Journal of the Korean Data & Information Science Society
2023, **34**(2), 203–214

http://dx.doi.org/10.7465/jkdi.2023.34.2.203
한국데이터정보과학회지

음성 위조 탐지를 위한 2단계 학습 모형 연구[†]

강태인[1] · 곽일엽[2]

[12]중앙대학교 응용통계학과

접수 2023년 1월 12일, 수정 2023년 2월 13일, 게재확정 2023년 3월 7일

요 약

본 연구에서는 음성 위조 탐지 문제에 있어서 딥러닝 모형들의 2단계 학습 모형에 대한 모델과 성능 결과를 제시하고자한다. 음성 위조 탐지는 실제 음성과 원래 음성과 다른 환경에서 복제된 위조 음성을 구별하는 과제이다. 음성비서와 같이 화자의 식별이 보안과 직접적으로 연관되는 문제들에서 음성 위조 탐지의 필요성이 커지고 있다. 제시하는 2단계 학습 모형은 Automatic Speaker Verification Spoofing (ASVSpoof) 2019 대회 LA 데이터 셋으로 연구된 여러 단일 음성 모형들의 임베딩 벡터들을 가져오고, 이를 합쳐서 새로운 피쳐로 정의한 후, 해당 피쳐에 딥러닝 네트워크를 구축하여 모형을 만들어 내는 방식이다. 다수의 모형들을 통해 결과를 도출한다는 면에서 유사성이 있는 기존 앙상블 기법들과 비교를 위해 음성위조 탐지문제 LA 데이터에 있어서 우수한 성능을 가진 단일 모형들을 이용하여 비교 분석한 결과를 살펴보았다. 여러 모형의 임베딩 조합으로 진행된 2단계 학습 모형은 Equal Error Rate (EER) 0.26 (%) 을 달성했다. 이는 앙상블 기법인 Voting의 최고 성능인 0.60 (%) 보다 0.34 (%p) 향상된 결과이며 단일모델 최고 성능 0.83 (%)과 비교해 0.57 (%p) 향상된 결과이다. 음성 위조 탐지 모형에서, 2단계 학습모형의 기초적인 모형을 제시했다는 것이 의미가 있으며 구조를 좀 더 고도화 시키는 후속 연구로 발전 시킬 수 있을 것이다.

주요용어: 딥러닝, 2단계 학습, 음성 위조 탐지, 임베딩.

1. 서론

우리는 일상 속에서 여러 소리를 듣고 구분할 수 있다. 그 중에서 목소리, 음성은 한 개인을 나타내는 소리로서, 우리에게 개인을 특정지을 수 있는 지표가 된다. 이러한 특성을 활용해 많은 딥러닝 연구에서도 음성을 통한 개인 식별 모형을 만들어내고 있다. 이를 발전시켜 여러 장치 등에 삽입하여 실생활에 유용한 기기를 생산하는데, 특히 음성비서의 경우 화자의 구별이 중요한 문제로 떠오르고 있다. 일례로, TV에서 나온 소리가 음성비서로 전달 되어 원하지 않은 품목이 온라인 커머스를 통해 구매가 된 사례가 있어 일각에서는 Automatic Speaker Verification (ASV) 시스템이 해킹에 취약하다는 의견이 나오고 있다. 사례에서는 의도적인 해킹 시도가 아니었지만 악의적인 목적을 가지고 해킹을 시도하고, 지속된다면 음성 보안의 신뢰가 떨어져 음성 보안을 활용한 기술의 확장성이 현저히 줄어들 것이다. 해킹 시도에 대한 사전 차단을 위해, 음성 위조 해킹 시도를 탐지할 수 있는 모형에 대한 연구가 필요하다.

학계에서는 음성 위조 공격 위협 차단에 대한 필요에 대응하여 여러 연구를 진행하고 있다. 그 중 Automatic Speaker Verification Spoofing (ASVSpoof) Challenge는 2015년부터 진행된 대회로, ASV

[†] 이 논문은 2022년도 중앙대학교 연구장학기금 지원에 의한 것임.
[1] (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 석사과정.
[2] 교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 조교수.
E-mail: ikwak2@cau.ac.kr

시스템의 보안 강화를 위해 실제 음성과, 실제 음성을 재생산한 위조 음성을 구별하는 모형을 평가하는 대회다 (Nautsch 등, 2021). 2015년 대회에서는 음성 변환, 합성과 같은 위조음성 데이터셋으로 음성 위조를 탐지 했고, 2017년에는 음성의 재생환경, 재생장치의 조건을 달리한 리플레이 음성에 대한 탐지를 중점으로 평가했다. 본 연구에서 사용한 2019년 대회의 데이터 셋은 2015년 이후의 4년 간 발전된 Text To Speech (TTS), Voice Conversion (VC), Voice Synthesis (VS)등의 기술이 추가된 위조 음성을 포함하고 있다. 2017년과 비교하면 부족했던 재생 환경 제어를 보완하여 다양한 실제 재생 장치와 시뮬레이션으로 채택된 제어된 재생 환경 제어를 통해 더 정교한 리플레이 음성을 만들었다.

머신러닝 및 딥러닝에서 주로 사용하는 앙상블 기법은 여러 분류기의 예측 결과들을 결합하여 단일 분류기 보다 높은 예측 성능을 얻는 것을 목표로 한다 (Dietterich, 2000). 테이블 데이터와 같은 정형 데이터의 분류 및 예측문제에 있어서는 머신러닝 기법들이 두드러진 성능을 나타내고 있으며, 이 때 모형의 앙상블이 성능 향상에 도움을 주고 있다 (Choi 와 Lim, 2013; Jeong 등, 2017). 주로 사용하는 앙상블 기법에는 크게 Voting, Stacking, Bagging, Boosting이 있다. Voting과 Stacking은 여러 알고리즘을 결합하여 성능을 내고, Bagging과 Boosting은 단일 알고리즘에서 샘플링을 달리 해 만든 여러 모형을 결합하는 방식으로 성능을 향상 시킨다. 하지만 Voting, Stacking, Bagging, Boosting 방법들은 각 모형들의 결과 확률정보만을 이용하게 되는데, 딥러닝 모형은 많은 경우 결과를 내기 전 단의 마지막 Embedding 층이 문제 해결에 주요한 feature 를 나타내도록 학습되는 경향이 있다. 최근 딥러닝 모형의 앙상블에서도 이러한 Embedding을 활용하기 위해 전이학습을 활용한 시도들이 이루어 지고 있다 (Heo 등, 2021).

본 연구에서는 딥러닝의 경우 마지막 Embedding 층의 더 많은 정보를 이용하면 더 좋은 결과를 얻을 수 있을 것이라는 아이디어를 떠올리게 되었고, 이를 음성위조 탐지 문제의 딥러닝 모형들에 적용하여 해당 아이디어의 적합성에 대해 살펴보고자 하였다. 음성위조 탐지 문제에 있어서 여러 딥러닝 모형들을 적용해보고 앙상블 방법과 제안된 2단계 학습모형을 수행해 보며 이의 결과를 비교하여 제안된 모형의 우수성을 보이고자 한다.

2. 연구 방법

여러 단일 모형들을 이용해 성능을 개선하는 기존의 앙상블 기법들을 기술하고 본 연구에서 제안하는 2단계 학습 모형에 대해 설명한다. 또한 실험에 사용된 음성 위조 탐지 단일 모형들 (AASIST, LCNN, ResMax, OFD) 의 각각의 구조와 특징, 사용한 임베딩층에 대한 정보를 제시한다 (Jung 등, 2022; Wu 등, 2018; Kwak 등, 2021; Choi 등, 2022). 단일 모형들의 성능을 기술하고 단일 모형들의 여러 조합으로 기존 앙상블 방법과 2단계 학습 모형을 활용해 결과를 제시하며 차이를 비교한다.

2.1. 앙상블

앙상블은 단일, 혹은 여러 알고리즘을 반복하거나 결합하여 더 높은 성능을 내기 위한 기법이다. 자주 쓰이는 앙상블 방법은 Voting, Stacking, Bagging, Boosting이 있고 Voting과 Stacking은 여러 알고리즘의 결합으로 사용되며 Bagging, Boosting은 표본의 재추출로 단일 알고리즘을 반복하여 분류기를 만들어 수행된다 (Breiman, 1996; Freund, 1996; Parhami, 1994; Syarif, 2012).

Voting의 종류는 Hard Voting과 Soft Voting이 있다. Hard Voting은 여러 단일 모형들의 분류 결과 중 다수의 모형들이 분류한 결과로 각 데이터의 최종 결과를 정하는 방식이다. Soft Voting은 단일 모형들이 구한 결과의 분류 확률을 평균 내어 최종 분류 확률을 결정한다. Stacking의 주요한 특징으로 Meta Learner가 있다. 여러 개의 단일 모형들이 예측한 확률값을 쌓아 Meta Learner의 학습 데이터

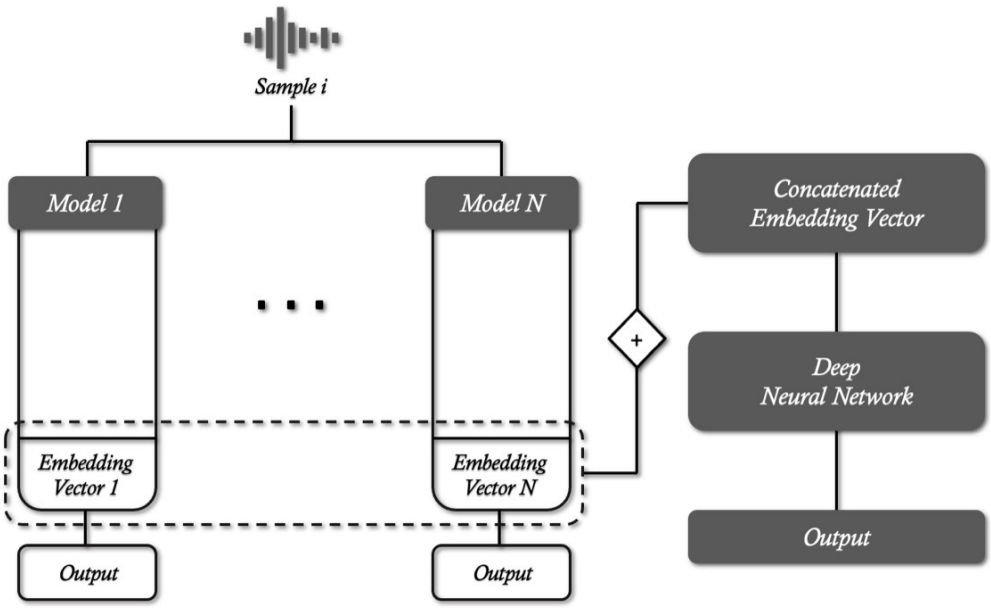


Figure 2.1 Overview of 2 stage training system

셋으로 사용하며 이 데이터셋으로 Meta Learner를 학습해 결과를 도출한다. 과적합을 방지하기 위해 K-Folds Cross Validation을 주로 사용한다.

Bagging은 Bootstrap을 통해 복원 추출한 데이터를 사용하여 단일 알고리즘으로 여러 약한 분류기들을 생성해 내고, 약한 분류기들의 결과로 Voting하여 최종 결과를 결정한다. Boosting은 Bagging과 비슷하지만 다른 점으로 학습을 순차적으로 하는 특징이 있다. 반복을 거듭할 때마다, 잘못 분류한 샘플에 가중치를 주어 이전 반복의 분류기와 함께 새로운 분류기로 분류 결과를 만든다. 반복이 종료되면 생성된 모든 분류기에 대한 결과를 Voting해 마지막 결과값을 결정한다.

2.2. 임베딩층을 활용한 2단계 학습모형

딥러닝은 정형데이터보다 비정형 데이터에서 높은 성능을 내는 경우가 많으며 머신러닝 방법과 학습 과정에 차이가 있다. 또한, 딥러닝 모형에서는 여러 층들을 거치는 과정이 보다 좋은 feature를 찾아주는 역할을 할다. 따라서, 기존의 앙상블 모형에서처럼 단순히 결과만 이용하는 것보다는 여러 층들을 거친 Embedding 층 자체를 이용하면 보다 많은 정보를 이용할 수 있으므로 성능 향상을 기대할 수 있을 것이다. 따라서 여러 단일 모형들의 임베딩을 가져와 Neural Network의 feature로써 학습해 높은 성능을 달성하는 임베딩층을 활용한 2단계 학습모형을 실험해 보고자 한다.

Figure 2.1에서 제안하는 임베딩층을 활용한 2단계 학습모형은 여러 단일 알고리즘의 결합으로 높은 성능을 내는 데에 목적이 있다. 사용된 단일 모형들은 ASVSpoof2019 Challenge의 LA 데이터셋을 사용해 개발된 음성 위조 탐지 모형이며 AASIST, LCNN, ResMax, OFD 총 네 가지 모형을 사용하였다. 먼저 각 모형들을 데이터셋에 대해 충분히 학습 시킨 후 최고 성능의 모델 가중치를 저장한다. 저장된 가중치에서 마지막 분류 단계 층을 없애고 직전 단계 층까지의 가중치를 사용하는데, 음성 샘플이 가중치를 통과했을 때 나오는 벡터를 임베딩이라고 정의한다. 그리고 모든 음성 샘플에 대해서 저장된 각각

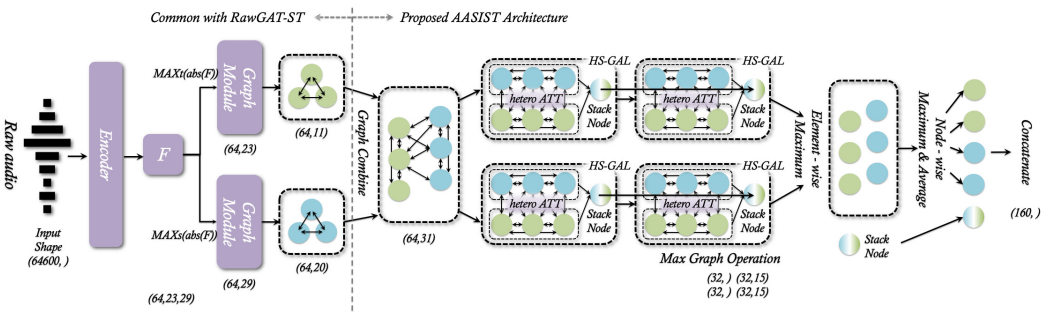


Figure 2.2 AASIST model

의 단일 모형 가중치에 통과 시켜 임베딩을 각출하고 저장한다. 단일 모형 임베딩들의 조합에 따라 결합된 임베딩 벡터의 길이가 달라지게 되고 결합된 임베딩 벡터를 Neural Network의 피쳐로 사용하여 학습한다. Neural Network의 구조로는 정해진 규칙이 없으며 본 연구에서는 기초적인 모형으로 간단한 DNN 모형을 제시해 분석했다.

2.3. 모형 구조

다음으로는 2단계 학습에 활용한 음성 위조 탐지 단일 모형들 (AASIST, LCNN, ResMax, OFD)에 대해 기술하였다.

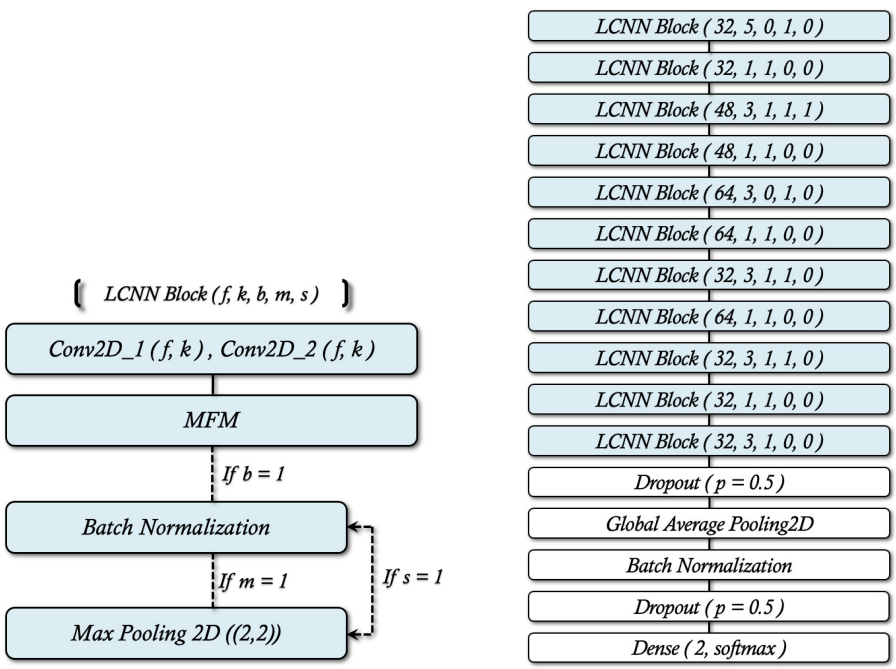
2.3.1. AASIST

Jung 등 (2022) 이 제안한 AASIST 모형은 Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks의 약자로 Figure 2.2 에 기술된 형태의 구조를 가진다. AASIST 모델은 Tak 등 (2021) 에서 제안된 RawGAT-ST 모델을 발전시킨 모형이다. RawGAT-ST 와 공통적인 부분은 Raw Data를 입력으로 넣어 Residual Block 6개로 구성된 RawNet2-based Encoder로 고차원 representation을 추출하고 이를 주파수 축, 시간 축을 기준으로 Graph attention network와 Graph Pooling으로 구성된 Graph Module에 각각 넣어 출력 graph 를 얻는다 (Jung 등, 2020; Tak 등, 2021; Velickovic 등, 2017; Lee 등 2019). AASIST에서 새롭게 제안하는 구조에서는 주파수 축의 출력 graph, 시간 축의 출력 graph를 Graph Combination으로 합쳐주어 통합된 출력 graph로 만들고 Heterogeneous Stacking Graph Attention Layer (HS-GAL)으로 stack node를 여러 개 만들어준 후 Max Graph Operation (MGO)으로 element-wise maximum을 출력한다. 뽑힌 출력값의 node-wise maximum, average와 뽑힌 출력값의 stack node를 concatenate하여 마지막 Linear layer로 최종 출력값을 생성한다. AASIST 단일 모형의 성능은 EER 0.83 (%) 이며, 이는 단일 모형으로 음성 위조 탐지 문제의 LA데이터에 대한 세계 최고 수준의 성능이다. 연구에서 사용한 AASIST의 임베딩은 마지막 Linear layer 전 concatenate된 벡터이며 길이는 160이다.

2.3.2. LCNN

Figure 2.3은 Light CNN-9 모델에서 몇 개의 layer를 추가한 Light-CNN (LCNN) 모델이다 (Wu 등, 2018). 32, 48, 64, 32, 32, 32 Convolution channel size와 Network-in Network (NIN) layer구조를 32, 48, 64, 64, 32의 size로 사용하여, 6개의 Convolution layer와 5개의 NIN layer를 반복하는

모델로 사전 학습 model을 만들었다 (Wu 등, 2018). Light CNN-9 모델에서와 같이 첫 번째 Convolution layer의 kernel size는 5로, 나머지 Convolution layer는 3으로 설정되어있다. 첫 번째, 세 번째 Convolution layer를 제외하고는 Batch Normalization을 적용하고 모든 NIN 레이어 뒤에는 Batch Normalization이 온다. Light CNN-9 모델에 정의된 Fully Connected layer를 사용하는 대신 Global Average Pooling layer, Batch Normalization 및 0.5의 확률로Dropout layer을 사용했다. LCNN 단일 모형의 성능은 EER 3.14 (%) 이다. 본 연구에서는 마지막 Linear layer 전 길이는 32의 임베딩 벡터를 사용했다.

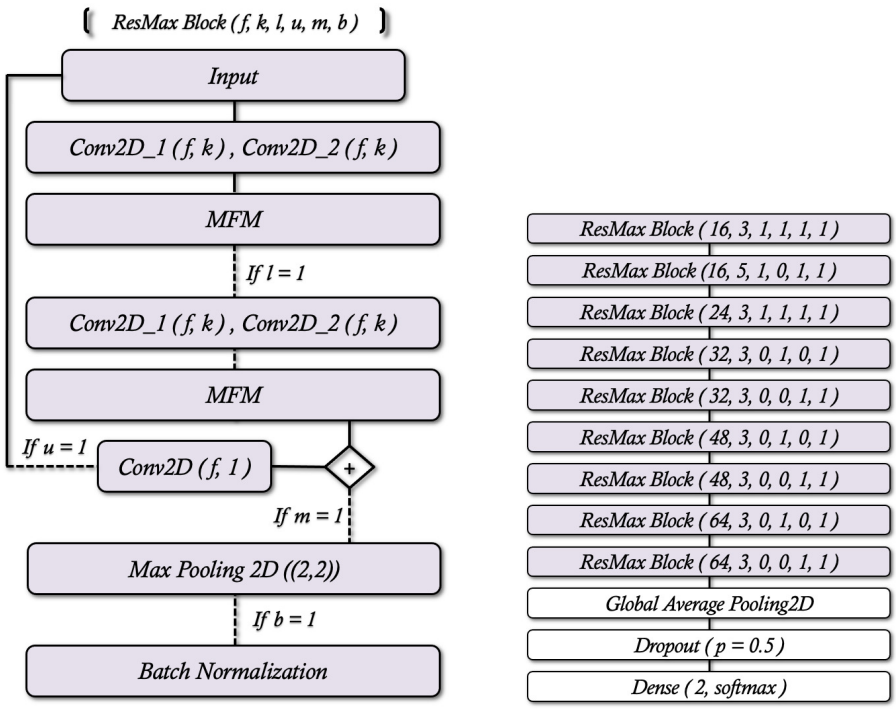


(a) LCNN block (b) LCNN architecture

Figure 2.3 LCNN model

2.3.3. ResMax

Kwak 등 (2021) 이 제안한 Residual Network with Max Feature Map (RexMax) 모델은 9개의 ResMax Block과 Global Average Pooling, Dropout, Linear Layer 으로 구성된다 (Figure 2.4 참고). RexMax Block은 Convolution layer 두 개에 같은 Input을 넣고 두 출력값을 Max Feature Map (MFM)에 통과시켜 Element-wise maximum operation을 수행한다. 그 다음 RexMax Block의 parameter l으로 이 작업을 한 번 더 수행할 지 결정한다. 그리고 입력에 대한 Residual을 차원을 맞춰 더해주고 RexMax의 다른 parameter m과 b로 Max Pooling을 적용할 지, Batch Normalization을 적용할 지 결정을 해준다. 마지막 Dropout은 0.5의 확률로 수행된다. ResMax 단일 모형의 성능은 EER 2.19 (%) 이다. 연구에 사용한 ResMax의 임베딩은 Global Average Pooling 전 마지막 벡터이며 길이는 64이다.



(a) Resmax block (b) Resmax architecture

Figure 2.4 Resmax model

2.3.4. OFD

Choi 등 (2022) 에서 제안된 Overlapped Frequency Distributed (OFD) 모델도 실험에 사용됐다. Figure 2.5은 OFD를 구성하는 OFD Block의 구조이며, OFD는 2개의 Stream이 있는 6개의 OFD Block으로 구성된다. 첫 번째 Stream은 주파수와 관련된 feature를 학습하는 것이다. 주파수 축을 따라 X구역을 n개로 나누고 각 n개 구역의 경계선과 중첩되게 Y구역을 $n-1$개의 구역으로 나눈다. 나눈 구역들을 kernel size 1의 Convolution layer, Batch Normalization, ReLU, 다시 kernel size 1의 Convolution layer, Batch Normalization 을 지나는 함수 f 를 거 친 후 element-wise maximum operation으로 겹치는 $f(x)$, $f(Y)$ 중 높은 요소를 취한다. 두 번째 Stream은 시간적 feature를 학습한다. 입력은 주파수 축을 따라 평균을 내고 Channel by Time 크기의 feature map을 만들고 Dilation 4의 Depthwise Convolution layer를 거친 후 Batch Normalization, Swish activation, point-wise Convolution, ReLU, spatial Dropout을 지나 시간에 대한 feature를 학습하게 된다. OFD 단일 모형의 성능은 EER 5.60 (%) 이다. 본 연구에서는 마지막 Linear layer 전 임베딩 벡터를 사용하는데, 길이는 128이다

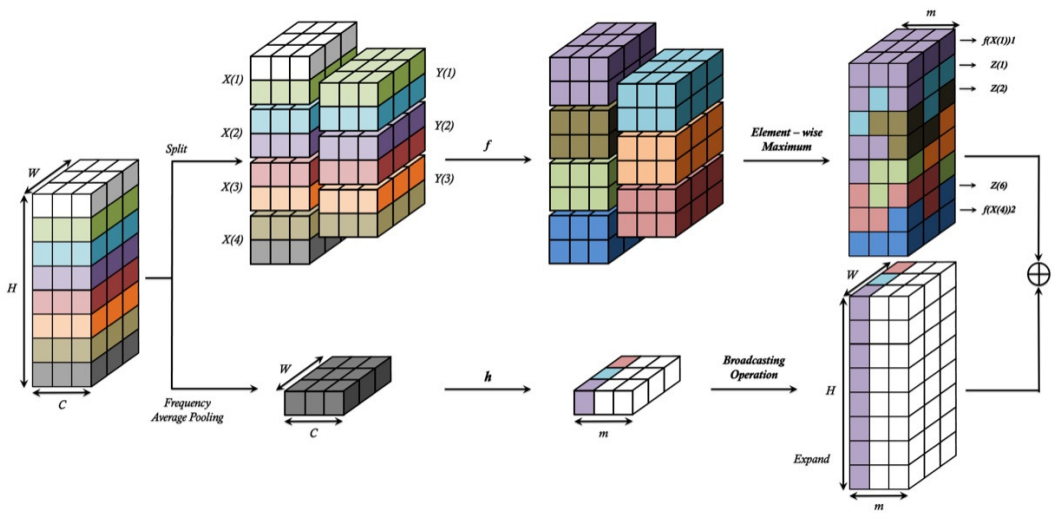


Figure 2.5 Overview of ofd block

3. 실험

3.1. 실험 설정

모든 모델은 NVIDIA Quadro RTX 8000 (48GB) GPU 및 INTEL i9-9900K CPU와 함께 Pytorch 버전 1.8.1, CUDA 버전 10.2를 사용하여 훈련된다. Adam 옵티마이저를 사용하여 학습하며, 배치 사이즈는 24로, Epoch는 50으로 설정하여 가장 좋은 모델을 사용해 검증했다. 실험에 사용한 각 모형들은 모형들이 제안된 페이퍼에 모두 LA 데이터에 대한 모형 학습 부분이 있었고, 모형들 별로 LA 데이터에 최적화된 하이퍼파라미터들을 그대로 가져와서 학습에 사용하였다. 평가지표로는 Equal error rate (EER)을 사용했다. EER은 생체인식 시스템의 성능 척도로서 자주 사용되며 오인식률 (False acceptance rate, FAR) 과 오거부률 (False rejection rate, FRR) 이 같아질 때의 값을 의미한다. 모형으로부터 음성위조에 대해 판별하는 threshold를 조절하면, FAR 과 FRR 은 Trade-off 관계에 있다. Threshold를 조절해서 FAR과 FRR이 가장 비슷해지는 시점에서의 FAR과 FRR의 평균값이 EER로 정의된다.

3.2. ASVSpoof2019 LA 데이터

Automatic Speaker Verification Spoofing (ASVSpoof) Challenge는 ASV 시스템의 취약점인 위조 음성 공격에 대한 대비로, ASV 시스템의 보안을 강화하기 위해 음성 위조 탐지 모형의 성능을 겨루는 대회이며 2015년, 2017년, 2019년, 2021년 등 매 2년마다 대회가 개최된다. 어떤 위조 음성 공격을 가정할 것인가가 매 대회 데이터셋의 특징이 된다. 2015년도 대회에서는 2015년도까지의 Text To Speech (TTS), Voice Conversion (VC), Voice Synthesis (VS)의 기술을 활용해 실제 음성을 변환 기술로 변환하여 위조 음성을 데이터 셋으로 제공했다. 2017년도 대회는 2015년도와 위조 음성을 다른 방식으로 생산했다. 음성 변환 기법이 아닌 실제 음성을 어떤 장치, 어떤 제어 환경에서 리플레이 했느냐를 위조 음성 생산 방식으로 택했다. 2019년 대회 데이터셋은 2015년도와 2017년도의 상위 단계로,

Table 3.1 Data description for asvspoof 2019 logical access (LA) dataset

Subsets	Male	Female	Bona fide	Spoof
Training set	8	12	2,580	22,800
Development set	8	12	2,548	22,296
Evaluation set	30	37	7,355	63,882

Table 4.1 Single classifier performance

Number	Classifier	EER (%)
1	AASIST	0.83
2	ResMax	2.19
3	LCNN	3.14
4	OFD	5.60

2015년도에서 4년 동안 발전한 음성을 변환하는 기법들을 적용해 ASV의 신뢰성에 대한 실질적인 위협이 되는 시나리오를 Logical Access (LA) DataSet으로 정했다. 또한, 2017년도 대회에서 대회를 진행한 후 대회측에서 다소 통제되지 않은 환경에서 음성 재현되어 복잡한 노이즈로 인해 결과를 분석하기가 다소 어려웠다는 점을 고려하여, 제어된 환경에서 생성된 음향 및 음성 재현을 통해 개선된 음성 시뮬레이션을 위조 음성으로 채택하여 센서, 음향 환경 수준에서 위조된 음성 재생 공격들을 Physical Access (PA) DataSet으로 정했다.

본 연구에서 활용한 DataSet은 ASVSpoof 2019 Challenge의 Logical Access (LA) DataSet이다. ASVspoof 2019 데이터베이스는 VCTK2라는 표준 다중 화자 음성 합성 데이터베이스를 기반으로 생산되었고 실제 음성 (Bona fide)은 107명의 화자 (남성 46명, 여성 61명) 로부터 배경 소음 없이 수집되었다. 위조 음성은 다양한 음성 위조 알고리즘을 사용하여 실제 데이터에서 생성된다. 음성 변환 또는 TTS에 사용되는 대상 화자와 관련하여 Training, Development, Evaluation Set 간에 겹치는 부분이 없고 각 하위 데이터에 사용된 화자 및 실제 음성, 위조 음성의 개수는 Table 3.1에서 확인 할 수 있다.

4. 실험결과

4.1. 임베딩층을 활용한 2단계 학습 모형 성능 비교

Table 4.1은 단일 모형의 음성 위조 탐지 결과를 보여준다. 네 개의 모형 AASIST, ResMax, LCNN, OFD 순으로 성능 척도 Equal Error Rate (EER)가 높으며 각각 0.83, 3.14, 2.19, 5.60 (%) 이다. Table 4.2는 Meta Learner 역할을 하는 Neural Network인 DNN의 기초적인 모형 2가지를 여러 조합의 임베딩 결합 피쳐에 학습시킨 결과이다. DNN의 모형 구조로 1개의 Linear Layer, 4개의 Linear Layer를 비교 실험 하였다. 결과는 4개의 Linear Layer로 DNN을 구성했을 때, AASIST와 LCNN, OFD의 임베딩을 합하여 피쳐로 학습시킨 2단계 학습모형이 0.26 (%)의 EER로 가장 높은 성능을 보였다. 마찬가지로 같은 임베딩 조합을 피쳐로 1개의 Linear Layer를 가지는 DNN을 학습시킨 2단계 학습모형이 EER 0.31 (%)로 다음으로 높은 성능을 차지했다.

4.2. 앙상블 방법과 성능 비교

기존 앙상블 방법들보다 더 많은 정보를 쓴다는 면에서 제안된 2단계 학습모형을 기존 앙상들 모형들과 같은 위상에서 비교하기에는 불공평한 면도 있지만, 여러 단일모형들의 결과를 종합하여 모형을 만

Table 4.2 Embedding ensemble performance

Number	Ensemble Embedding	EER (%)	
		DNN	
		1 Linear	4 Linear
1	AASIST + LCNN	0.44	0.38
2	AASIST + ResMax	0.39	0.41
3	AASIST + OFD	0.76	0.54
4	AASIST + LCNN + ResMax	0.35	0.33
5	AASIST + LCNN + OFD	**0.31**	**0.26**
6	AASIST + ResMax + OFD	0.40	0.34
7	AASIST + LCNN + ResMax + OFD	0.33	0.35

Table 4.3 Soft voting ensemble performance

Number	Ensemble Probability	EER (%)
1	AASIST + LCNN	0.71
2	AASIST + ResMax	0.73
3	AASIST + OFD	0.78
4	AASIST + LCNN + ResMax	0.65
5	AASIST + LCNN + OFD	0.69
6	AASIST + ResMax + OFD	0.69
7	AASIST + LCNN + ResMax + OFD	**0.60**

든다는 점에 유사성이 있어 제안된 방법과 앙상블 결과들을 비교하였다. Table 4.3에서는 여러 단일 모형의 확률 값을 조합하여 Soft Voting의 결과를 확인했다. 모형의 수를 늘릴수록 Voting의 결과가 높게 나왔고 4가지 모형 모두의 확률 값을 Voting한 결과가 0.60 (%)의 EER으로 성능이 제일 높았다. Table 4.2와 Table 4.3을 비교하여 2단계 학습모형과 앙상블의 효과 차이를 본다면, 같은 조합의 경우이더라도 2단계 학습모형이 최소 0.02 (%p), 최대 0.43 (%p) 더 높은 결과를 가지는 것을 확인 할 수 있었다.

5. 결론

본 연구에서는 음성 위조 탐지 문제에서 딥러닝 모형들에 대한 임베딩층을 활용한 2단계 학습모형을 제시하고 딥러닝에 사용할 수 있는 기존의 앙상블 방법들과 비교하였다. 임베딩층을 활용한 2단계 학습모형은 여러 단일 모형의 학습된 파라미터로 각각의 임베딩 피쳐들을 출력해 결합함으로써 Meta Learner역할을 하는 Deep Neural Network (DNN)에 입력으로 들어가 DNN을 학습시킨 후 최종 결과를 도출한다. 실험을 위해 ASVSpoof2019 Challenge Logical Access (LA) DataSet을 활용하였고, 이는 2019년까지의 최신 음성 위조 변환 기술을 적용해 위조 음성을 만들어 낸 데이터이다. Training, Development, Evaluation의 하위 DataSet으로 학습 및 평가를 실시하였으며 각 하위 집합에서 중복되는 화자의 음성은 없다. 2단계 학습에 필요한 임베딩을 추출하기 위해, 사전학습된 AASIST, LCNN, ResMax, OFD 모형을 활용하였고, AASIST의 단일 모형 성능이 가장 높기 때문에 AASIST 모형의 임베딩은 모든 임베딩 결합 조합에 포함되었다. AASIST의 임베딩을 포함하는 모든 임베딩 결합을 입력 피쳐로 사용해 간단한 DNN 모형 2가지 (1 Linear Layer, 4 Linear Layer)에 대해 비교실험을 하였다. 이를 기존의 앙상블 방법과도 비교하기 위해 단일 모형들의 확률값의 조합으로 Soft Voting을 적용하여 성능을 살펴보았다.

DNN 모형의 비교 결과는 AASIST,LCNN,OFD 모형의 임베딩 결합을 입력 피쳐로 사용할 때 4 Lin-

ear Layer의 DNN 구조를 학습시키면 가장 높은 성능을 얻을 수 있었다. 이를 Soft Voting을 사용한 것과 비교 했을 때, Soft Voting에서 최고 성능을 보여준 4모형 모두의 앙상블 결과보다 딥러닝의 임베딩을 활용한 2단계 학습 모형이 더 좋은 결과를 보여주는 것을 확인했다. 따라서 음성 위조 탐지 문제에서 본 논문이 제시하는 딥러닝 앙상블의 효과가 뛰어나다는 결론을 내릴 수 있다. 또한 본 연구에서 DNN의 구조가 고도화 됨에 따라 성능의 차이가 나는 결과도 얻을 수 있었는데, 이것은 기초적 모형보다 발전된 모형을 사용했을 때에 더 나은 연구 결과를 얻을 수 있을 가능성을 내포하고 있다. 향후 연구에서는 본 연구에서 자세히 살펴보지 못한 Neural Network의 다양한 변화를 통해 2단계 학습 모형의 발전 가능성을 살펴 보고, 음성 위조 탐지에 적절한 Softmax 방법을 적용해 고도화된 2단계 학습 모형의 구조를 연구할 예정이다. 또한, LA 데이터 뿐만 아니라 PA 데이터에 대한 확장, 그리고 단순히 음성 위조 탐지 문제에만 국한하지 않고 다양한 데이터와 문제들에 대한 연구를 진행한다면 임베딩을 활용한 2단계 학습 모형의 일반화 가능성이 높아질 것으로 기대된다.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Choi, H. N. and Lim, D. H. (2013). Bankruptcy prediction using ensemble SVM model. *Journal of the Korean Data & Information Science Society*, **24**, 1113-1125.

Choi, S., Kwak, I. and Oh, S. (2022). Overlapped frequency-distributed network: frequency-aware voice spoofing countermeasure. *Proc. Interspeech*, 3558-3562, Incheon, Korea.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15. Springer, Berlin, Heidelberg.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Icml*, **96**, 148-156.

Heo, J. H., I, S. B., Yang, W. H. and Lim, D. H. (2021). Transfer learning-based ensemble deep learning for image classification of COVID-19 patients *Journal of the Korean Data & Information Science Society*, **32**, 1219-1235.

Jeong, J., Cha, S., Kim, M., Kim, G., Lim, Y. J. and Lee, K. E. (2017). Drought index forecast using ensemble learning. *Journal of the Korean Data & Information Science Society*, **28**, 1125-1132.

Jung, J., Heo, H. S., Tak, H., Shim, H. , Chung, J. S., Lee, B. , Yu, H. and Evans, N. (2022). Aasist: audio anti-spoofing using integrated spectro-temporal graph attention networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6367–6371, Singapore, Singapore.

Jung, J., Kim, S., Shim, H., Kim, J. and Yu, H. (2020). Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Proc. Interspeech*, 1496–1500, Shanghi, China.

Kwak, I., Kwag, S., Lee, J., Huh, J. H., Lee, C. , Jeon, Y., Hwang., J. and Yoon, J. W. (2021). Resmax: detecting voice spoofing attacks with residual network and max feature map. *25th International Conference on Pattern Recognition (ICPR)*, 4837–4844. Milan, Italy.

Lee, J., Lee, I. and Kang, J. (2019). Self-attention graph pooling. *International Conference on Machine Learning*, **68**, 3734–3743. PMLR.

Nautsch, A., Wang, X., Evans, N., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M. Yamagishi, J. and Lee, K. A. (2021). Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, **3**, 252–265.

Parhami, B. (1994). Voting algorithms. *IEEE Transactions on Reliability*, **43**, 617-629.

Syarif, I., Zaluska, E., Prugel-Bennett, A. and Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, **7376**, 593-602.

Tak, H., Jung, J., Patino, J., Kamble, M., Todisco, M. and Evans, N. (2021). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv Preprint arXiv*:2107.12710.

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N. and Larcher, A. (2021). End-to-end anti-spoofing with rawnet2. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6369–6373, Toronto, Ontario, Canada.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y. (2017). Graph attention networks. *arXiv Preprint arXiv*:1710.10903.

Wu, X., He, R., Sun, Z. and Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, **13**, 2884–2896.

Journal of the Korean Data & Information Science Society
2023, **34**(2), 203–214

http://dx.doi.org/10.7465/jkdi.2023.34.2.203
한국데이터정보과학회지

A two-stage training approach for voice spoofing detection[†]

Taein Kang[1] · Il-Youp Kwak[2]

[12]Department of Applied Statistics, Chung-Ang University

Received 12 January 2023, revised 13 February 2023, accepted 7 March 2023

Abstract

A novel 2-stage training method for voice spoofing detection is presented in this work along with performance experiments. The challenge of voice spoofing detection is to tell a real voice from a spoof that has been replicated in a setting other than the original voice. In areas where speaker identification is crucial to security, such as voice assistants, the demand for speech forgery detection is on the rise. The proposed 2-stage training model imports the embedding vectors of several single speech models studied with the Automatic Speaker Verification Spoofing (ASVSpoof) 2019 competition LA data set, combines them to define a concatenated embedding feature, and then builds a deep learning network on the concatenated embedding feature to create an ensemble model. We examined the analysis results based on the fusion of embedding vectors from various single models and modifications to deep learning networks for comparison with existing ensemble methodologies. The 2-stage training model produced an EER of 0.26 (%) by combining a number of models. This is a 0.34 (%p) improvement over the ensemble technique (Voting method) of 0.60 (%) and a 0.57 (%p) improvement over the single model's highest performance of 0.83 (%).

Keywords: Deep learning, two-stage training, voice spoofing detection, embedding.

[†] This research was supported by the Chung-Ang University Research Scholarship Grants in 2022.
[1] Graduate student, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea.
[2] Corresponding author: Assistant professor, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: ikwak2@cau.ac.kr