The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 435–446

DOI: https://doi.org/10.5351/KJAS.2021.34.3.435
Print ISSN1225-066X / Online ISSN 2383-5818

Data augmentation in voice spoofing problem

Hyo-Jung Choi[a], Il-Youp Kwak[1,a]

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

Abstract

ASVspoof 2017 deals with detection of replay attacks and aims to classify real human voices and fake voices. The spoofed voice refers to the voice that reproduces the original voice by different types of microphones and speakers. data augmentation research on image data has been actively conducted, and several studies have been conducted to attempt data augmentation on voice. However, there are not many attempts to augment data for voice replay attacks, so this paper explores how audio modification through data augmentation techniques affects the detection of replay attacks. A total of 7 data augmentation techniques were applied, and among them, dynamic value change (DVC) and pitch techniques helped improve performance. DVC and pitch showed an improvement of about 8% of the base model EER, and DVC in particular showed noticeable improvement in accuracy in some environments among 57 replay configurations. The greatest increase was achieved in RC53, and DVC led to an approximately 45% improvement in base model accuracy. The high-end recording and playback devices that were previously difficult to detect were well identified. Based on this study, we found that the DVC and pitch data augmentation techniques are helpful in improving performance in the voice spoofing detection problem.

Keywords: voice spoofing detection, data augmentation, deep learning, audio data

1. 서론

2017년 '국제 전자제품 박람회(Consumer Electronics Show; CES)' 이후 음성비서 시장이 빠르게 성장해왔다. 많은 기업이 음성비서 산업에 뛰어들었고 오늘날 음성비서는 핸드폰, 자동차, 냉장고, 스마트 TV 등에 내장되어 우리에게 익숙한 기술로 자리 잡아가고 있다. 이는 음성비서가 우리의 일상 속 깊숙이 들어왔다는 의미이기도 하다. 하지만 그만큼 음성비서의 음성인식 기능을 노리는 해킹 공격 위협이 부상하고 있다. 해킹 공격은 사회적 혼란과 사생활 침해까지 야기할 수 있어 사전에 방지할 수 있는 방안이 필요하다.

Automatic speaker verification (ASV)는 지난 수십 년 동안 상당히 성숙해왔다. ASV 기술은 개인 인증을 위한 효율적이고, 편리하고, 신뢰할 수 있는 솔루션으로서 점점 더 많은 상업용 제품과 서비스를 제공하고 있다 (Delgado 등, 2018). 하지만 위조(spoofing) 음성에 취약하다는 우려도 있다. 위조 음성이란 조작된 신호를 사용하여 ASV 시스템의 정상적인 작동을 방해하려는 공격 시도를 뜻하고 이는 소리를 흉내 내거나, 녹음 후 재생시키거나, 음성 합성 및 변환을 하는 모든 취약점을 포함한다. 음성인식 기기는 사용자의 음성에만 작동해야 하므로 해결되지 않을 경우 신뢰도를 떨어뜨릴 수 있어 연구계는 ASV 기술이 위조 음성에 취약하다는

This research was supported by the National Research Foundation of Korea (NRF) grant funded by Ministry of Science and ICT (2020R1C1C1A01013020).

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjakgu, Seoul 06974, Korea. E-mail: ikwak2@cau.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr

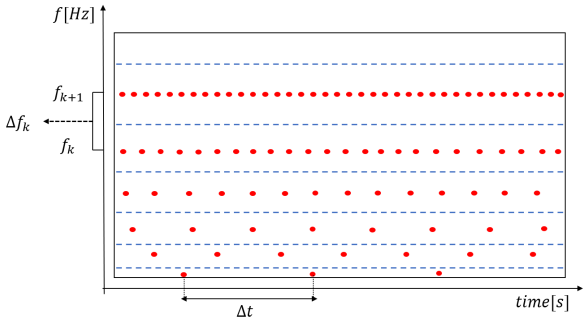


Figure 1: *In CQT, the difference between frequencies ($\triangle f_k$) and the difference between times ($\triangle f_k$) varies to make the Q value constant.*

우려에 대해 대응을 해왔다. 또한 음성인식이 거래와 관련된 산업으로까지 확장되고 있는 만큼 화자 인식의 중요성이 대두되고 있다. ASVspoof challenge는 위조 음성 공격에 대비한 ASV 시스템의 보안 강화를 목표로 한다. ASVspoof challenge는 2015년부터 시작된 챌린지로 2015년에는 음성 합성 및 변환과 같은 인공음성 탐지에 초점을 맞췄었다면 2017년은 재생 환경이나 재생 장치의 조건을 다르게 한 리플레이(replay) 공격에 초점을 맞췄다. 그런 의미에서 ASVspoof 2017은 음성을 녹음한 후 다시 재생시키는 리플레이 공격 검출의 실행적 한계를 평가하고 다양한 음향 조건에서 리플레이 공격을 탐지할 수 있는 대응책 개발을 촉진하는 기여를 하고자 한다 (Kinnunen 등, 2017).

데이터 증강 기법은 데이터가 부족한 상황에서 데이터를 증강시켜 더 강건한 모형을 학습하기위해 많이 사용되어 왔다. 특히, 이미지 데이터에 대한 데이터 증강기법 연구가 많이 되어 왔고 (Perez와 Wang, 2017; Mikolajczyk와 Grochowsk, 2018), 그림이 몇 장 되지 않는 퓨샷러닝(few shot learning) 또는 그림이 한 장만 있는 원샷러닝(one shot learning)의 문제에서 데이터 증강기법을 통해 강건한 모형을 학습시킬 수 있다는 것이 연구되었다 (Alfassy 등, 2019; Chen 등, 2019).

음성 데이터에도 증강기법을 활용한 여러 연구가 이루어져 왔다. 관련 연구인 Salamon과 Bello (2017)는 데이터 증강기법마다 어느 정도 변화시킬지 결정짓는 요인(factor)을 각 4가지씩 적용하여 늘어난 훈련데이터로 소리 환경을 구분하였고 Ko 등 (2017)은 원본 데이터와 요인 2가지의 증강기법을 적용한 데이터를 합쳐 훈련데이터를 3배로 늘린 후 성능향상을 이끌어냈다. 하지만 리플레이 공격 탐지 모형에는 어떤 증강기법이 좋은 영향을 미치는지에 대한 연구가 없는 실정이다.

본 논문에서는 리플레이 공격 문제에서 데이터 증강기법이 성능에 어떤 영향을 미치는지 알아보았다. 2장에서는 연구에 사용한 오디오 데이터의 전처리 작업과, 많이 쓰이는 light convolutional neural network (LCNN) 모형, 그리고 7가지 유형의 오디오 변형 기법에 대해 기술하였다. 3장에서는 우리가 제안한 증강기법별 LCNN 모형의 정확도 변화에 대해 살펴보았다.

2. 모형

2.1. 오디오 데이터 전처리

오디오 데이터를 모형에 넣기에 앞서 컴퓨터가 처리할 수 있는 데이터의 형태로 바꿔주는 과정이 필요하다. 오디오의 대표적인 특징값 추출 방법으로는 mel-frequency cepstral coefficients (MFCC), short-time fourier transform (STFT), constant-Q-transform (CQT) 등이 있으며, 딥러닝 모형에서는 STFT, CQT가 사용될 때 성

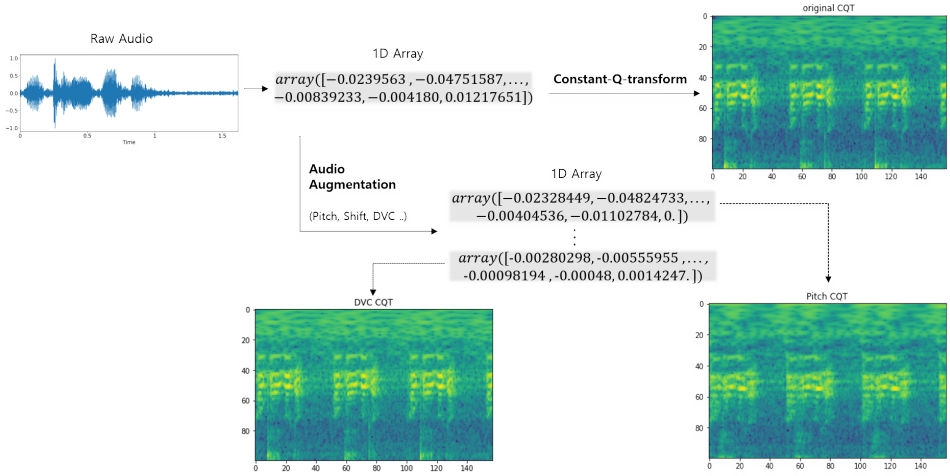


Figure 2: *Read the raw audio files (.wav, .flac) as 1D array, and apply constant Q transform. Top right figure is the original CQT spectrogram without data augmentation. Bottom left and right figures are augmented images using DVC and Pitch shifting.*

능이 더 좋은 편이다. 본 논문에서는 리플레이 공격 탐지에 유용하다고 알려진 CQT를 사용하였다 (Todisco 등, 2017). K개의 주파수 대역에서 데이터를 추출한다고 할 때, STFT는 각 주파수 들간 거리($\triangle f_k$, band width)가 고정된 상수지만, CQT는 Figure 1과 같이 주파수 대역별로 주파수 들간 거리를 다르게 하여 데이터를 추출한다. 이는 아래 수식에 정의된 Q를 고정된 상수(constant)로 만들어 주기 위함이다.

$$Q = \frac{f_k}{\triangle f_k}.$$

그러므로 STFT는 주파수(f_k)가 높아짐에 따라 Q가 증가하게 되지만 CQT는 $\triangle f_k$를 주파수 대역별로 조정하여 Q를 고정상수(Constant)로 만든다 (Todisco 등, 2017). Figure 1에서 주파수가 작아질수록 $\triangle f_k$가 함께 작아지는 것을 볼 수 있다. 이러한 CQT는 음악 처리 분야에서 처음 제안되었고, 많이 사용되고 있다 (Brown, 1991).

1차원의 연속적인 음성신호를 숫자로 표현하기 위해서 파동의 높이를 저장하게 되고 이를 샘플링(sampling)이라고 한다. 초당 샘플링 횟수를 의미하는 샘플링 레이트(sampling rate)는 16khz로 진행하였다. CD의 경우 보통 44.1khz(초당 44,100번)로 샘플링 되고 라디오는 22.05khz(초당 22,050번)의 샘플링을 한다. 샘플링 레이트가 높을수록 아날로그와 유사한 값을 갖기 때문에 더 좋은 음질을 얻을 수 있다. 모든 음성의 길이를 5초로 맞춰주었으며 5초보다 짧은 음성은 뒤로 이어 붙이고 5초보다 긴 음성은 잘라주는 전처리 과정을 거쳐 일정한 길이로 맞춰주었다. Figure 2는 음성 데이터에 CQT를 적용해 이미지 데이터로 바꿔주는 전처리 과정을 도식화 하였다.

2.2. Light convolutional neural network

Convolutional neural network (CNN)은 지난 10년간 컴퓨터 비전 문제를 해결하는 데 있어 가장 많이 사용된 기법이며 주로 이미지, 오디오, 비디오를 분류할 때 쓰인다. CNN 알고리즘은 각 이미지에서 특징을 활성화하는 합성곱 필터 집합에 이미지를 입력시키고 활성화 함수를 통해 특징을 다음 계층으로 전달시킨다. 그 후 풀링층에서 다운샘플링을 해 출력을 간소화 시키는 과정을 반복하여 각 계층이 여러 특징을 검출하는 방법을 학습시키는 과정이다. 신경망에서의 활성화 함수는 입력받은 신호를 다음 층으로 얼마나 출력할지를 결정하고

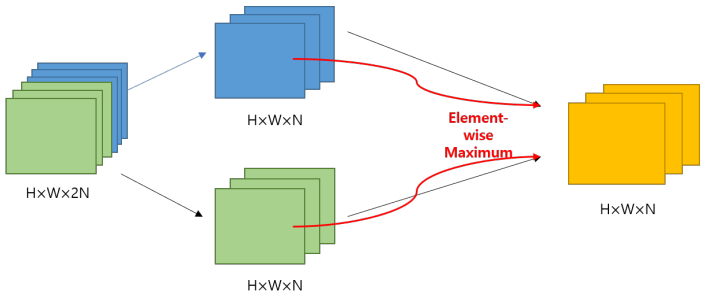


Figure 3: *Max-feature-map.*

비선형성을 표현 할 수 있도록 해준다. 활성화 함수가 선형이라면 신경망 층을 깊게 쌓는 것이 의미가 없어지기 때문에 중요한 역할을 한다. 대표적으로 rectified linear uni (ReLU), Sigmoid, Tanh (Hyperbolic tangent) 함수 등이 있다.

본 논문에서는 Lavrentyeva 등 (2017)이 제안한 light convolutional neural network (LCNN) 모델로 분석을 진행하였다. LCNN 모형의 특징으로는 max-feature-map (MFM) 활성화 함수를 사용하여 CNN 모형을 단순화 시키는 것이다. Goodfellow 등 (2013)이 제안한 maxout 활성화 함수를 응용한 MFM은 $H \times W \times 2N$ 차원의 데이터를 두개의 $H \times W \times N$ 차원 데이터로 나누어 채널(channel)방향으로 최댓값을 취하며, Figure 3에 도식화 하였다. MFM은 네트워크를 가볍게 하고 네트워크 수렴 속도를 높일 수 있을 뿐만 아니라 성능 역시 향상시킨다 (Wu 등, 2018). MFM은 잡음과 유익한 신호를 분리 할 수 있고 두 레이어의 동일한 위치에서 최댓값을 각각 선택한다. 이는 모델을 가볍고 강건하게 만들어준다.

사용한 LCNN 모형의 아키텍처는 Table 1과 같다. 훈련 데이터에 과대적합하는 것을 방지하기 위해 dropout 0.7과 dropout 0.5를 사용해주었다. Srivastava 등 (2014)이 제안한 Dropout은 신경망에서 뉴런들을 랜덤하게 제거하여 과대 적합을 방지해 주는 역할을 한다. Dropout이 0.5이면 절반의 뉴런을 랜덤하게 제거해 준다는 의미이며 1에 가까울 수록 뉴런을 많이 제거하지 않는 것이다. 맥스 풀링층은 2×2 커널과 스트라이드 2를 사용하여 차원을 줄여나갔고 FC7 층에서는 진짜 음성과 재생 음성을 구별하기 위해 Sigmoid 활성화 함수를 사용하였다. 옵티마이저(Optimizer)로는 ADAM($learning\_rate = 10^{-5}$), 평가지표(Metric)로는 area under curve (AUC)를 사용해 검증 데이터의 AUC가 가장 클 때의 가중치들을 저장하였다. 본 실험은 Tensorflow 2.2.0과 Keras 2.3.0 환경에서 실행되었다.

2.3. 데이터 증강

데이터 증강은 실제로 새로운 데이터를 수집하지 않고도 훈련 모델에 사용할 수 있는 데이터의 다양성을 크게 높일 수 있는 방법이다. 고양이 이미지(image)를 회전시키거나 미러링, 스케일링시켜도 여전히 고양이 이미지인 것 처럼 라벨(label)은 유지하면서 추가적인 훈련 데이터를 생성한다. 증강된 데이터를 훈련시키면서 네트워크는 데이터의 변형에 불변하게 되고 이전에 보지 못했던 데이터들에 대해서도 일반화가 더 잘 될 것을 기대할 수 있다 (McFee 등, 2015). 음성 데이터에서는 pitch, shift, speed, noise 등과 같은 데이터 증강 기술이 과대 적합을 방지하는 데 도움을 줄 수 있다.

2.3.1. Pitch

악기에 사용되는 피치 스케일링을 구현한 것이다. 음의 속도에 영향을 주지 않고 음의 음조를 바꾸는 과정이다. librosa의 pitch_shift를 사용하였으며 pitch_shift 함수는 오디오 샘플, 샘플링 레이트 및 피치 스텝 수를

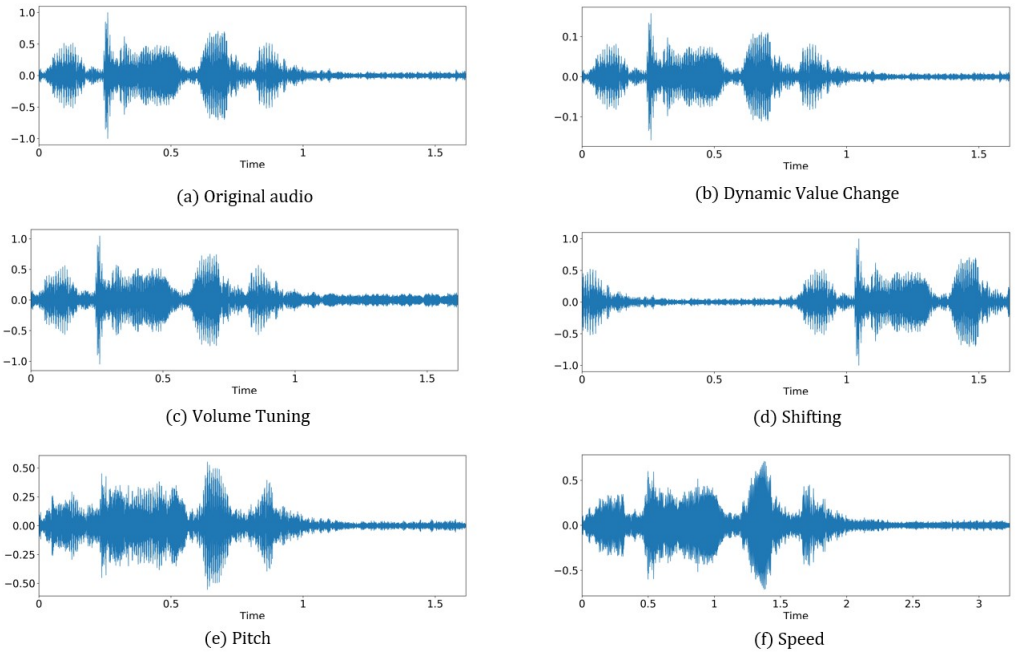
Table 1: Light convolutional neural network model

Type	Filter/Stride	Output	#Params
Conv1	5 × 5 / 1 × 1	100 × 157 × 32	832
MFM1	-	100 × 157 × 16	-
MaxPool1	2 × 2 / 2 × 2	50 × 79 × 16	-
Conv2a	1 × 1 / 1 × 1	50 × 79 × 32	544
MFM2a	-	50 × 79 × 16	-
Conv2b	3 × 3 / 1 × 1	50 × 79 × 48	6960
MFM2b	-	50 × 79 × 24	-
MaxPool2	2 × 2 / 2 × 2	25 × 40 × 24	-
Conv3a	1 × 1 / 1 × 1	25 × 40 × 48	1200
MFM3a	-	25 × 40 × 24	-
Conv3b	3 × 3 / 1 × 1	25 × 40 × 64	13888
MFM3b	-	25 × 40 × 32	-
MaxPool3	2 × 2 / 2 × 2	13 × 20 × 32	-
Dropout	-	13 × 20 × 32	-
Conv4a	1 × 1 / 1 × 1	13 × 20 × 64	2112
MFM4a	-	13 × 20 × 32	-
Conv4b	3 × 3 / 1 × 1	13 × 20 × 32	9248
MFM4b	-	13 × 20 × 16	-
Maxpool4	2 × 2 / 2 × 2	7 × 10 × 16	-
Dropout	-	7 × 10 × 16	-
Conv5a	1 × 1 / 1 × 1	7 × 10 × 32	544
MFM5a	-	7 × 10 × 16	-
Conv5b	3 × 3 / 1 × 1	7 × 10 × 32	4640
MFM5b	-	7 × 10 × 16	-
Maxpool5	2 × 2 / 2 × 2	3 × 5 × 16	-
Dropout	-	3 × 5 × 16	-
GlobalAveragePool	-	16	-
FC6	-	64	1088
MFM	-	32	-
Dropout	-	32	-
FC7	-	1	33
Total			41089

필요로 한다. 피치 스텝 수를 [–5, 5] 범위 내에서 실험해보았으며 본 데이터 집합에서는 –1에서 0 사이의 스텝 수가 가장 좋은 성능을 보였다. Figure 4에서 원본 오디오 (a)의 진폭 값 범위는 [–1, 1]이지만 pitch_factor 1을 적용한 (e)의 진폭 값은 [–0.5, 0.5]로 바뀌게 된다.

2.3.2. Shifting

시간 축을 따라 주어진 수 만큼 오른쪽으로 파형을 이동시키는 기법으로 numpy의 roll 함수를 사용하였다. shift_factor를 [0, 2] 범위 내에서 실험해보았으며 파형을 많이 이동시킬수록 성능이 좋지 않았다. shifting 시킨 Figure 4의 (d)와 원본 오디오 (a) 파형과 비교해봤을 때 오른쪽으로 밀린 음성이 앞으로 붙은 것을 확인할 수 있다.



(a) Original audio (b) Dynamic Value Change

(c) Volume Tuning (d) Shifting

(e) Pitch (f) Speed

Figure 4: *Audio augmentation.*

2.3.3. Speed

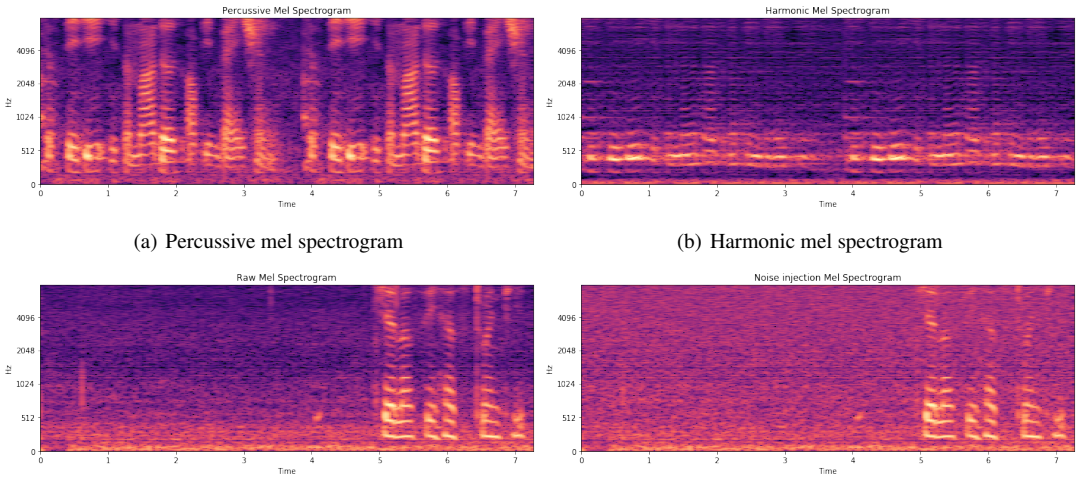
파형을 늘여 음성의 속도를 느리게 만들거나 파형을 압축 시켜 음성의 속도를 빠르게 만드는 기법이다. 파형을 늘이면 진동수가 작아져 낮은 소리가 나고 파형을 압축시키면 진동수가 많아져 높은 소리가 난다. librosa의 time_stretch를 사용하였으며 speed_factor를 [0, 2] 범위 내에서 실험해보았다. [0, 0.5] 사이의 성능이 가장 좋았으며 느린 음성이 빠른 음성보다 좋은 성능을 보였다. speed_factor를 0.5로 지정해준 파형 예시는 Figure 4의 (f)와 같다.

2.3.4. Volume tuning

이 데이터 증강 기법은 음성의 볼륨을 조절해주는 방법이다. 원 음성보다 볼륨을 낮춰서 실험을 진행하였으며 오디오 파일을 실수 형태로 읽어온 후 numpy의 random.uniform을 사용해 [0,0.5] 사이 랜덤한 수로 빼주었다. 파형 예시는 Figure 4의 (c)와 같다.

2.3.5. Dynamic value change

Dynamic value change (DVC)는 base의 파형과 똑같은 형태를 보이지만 진폭 값의 범위가 바뀌는 기법이다. 오디오 파일을 읽어올 때면 진폭 값이 [–1, 1] 범위의 실수로 변환되어 Figure 4의 (a)와 같은 진폭 값을 갖게 되는 게 일반적이다. 하지만 DVC를 적용하게 되면 지정해 준 범위 내 랜덤한 수로 그 실수를 나눠주게 되고 진폭 값의 범위는 [–1, 1]보다 더 좁아지게 된다. 원본 오디오 (a) 진폭 값의 범위가 [–1, 1]인 반면에 DVC (b) 진폭 값의 범위는 [-0.1, 0.1]가 되는 것을 확인 할 수 있으며 이는 실수 형태로 오디오를 읽어온 뒤 약 5.973으로 나눴을 때의 파형 예시이다. 본 연구에서는 [2, 10] 범위의 랜덤한 값들로 나눠주었을 때 가장 좋은 성능을



(a) Percussive mel spectrogram (b) Harmonic mel spectrogram

(c) Original mel spectrogram (d) Mel spectrogram with additional noise

Figure 5: *Mel-Spectrogram.*

보여줬으며 DVC를 적용한 후에도 진폭 값만 달라질 뿐이지 음성을 재생해보면 기법을 적용하기 전과 후의 소리가 같았다.

2.3.6. Harmonic percussive sound separation

Harmonic percussive sound separation (HPSS)는 하모닉(Harmonic, 음조 악기)과 퍼커시브(Percussive, 음조 악기가 아닌 것, 타악기)의 분리를 뜻하며 원본 음악 신호를 신호의 하모닉 부분과 퍼커시브 부분으로 분해하는 것이다. Figure 5와 같이 세로 패턴의 퍼커시브 (a), 가로 패턴의 하모닉 (b)로 분리된다. 이러한 방법은 오디오 믹싱 소프트웨어에 적용하거나 리듬 분석이나 화음/음향 인식과 같은 음악분야에 대한 전처리 방법으로도 사용할 수 있다. 본 실험에서는 하모닉 부분만을 사용해주었다.

2.3.7. Noise

원 음성에 소음을 추가해주는 증강 기법이다. 평균이 0이고 표준 편차가 1인 표준정규분포를 따르는 백색소음을 추가해준다. 이를 위해 numpy의 random.randn을 사용하여 오디오 샘플 데이터에 추가해주었으며 잡음에 강인한 모델을 만들 수 있게 해준다. 백색소음을 추가한 뒤 오디오를 재생해보면 음질이 좋지 않은 것을 확인할 수 있었으며 Figure 5의 멜 스펙트로그램(mel-spectrogram) (d)처럼 원본 오디오 (c)에 비해 경계가 흐려진 것을 확인 할 수 있다.

3. 분석결과

3.1. 데이터 설정

ASVspoof 2017 데이터는 전 세계의 지원자(대부분 ASV 연구원)들이 안드로이드 스마트폰을 이용해 수집한 *RedDots corpus*에서 유래한다. *RedDots Datasets*를 녹음한 후 다시 재생시킨 음성들은 공격자가 다양한

Table 2: The description of the ASVspoof 2017 dataset

Subset	# Speakers	# Non-replay	# Replay
Train	10	1508	1508
Dev	8	760	950
Eval	24	298	12008
Total	42	3566	14466

Figure 6: *The graphical description of how spoofing data is made. We considered multiple recording conditions (RC) that consist of recording devices, playback devices, and recording background conditions. The large number of RC could cover diverse range of spoofing attack scenarios.*

품질의 변환기를 통해서 화자의 원본 음성을 디지털 복제 할 수 있다는 위조(spoofing) 시나리오의 예시가 된다. ASVspoof 2017의 위조 음성이 만들어지는 과정은 Figure 6과 같다. 위조 음성 파일의 57% 는 EU Horizon 2020가 지원한 OCTAVE project 참여자 4명으로부터 수집되었으며, 나머지 43%는 다른 기여자들로부터 수집되었다. 평가 데이터 중 진짜 사람의 음성은(genuine/non-replayed) 7명의 새로운 화자들에게 수집된 음성들로 보완되었고 ASVspoof 2017가 제공하는 데이터는 훈련, 검증, 평가로 나뉘어있다. 전체 데이터는 총 14,466개의 음성 파일로 구성되어 있으며 세부내용은 Table 2와 같다.

훈련, 검증, 평가 데이터의 화자는 겹치지 않는다. 데이터 수집 사이트(web-site)도 분리되어 있는데 훈련데이터는 단일 사이트, 검증데이터는 훈련데이터 사이트와 2개의 추가적인 사이트에서 수집되었다. 또한 평가 데이터는 검증 데이터를 수집한 3개의 사이트 이외에 새로운 2개의 사이트에서 추가로 데이터를 수집하였다. 만약 서로 같은 사이트의 데이터라 하더라도 녹음·재생 장치, 환경을 모두 다르게 수집하였다 (Kinnunen 등, 2017). 이런 데이터의 이질성은 신뢰할 수 있는 spoofing 대응책 개발에 필수적인 것으로 입증되었다 (Lavrentyeva 등, 2017; Korshunov과 Marcel, 2016).

3.2. 데이터 증강기법 별 성능

Equal error rate (EER)은 생체 인식 성능을 표시하는데 사용되는 지표로 False acceptance rate (FAR)과 False rejection rate (FRR)에 대한 임계값을 예측하기 위해 사용된다. EER은 FAR과 FRR이 같아지는 지점과 동일하다. EER이 낮은 기기 일수록 더 정확한 것으로 간주된다 (Kinnunen 등, 2017).

$$\mathrm{FAR} = P_{\mathrm{fa}} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} = \frac{\#(\text{replay trials with score} > \theta)}{\#(\text{total replay trials})},$$

$$\mathrm{FRR} = P_{\mathrm{miss}} = \frac{\mathrm{FN}}{\mathrm{TN}+\mathrm{FN}} = \frac{\#(\text{non-replay trials with score } \leq \theta)}{\#(\text{total non-replay trials})}.$$

어떠한 증강기법도 적용하지 않은 기본 모델의 성능에 비해 DVC, Pitch 증강기법을 적용했을 때 성능이

Table 3: The performance by data augmentation techniques

Augmentation	Eval EER Mean	Eval EER SD	*p*-value
Base	23.80	0.0067	
DVC	21.92	0.0078	1.461e-05 **
Pitch	21.97	0.0045	4.516e-06 **
HPSS	25.71	0.0034	
Volume -	27.88	0.0030	
Shift 0.5	26.93	0.0057	
Shift 1.5	30.57	0.0054	
Speed 0.5	44.79	0.0025	
Speed 0.7	48.07	0.0053	
Speed 2.0	50.29	0.0080	
Noise	46.69	0.0067	

향상되었다. 증강기법들로 이루어낸 성능 변화가 통계적으로 유의한지를 보이기 위해 오디오 데이터를 각 10회씩 복원 추출하였고 기법마다 측정된 EER에 대응표본 t-test를 진행하였다. 그 결과는 Table 3과 같으며 집단의 전 후 평균차이가 유의하였다. p-value < 0.05, with Bonferroni correction). DVC와 Pitch 증강기법을 적용했을 때의 EER이 기본 모델보다 감소하였음을 통계적으로 확인 할 수 있었고 DVC와 Pitch를 제외한 나머지 기법들은 성능 향상에 도움이 되진 않았다.

3.3. 환경 별 증강기법의 성능

ASVspoof 2017의 데이터는 환경변수(replay configuration) 별로 녹음 상태가 다르다. 잡음이 많은 환경에서 녹음되어 비교적 위조임을 쉽게 탐지할 수 있는 환경과 잡음이 거의 없이 녹음되어 위조인지 탐지하기 어려운 환경, 중간 정도의 잡음에서 녹음된 환경으로 이루어져 있다. 또한 음성의 녹음과 재생 장치를 성능별로 다양하게 사용하였으며 고사양, 중사양, 저사양의 장치로 나뉘어 있다.

Delgado 등 (2018)에 의해 각 음향 환경과 장치들의 조합이 57개의 환경변수로 재정의되었다. 본 연구에서는 성능향상을 보였던 2가지의 증강기법이 환경변수 별로는 어떠한 영향을 미칠지 알아보기 위해 환경변수별 정확도를 측정하였다. Keras의 model.predict_generator을 사용해 예측된 스코어가 EER_threshold(θ_{EER}) 보다 작으면 Spoof(= 0), θ_{EER} 보다 크면 Genuine(= 1)으로 판단하였다. 환경변수별 정확도를 구하는 수식은 다음과 같다.

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}}, \\ &= \frac{\#(\text{predict spoof}|\text{true spoof}) + \#(\text{predict genuine}|\text{true genuine})}{\#(\text{total trials})}.\end{aligned}$$

성능 향상을 이끌어냈던 2가지의 증강기법을 각 환경변수마다 적용한 결과 기본 모델에 비해 향상된 정확도를 Figure 7에서 시각적으로 확인할 수 있다. DVC의 경우 큰 폭으로 성능이 향상된 환경변수들이 있어 주목할 만 하다. 기본 모델의 정확도가 거의 1에 근사하는 경우 (탐지하기 쉬운 경우) 증강기법의 효과를 확인하기 어려웠지만, 기본 모델의 정확도가 낮은 (탐지하기 어려운) 배치에서 증강 기법의 성능 향상을 확인 할 수 있었다. 가장 큰 증가 폭을 보인 RC40-RC53은 고사양의 녹음·재생 장치로 만들어진 음성들로 진짜 사람 음성과의 구분이 어려웠는데 데이터 증강기법을 통해 성능 향상을 이끌어 낸 것을 확인 할 수 있다. 특히, 가장 큰 폭으로 증가한 RC53의 경우 기본 모델의 약 45%의 성능 향상을 이끌어냈다.

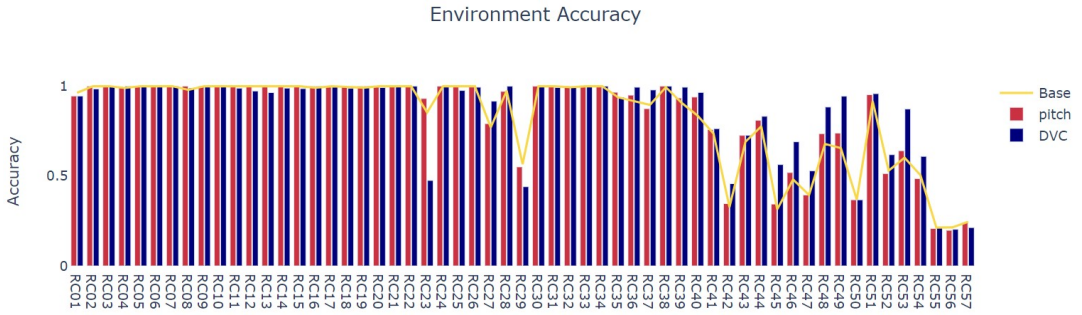


Figure 7: *The barplot of accuracy for each recording conditions (RC). Each RC consists of one replay device, one recording device and one recording background.*

4. 결론 및 시사점

본 논문에서는 ASVspoof 2017 데이터에 7가지의 많이 알려진 증강 기법을 적용해 성능 향상 여부에 대해 실험하였다. 그 결과 Pitch, DVC를 적용시켰을 때 유의미한 성능 향상이 있음을 확인하였다. 하지만, Pitch, DVC를 제외한 5가지 증강기법은 성능 향상에 도움이 되지 않았다. 기존의 음성위조탐지 대회에서 제출된 모형들에 데이터 증강기법이 많이 사용되고 있지 않았는데, 이러한 연유로 많은 팀이 데이터 증강기법을 사용하지 않은 것으로 생각된다. 본 연구는 5초간의 음성에 단일 증강 기법으로 실험하였으며, 더 긴 시간의 음성에 적용하거나, 여러 증강 기법 적용한 데이터들을 합쳐 훈련 데이터의 양을 늘리게 된다면 어떻게 될지에 대한 추가적인 연구들이 필요할 것으로 생각된다. 특히 DVC의 경우 훈련데이터를 더 많이 만들어 낼수록 좋은 결과를 보여주었다. 또한 Figure 7에서도 알 수 있듯이 본 논문에서 적용한 증강기법은 기존에 음성 위조를 탐지하기 어려운 장치나 환경들에서 큰 폭의 증가를 보였다. 이는 더 많은 훈련데이터를 수집하지 않고도 데이터 증강기법을 통해 성능을 향상 시킬 수 있다는 것을 의미하며, 기존에 연구가 이루어지지 않았던 음성위조탐지 문제에 DVC와 Pitch 증강기법 적용과 응용에 관한 추가적인 연구가 필요할 것으로 보인다.

References

Alfassy A, Karlinsky L, Aides A, Shtok J, Harary S, Feris, R, Giryes R, and Bronstein AM (2019). Laso: label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6548–6557.

Brown JC (1991). Calculation of a constant Q spectral transform, *The Journal of the Acoustical Society of America*, **89**, 425–434.

Chen Z, Fu Y, Wang YX, Ma L, Liu W, and Hebert M (2019). Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8680–8689.

Delgado H, Todisco M, Sahidulla Md, Evans N, Kinnunen T, Lee K, and Yamagishi J (2018). *Asvspoof 2017 Version 2.0: Meta-Data Analysis and Baseline En-hancements*.

Goodfellow I, Warde-Farley D, Mirza M, Courville A, and Bengio Y (2013). Maxout networks. In *International Conference on Machine Learning, PMLR*, 1319–1327.

Kinnunen T, Sahidulla Md, Delgado H, Todisco M, Evans N, Yamagishi J, and Lee K (2017). The asvspoof 2017 challenge: assessing the limits of replayspoofing attack detection, *Interspeech 2017*, 2–6.

Ko T, Peddinti V, Povey D, and Khudanpur S (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Korshunov P and Marcel S (2016). Cross-database evaluation of audio-based spoofing detection systems, *In Interspeech*.

Lavrentyeva G, Novoselov S, Malykh E, Kozlov A, Kudashev O, and Shchemelinin V (2017). Audio replay attack detection with deep learning frameworks, *In Interspeech*, 82–86.

McFee B, Humphrey EJ, and Bello JP (2015). A software framework for musical data augmentation, *ISMIR*, 248–254.

Mikołajczyk A and Grochowski M (2018). Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE*, 117–122.

Perez L and Wang J (2017). *The Effectiveness of Data Augmentation in Image Classification Using Deep Learning*, arXiv preprint arXiv:1712.04621.

Salamon J and Bello JP(2017). Deep convolutional neural networks and data augmentation for environmental sound classification, *IEEE Signal Processing Letters*, **24**, 279–283.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R (2014). Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, **15**, 1929–1958.

Todisco M, Delgado H, and Evans N (2017). Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification, *Computer Speech & Language*, **45**, 516–535.

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.

Received January 11, 2021; Revised January 28, 2021; Accepted February 6, 2021

데이터 증강기법을 이용한 음성 위조 공격 탐지모형의 성능 향상에 대한 연구

최효정[a], 곽일엽[1,a]

[a]중앙대학교 응용통계학과

요 약

본 논문에서는 음성위조공격탐지(Voice spoofing detection) 문제에 데이터 증강을 적용한다. ASVspoof 2017은 리플레이 공격 탐지에 대해 다루며 진짜 사람의 음성과 환경이나 녹음·재생 장치의 조건들을 다르게 하여 위조한 가짜 음성을 분류하는 것을 목적으로 한다. 지금까지 이미지 데이터에 대한 데이터 증강 연구가 활발히 이루어졌으며 음성에도 데이터 증강을 시도하는 여러 연구가 진행되어왔다. 하지만 음성 리플레이 공격에 대한 데이터 증강시도는 이루어지지 않아 본 논문에서는 데이터 증강기법을 통한 오디오 변형이 리플레이 공격 탐지에 어떠한 영향을 미치는지에 대해 탐구해본다. 총 7가지의 데이터 증강기법을 적용해보았으며 그 중 DVC, Pitch 음성 증강기법이 성능향상에 도움되었다. DVC와 Pitch는 기본 모델 EER의 약 8% 개선을 보여주었으며, 특히 DVC는 57개의 환경변수 중 일부 환경에서 눈에 띄는 정확도 향상이 있었다. 가장 큰 폭으로 증가한 RC53의 경우 DVC가 기본 모델 정확도의 약 45% 향상을 이끌어내며 기존에 탐지하기 어려웠던 고사양의 녹음·재생 장치를 잘 구분해냈다. 본 연구를 토대로 기존에 증강기법의 효과에 대한 연구가 이루어지지 않았던 음성 위조 탐지 문제에서 DVC, Pitch 데이터 증강기법이 성능 향상에 도움이 된다는 것을 알아내었다.

주요용어: 음성 위조 탐지모형, 데이터 증강기법, 딥러닝, 음성자료

이 성과는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1C1C1A01013020).

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: ikwak2@cau.ac.kr