel-level data augmentation”

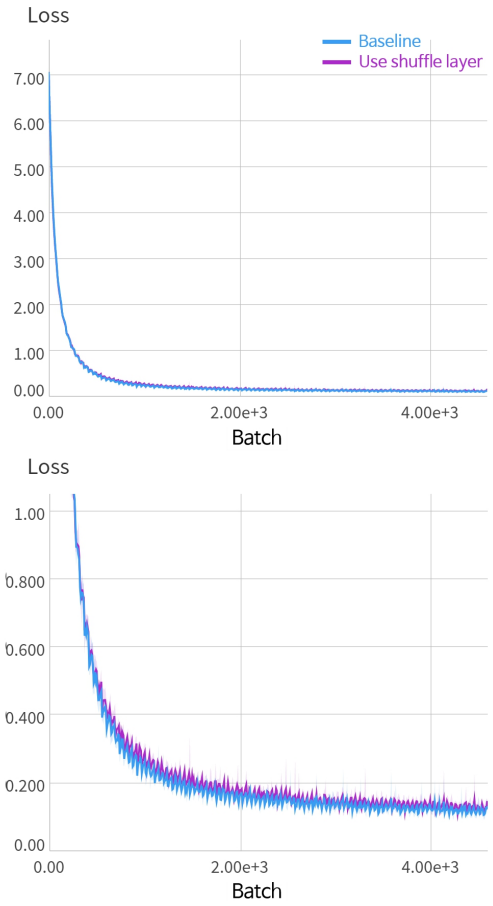


Fig. 4. (Color available online) CCE loss trend according to the application of segment unit shuffle layer.

Table 2. The equal error rates of the segment unit shuffling layer and the shuffling data augmentation.

Segment unit shuffling layer	Frame-level data augmentation	Channel-level data augmentation
5.26 %	5.73 %	5.73 %

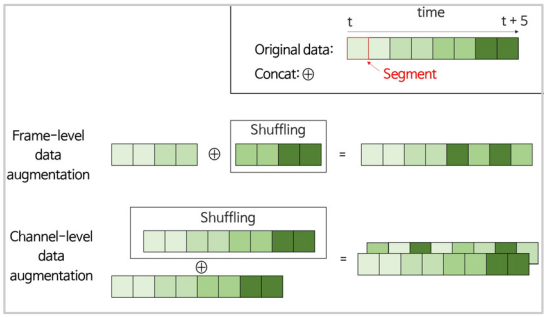


Fig. 5. (Color available online) Augmentation strategy for frame-level data augmentation and channel-level data augmentation.

은 입력 음향 특징 원본에 이를 60프레임 단위로 혼합한 음향 특징을 부가적인 채널로 추가하여 입력으로 사용한 실험의 결과이다. 실험 결과, 혼합을 데이터 증강 기법으로 사용했을 때 두 실험 모두 동일오류율이 5.73 %인 반면, 제안한 혼합 계층을 사용했을 때 동일 오류율이 5.26 %임을 확인하였다. 즉, 입력층에서의 혼합에 비해 은닉층에서의 혼합이 실제로 효과적인 것을 실험적으로 확인하였다.

V. 결 론

본 논문에서는 문장 독립 화자 인증의 일반화 성능을 개선하기 위한 세그멘트 단위 혼합 계층을 제안하였다. 제안한 세그멘트 단위 혼합 계층은 시계열 정보 혼합을 통해 심층 신경망이 동일한 시계열 정보를 반복 학습하지 못하도록 동작하는 계층이다. 이는 심층 신경망이 문장 정보에 과적합 되는 것을 방지하여 일반화 성능을 향상시킬 수 있는 것으로 해석할 수 있다. 제안한 방법은 은닉층에서 데이터를 혼합하기 때문에 데이터 증강과 함께 적용이 가능하다. 본 논문에서는 세그멘트 단위 혼합 계층을 적용할 경우, 문장 독립 화자 인증 성능을 개선할 수 있음을 실험적으로 확인하였다. 향후 연구에서는 세그멘트 단위 혼합 계층을 여러 은닉층에 적용할 뿐만 아니라 제안한 기법과 데이터 증강 기법과 동시에 사용한다면 추가적인 성능 향상을 확인할 수 있을 것으로 기대된다. 또한 상황 식별, 오디오 위변조 탐지 등에서도 세그멘트 단위 혼합 계층이 유효한지 확인하는 연구를 수행할 계획이다.

감사의 글

이 논문은 2020년도 서울시립대학교 교내학술연구비에 의하여 지원되었음.

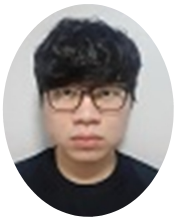
References

1. J. -w. Jung, H.-j. Shim, J.-h. Kim, and H.-J. Yu, “α-feature map scaling for raw waveform speaker verification” (in Korean), J. Acoust. Soc. Kr. **39**, 441-446 (2020).
2. D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmie, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” Proc. IEEE SLT. 165-170 (2016).
3. D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” Proc. Interspeech, 999-1003 (2017).
4. E. Varian, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” Proc. ICASSP. 4080-4084 (2014).
5. S. Shon, H. Tang, and J. R. Glass, “Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model,” Proc. IEEE SLT. 1007-1013 (2018).
6. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” JMLR. **15**, 1929-1958 (2014).
7. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proc. PMLR. 448-456 (2015).
8. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779 (2019).
9. T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, and R. Tachibana, “Shuffling and mixing data aug-mentation for environmental sound classification,” Proc. of the DCASE. 109-113 (2019).
10. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT press, Cambridge, 2016), pp. 236-239.
11. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” Proc. ICASSP. 5329-5333 (2018).
12. Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, “Improved relation classification by deep recurrent

neural networks with data augmentation," arXiv preprint arXiv:1601.03651 (2016).
13. Z. Wu, S. Wang, Y. Qian, and K. Yu, "Data augmentation using variational autoencoder for embedding based speaker verification," Proc. Interspeech, 1163-1167 (2019).
14. L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621 (2017).
15. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. CVPR. 7132-7141 (2018).

저자 약력

▸ 허 정 우 (Jungwoo Heo)

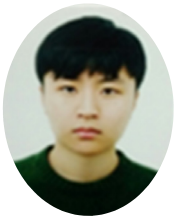


2018년 ~ 현재 : 서울시립대학교 컴퓨터과학부 학사

▸ 심 혜 진 (Hye-jin Shim)

2017년 2월 : 광운대학교 동북아통상학부 학사
2019년 2월 : 서울시립대학교 컴퓨터과학부 석사
2019년 ~ 현재 : 서울시립대학교 산업기술연구소 전임연구원
2020년 ~ 현재 : 서울시립대학교 컴퓨터과학부 박사 과정

▸ 김 주 호 (Ju-ho Kim)



2019년 2월 : 서울시립대학교 컴퓨터과학부 학사
2019년 ~ 현재 : 서울시립대학교 컴퓨터과학부 석사 과정

▸ 유 하 진 (Ha-Jin Yu)

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