한국음향학회지 제37권 제6호 pp. 475~482 (2018)
The Journal of the Acoustical Society of Korea Vol.37, No.6 (2018)
https://doi.org/10.7776/ASK.2018.37.6.475

pISSN : 1225-4428
eISSN : 2287-3775

Deep neural network 기반 오디오 표식을 위한 데이터 증강 방법 연구

Study on data augmentation methods for deep neural network-based audio tagging

김범준,[1] 문현기,[2] 박성욱,[3] 박영철[4†]

(Bum-Jun Kim,[1] Hyeongi Moon,[2] Sung-Wook Park,[3] and Young cheol Park[1†])

[1]연세대학교 전산학과, [2]연세대학교 전기전자공학과, [3]강릉원주대학교 전자공학과, [4]연세대학교 컴퓨터정보통신공학부

(Received September 14, 2018; accepted November 21, 2018)

초 록: 본 논문에서는 DNN(Deep Neural Network) 기반 오디오 표식을 위한 데이터 증강 방법을 연구한다. 본 시스템에서는 오디오 신호를 멜-스펙트로그램으로 변환하여 오디오 표식을 위한 심층신경망의 입력으로 사용한다. 적은 수의 훈련 데이터를 사용하는 경우 발생하는 문제를 해결하기 위해, 타임 스트레칭, 피치 변화, 동적 영역 압축, 블록 혼합 등의 방법을 사용하여 훈련 데이터를 증강시켰다. 사용된 데이터 증강 기법의 최적 파라미터와 최적 조합을 오디오 표식 시뮬레이션을 통해 확인하였다.
핵심용어: 오디오 표식, 인공신경망, 데이터 증강, 파라미터 조정

ABSTRACT: In this paper, we present a study on data augmentation methods for DNN (Deep Neural Network)-based audio tagging. In this system, an audio signal is converted into a mel-spectrogram and used as an input to the DNN for audio tagging. To cope with the problem associated with a small number of training data, we augment the training samples using time stretching, pitch shifting, dynamic range compression, and block mixing. In this paper, we derive optimal parameters and combinations for the augmentation methods through audio tagging simulations.
Keywords: Audio tagging, DNN (Deep Neural Network), Data augmentation, Parameter tuning

PACS numbers: 43.60.Bf, 43.60.Uv

I. 서 론

인터넷 기반의 멀티미디어 서비스가 보편화되고 가정용 인공지능 스피커가 확산되면서, 오디오와 관련된 정보를 이용하여 콘텐츠를 검색하는 기능과 인공지능 스피커가 오디오 신호를 기반으로 주변상황을 판단하고 적절히 반응 하는 서비스들이 점차 주목 받고 있다. 이와 같은 서비스를 제공하기 위해서는 오디오 신호를 분석하여 어떤 사건(혹은 내용)이 포함되어 있는지를 분류하는 기능이 필요한데 이를 오디오 표식 기능이라고 한다. 오디오 표식은 정보 검색,[1] 음향 분류,[2] 추천 시스템[3]과 같은 다양한 응용 분야에 적용된다. 이를 이용하여 개인용 미디어 관리 기술에 적용할 수 있다.

오디오 표식 알고리즘은 입력된 오디오 신호에 포함된 이벤트 종류를 출력한다. 오디오 표식에는 GMM(Gaussian Mixture Model) 알고리즘이 사용되었으나[4] 최근 연구들은 이미지 및 음성 신호 처리에서 높은 성능을 보여준 인공 신경망(Deep Neural Network, DNN) 구조를 오디오 표식에 적용하여 더욱 높은 성능을 달성하였다.

인공 신경망은 여러 비선형 변환기법의 조합을 통

†**Corresponding author:** Young cheol Park (young00@yonsei.ac.kr)
Division of Computer and Telecommunication Engineering, Chang jo room 269, Yonsei University, 1 Yonseidae-gil, Wonju, Gangwon-do 26493, Republic of Korea
(Tel: 82-33-760-2756, Fax: 82-33-763-4323)
"이 논문은 2018년도 한국음향학회 음성통신 및 신호처리 학술대회에서 발표하였던 논문임."

해 추상화를 시도하는 기계학습의 종류이다. 특징 추출을 자동으로 수행하고 훈련 데이터의 양이 많아질수록 성능이 좋아진다. 또한 예측력이 다른 머신러닝 기법들에 비해 상대적으로 우수하다.

오디오 표식의 성능을 높이기 위하여 여러 방법들이 사용되었으나,[4] 오디오 표식에서 가장 높은 성능을 보여준 인공신경망 구조는 CRNN(Convolutional Recurrent Neural Networks) 이다. CRNN은 높은 오디오 표식 성능을 위하여 깊은 합성곱 신경망(Convolutional Neural Network, CNN)을 갖고 있을 뿐만 아니라 순환 신경망(Recurrent Neural Network, RNN), 완전 접속망(Fully-connected Neural Network, FNN)이 결합된 복잡한 구조를 가진다. 큰 규모의 네트워크를 훈련하기 위하여 훈련 데이터의 크기가 커져야 하는 것이 일반적인 DNN의 특성이다. 하지만 훈련에 사용가능한 양질의 데이터를 다수 확보하기는 쉽지 않으며 이러한 경우에는 데이터 증강방법이 필수적이다.

본 논문에서는 DNN의 성능을 올리기 위하여 훈련 데이터를 증가하는 방법을 사용하였다. 데이터 증강 방법으로는 원 데이터를 신호처리 기법으로 변형하여 추가 데이터를 생성하는 방식[5]을 사용하였다. 최근 생성 모델(Generative DNN Model)을 이용하여 원 데이터와 비슷한 데이터를 생성하는 방식[6]들이 제안되었지만, 안정적으로 다량의 데이터를 생성하기에는 무리가 있어 본 연구에서는 포함하지 않았다.

본 논문에서는 원 데이터를 신호처리 기법으로 변형하여 데이터를 증강하는 방법을 사용하였고, 이 기법에 적용된 파라미터들을 조정하면서 성능의 변화를 확인하였다. 심층 신경망 구조로는 state-of-the-art 구조인 CRNN을 사용하였으며 데이터 증감 방법으로는 타임 스트레칭(Time Stretching, TS), 피치 변화(Pitch Shifting, PS), 동적 영역 압축(Dynamic Range Compression, DRC), 블록 혼합(Block Mixing, BM)을 사용하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 사용된 DNN 구조를 소개한다. 3장에서 오디오 표식의 성능을 향상시키기 위해 제안된 데이터 증강 방법을 소개한다. 4장에서 알고리즘 구현을 소개하여 동작하는 구조 및 평가 방법을 설명한다. 5장에서 제안된 알고리즘에 따른 결과를 소개한다. 마지막으로 6장에서 결론으로 마무리한다.

II. 심층신경망 구조

본 논문에서 Fig. 1과 같은 CRNN 구조를 사용하였다.[7] 사용된 DNN 구조는 3 개의 CNN 계층, 1 개의 양방향 회귀 신경망(Bidirectional Neurual Network, Bi-RNN), 1 개의 FNN, 전역 평균 풀링(global average pooling)으로 이루어진다.

본 구조는 시간 영역의 입력 신호를 프레임 단위로 멜-스펙트로그램으로 변환하여 입력으로 사용한다. CNN은 주어진 입력에 대한 특징을 추출한다. 각 CNN 블록은 ReLU 활성 함수와 최대 풀링(max pooling)을 사용한다. Bi-RNN은 과거와 미래 상태를 고려한 확장된 형태를 가지고 있다.[8] 입력 데이터의 사전/후 연관성을 고려하여 입력 데이터에 어떤 표식을 부여하는 것이 적정한지를 추정한다. 그리고 FNN을 통해 Bi-RNN의 결과를 종합한다. 오디오 클립 전체에 대한 표식을 부여하기 위하여, 프레임 단위로 추정한 결과들을 모아서 종합적인 판단을 내릴 필요가 있다. 전역 평균 풀링이 이 역할을 담당한다.

훈련 시 문제점인 과적합, 기울기 소실 문제(vanishing

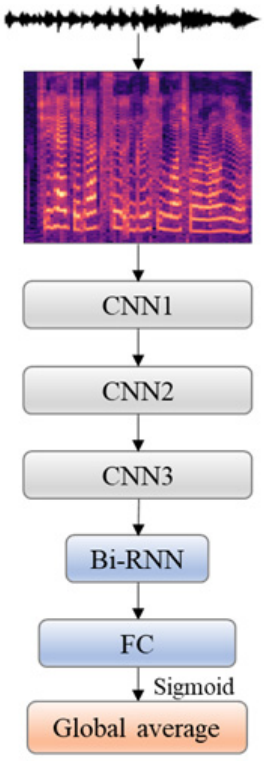

Fig. 1. Block diagram of the DNN structure.

gradient problem) 등을 해결하고자, 각 레이어 사이에 드롭아웃, 배치 정규화 방법을 적용하였다.[9]

III. 데이터 증강

주어진 표식이 붙은 데이터가 충분하지 못한 경우가 많기 때문에 적절한 훈련을 위하여 데이터 증강 방법을 적용한다. 오디오 표식을 위한 데이터 증강 방법으로 널리 사용되는 타임 스트레칭, 피치 변화, 동적 영역 압축, 블록 혼합 등 네 가지 방법을 적용한다.[5]

3.1 타임 스트레칭

타임 스트레칭은 오디오 신호의 피치에는 영향을 주지 않으면서 오디오의 속도 또는 지속 시간을 바꾸는 방법이다. 주어진 오디오 신호에 대하여 리샘플링하는 형태로 수행 한다.본 논문에서는 stretch factor로 [0.6 ~ 3.4] 범위에 대하여 타임 스트레칭을 적용하였다. 초기 오디오 신호를 1로 하여 0.2배씩 변경하며 14가지 경우에 대하여 수행 하였다.

3.2 피치 변화

피치 변화는 오디오 신호의 속도 또는 지속 시간에는 영향을 주지 않으면서 피치를 바꾸는 방법이다. 주파수 영역에서 한 옥타브를 12개의 반음 단위로 구분되어 변화가 수행된다. pitch shift factor로 [-4 ~ 4] 범위에 대하여 피치 변화를 수행하며 반음 단위로 16 가지 경우에 대하여 결과를 확인하였다.

3.3 동적 영역 압축

오디오 신호의 크기를 줄이거나 증폭 시키는 것이다. 이를 통해 오디오 신호의 동적 영역을 압축하게 된다. DRC 동작을 함에 있어 DRC 곡선을 사용한다. DRC 곡선은 신호의 입력 레벨에 대한 출력 레벨을 나타내며 그 차이가 실제 적용될 이득이 된다.

사용되는 이득곡선은 Fig. 2와 같은 형태로 주어진다. Boost Range는 입력 신호가 증폭되어 출력되는 영역을 나타낸다. Null Band는 입력 신호의 레벨 그대로 출력이 되는 구간을 나타낸다. Cut Range는 입력 신호보다 작은 레벨의 출력이 나가는 구간이다. 입력 신호의 레벨에 대해 증폭 또는 압축됨에 따라 동적 영역이 줄어들게 된다.

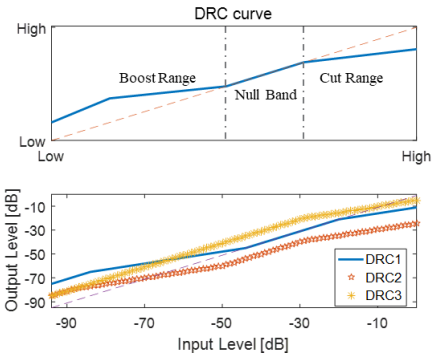

Fig. 2. Example of DRC curve.

본 논문에서는 3 가지 DRC curve[10,11]를 사용하여 결과를 비교한다. [Boost Range, Null Range, Cut Range] 순으로 입력되며, [-65, -41, -21], [-80, -60, -40], [-20, -10, 0]의 세 가지 곡선에 대해 동작한다.

3.4 블록 혼합

블록 혼합은 두 개의 다른 오디오 신호를 특정 조건에 대하여 혼합을 한다. 블록 혼합을 통해 훈련에 사용되는 DNN이 일반화가 되도록 한다. 본 논문에서 표식 관점에서 두 가지 조건에 대하여 혼합을 수행한다. 신호 관점에서의 혼합 방법도 달리하여 결과를 확인한다.

블록 혼합을 위해 라벤 관점의 두 가지 조건을 제시한다. 첫 번째 조건으로 단일 표식을 가지는 오디오 클립들에 대하여, 서로 다른 두 클립의 표식이 동일한 경우다. 두 번째 조건으로는 두 개의 표식을 가지는 오디오 클립들에 대하여, 한 개 또는 두 개의 표식이 동일한 경우다.

$$M(\alpha) = \alpha \times F + (1-\alpha) \times S. \tag{1}$$

$$M_{norm}(\alpha) = \frac{M(\alpha)}{\max(|M(\alpha)^{\forall}|)}. \tag{2}$$

신호 관점에서의 혼합 방법으로 크게 두 가지의 방법을 사용한다. F는 첫 번째로 선택된 오디오 클립

을, S는 두 번째로 선택된 오디오 클립을 나타낸다. M은 혼합된 오디오 클립을 의미한다. α는 두 번째 방법에서 사용될 비율을 나타낸다. 혼합 방법으로-진폭에 대한 정규화를 사용한 Eq. (2)를 사용한다. 블록 혼합의 효과를 보기 위하여 파라미터는 α = [0.5, 0.6, 0.7]을 사용한다.

IV. 알고리즘 구현

4.1 데이터 세트

사용된 데이터 세트는 구글에서 제공하는 오디오 세트 중 일부를 사용하였다.[12] 10 종류의 음향 표식을 가지며 하나 이상의 표식을 갖는 멀티 표식이 된 오디오 클립들로 이루어져있다. 1,578개의 표식이 붙은 훈련용 오디오 클립들, 288개의 테스트용 오디오 클립들로 이루어진다. 표식이 있는 오디오 클립들의 표식 분포는 Table 1과 같다. 불균형한 분포를 띄우고 있으며 특히 'Speech'에 대해 치중되어 있다.

데이터의 수가 부족하여 충분히 훈련되지 않는 문제를 해결하기 위하여 3 장에서 제시한 데이터 증강 방법을 사용하여 데이터를 증강하였다.

4.2 훈련 구조

Fig. 1의 DNN 구조를 사용한다. 전반적인 훈련 진행은 Fig. 3과 같이 수행한다. 주어진 음향 전체에 표식이 붙은 데이터 중 80 %를 훈련 데이터로 사용하며, 20 %는 검증 데이터로 사용한다. 데이터 증강 방법은 훈련 데이터에 대해서만 이루어진다.

Table 1. Distribution of weakly labeled data each class.

Class	# of clips
Speech	550
Dog	214
Cat	173
Alarm / bell / ringing	205
Dishes	184
Frying	171
Blender	134
Running water	343
Vacuum cleaner	167
Electric shaver / toothbrush	103
Total	2244

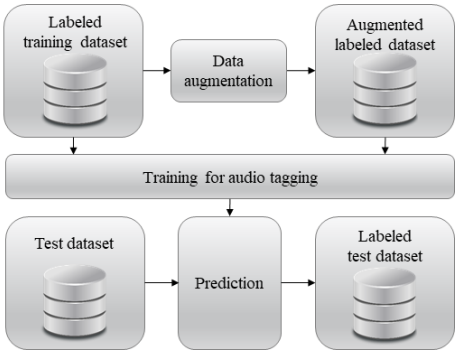

Fig. 3. Block diagram for overall structure.

각 데이터 증강 방법의 파라미터를 바꾸어 가며 결과를 확인한다. 각 방법 별 파라미터의 성능 차이, 방법에 따른 성능 차이를 확인하였다.

또한 데이터 증강 방식과 파라미터를 복합적으로 사용할 경우의 성능 차이도 확인하였다. 구체적으로는 하나의 증강 방법에서 좋은 성능을 내는 상위 2 개의 파라미터로 생성한 데이터로 훈련 데이터를 구성하였다. 또한 이렇게 상위 2개의 파라미터로 생성한 데이터를 다른 증강 방법으로 생성한 데이터와 합하여 훈련 데이터를 구성하기도 하였다. 이와 같은 방법으로 본 논문에서 사용한 데이터 증강 방법들의 모든 조합에 해당하는 훈련 데이터들을 생성하였다.

4.3 성능 평가 방법

성능을 평가할 지표로서 f-score가 있다. 이때 실제 이벤트가 있을 때 있다고 판별하는 TP(True Positive), 이벤트가 있지만 없다고 판별하는 FN(False Negative), 이벤트가 없을 때 없다고 판별하는 TN(True Negative), 이벤트가 없지만 있다고 판별하는 FP(False Positive) 등 네 가지 요소가 사용된다.

F-score는 검사의 정밀도에 대한 척도를 나타낸다. 보통 f1-score를 계산하며, 정보 검색, 문서 분류 등에 사용된다.[13] 선택된 항목 중 정답과 관련된 항목이 얼마나 있는지를 나타내는 정밀도(precision), 정답과 관련된 항목 중 얼마나 선택이 되었는지를 나타내는 재현율(recall) 등 두 요소를 통해 계산을 한다.

$$precision = \frac{TP}{TP+FP}. \quad (3)$$

$$recall = \frac{TP}{TP+FN}. \quad (4)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (5)$$

Eq. (5)를 통해 구해진 f-score는 확률의 값을 가지며 0 %~100 %의 값을 갖는다. 100 %에 가까운 값을 가질수록 예측이 잘 되었다고 판단할 수 있으며 추후 시뮬레이션 결과를 확인함에 있어 검증 데이터 또는 시험 데이터에 대한 f-score를 통해 비교한다.

V. 실험 결과

훈련을 위해 앞서 언급된 구글에서 제공되는 오디오 세트를 사용한다. 제공된 오디오 세트는 44.1 kHz의 샘플링 율을 갖는다. DNN의 입력으로 64 멜-스펙트로그램을 사용한다. CNN 블록 및 Bi-RNN 블록 등에서 30 % 드롭아웃을 적용한다. 각 훈련은 최대 100 에폭까지 동작한다.

5.1 데이터 증강 방법 간 파라미터 별 성능

Fig. 4는 TS, PS의 파라미터에 따른 성능을 나타낸다. 하늘색 선은 65.45 %의 데이터 증강을 사용하지 않았을 때의 성능을 나타내며, 주황색 선은 각 방법 간 파라미터에 따른 성능을 나타낸다. DRC에 사용된 curve 및 BM 조건과 혼합 방법에 따른 결과는 Fig. 5에서 확인할 수 있다. Fig. 6을 통해 데이터 증강 방법에 따른 전반적인 성능을 확인할 수 있다.

Fig. 4(a)를 보면 TS는 최소 63.59 %, 최대 71.61 %의 성능 향상을 보인다. 기존 음원에 대해 큰 변화를 주지 않은 0.8, 1.2의 파라미터 값을 적용할 때에 대하여 가장 큰 성능을 보인다. 이는 과도한 TS를 사용하면 오히려 성능이 하락하는 문제가 있음을 알 수 있다. 다른 값을 사용할 경우 대체로 60 %중, 후반의 성능을 보이며 사용된 데이터 증강 방법 중 가장 편차가 크면서 중앙값이 67.35 %로 가장 안 좋은 성능을 보인다.

Fig. 4(b)를 보면 PS는 최소 65.94 %, 최대 74.31 %의 가장 성능을 보인다. 대체로 70 % 초반의 값을 가지며, 반음을 낮춘 경우 성능이 보다 개선됨을 확인 할 수 있다. 중앙값도 71.58 %로 데이터 증강을 사용하지 않은 경우와 비교하여 PS만을 사용할 경우 약 6 % 정도의 성능 향상을 기대할 수 있다.

Table 2는 Fig. 5에서 사용된 파라미터를 나타낸다.

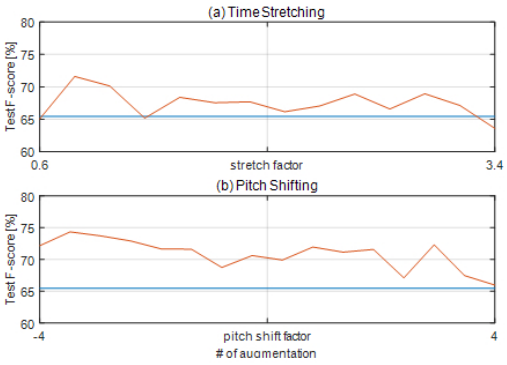

Fig. 4. Performance according to parameters of time stretching and pitch shifting. (a) Time stretching, (b) Pitch shifting.

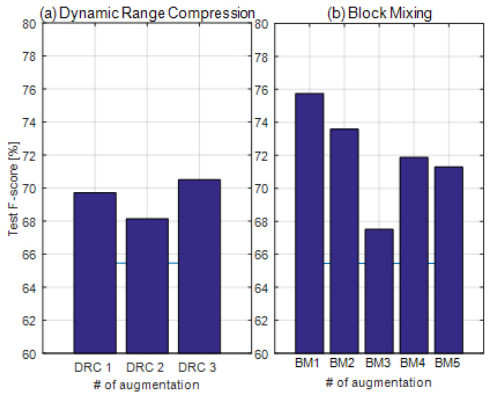

Fig. 5. Performance according to DRC curve and block mixing method. (a) Dynamic range compression, (b) Block mixing.

Table 2. Parameters of block mixing and dynamic range compression.

Aug.	Param.
BM1	cond1, α = 0.5, with norm
BM2	cond2, α = 0.5, with norm
BM3	cond1, α = 0.5, w/o norm
BM4	cond1, α = 0.6, w/o norm
BM5	cond1, α = 0.7, w/o norm
DRC1	[-65, -41, -21]
DRC2	[-80, -60, -40]
DRC3	[-20, -10, 0]

Fig. 5(a)를 보면 DRC curve에 따라 68.14 %에서 70.50 %까지의 성능 향상을 보인다. 블록 혼합에 비해 시험 데이터에 대한 성능의 편차가 적은 것이 확인된다.

Fig. 5(b)는 블록 혼합 방식을 적용하였을 때 성능의 변화를 나타낸다. 블록 혼합의 파라미터 중 'cond1'은 3장의 단일 표식 조건, 'cond2'는 다중 표식 조건을 나타내며, 혼합 방법은 정규화가 적용된 Eq. (2)를 적용하였다. 두 조건에 대하여 75.75 %, 73.59 %로 'cond1'가 2.16 %만큼 우월한 결과를 보인다. 'cond1'의 조건을 가지면서 Eq. (1)과 같이 정규화를 하지 않으며, α를 바꾸어 가며 동작을 한 결과 0.6의 값을 가질 때 가장 성능이 좋았으며, 0.7일 때와 큰 차이가 없었다. 0.5의 값을 갖는 경우 이전 Eq. (2)의 방법으로 혼합을 한 결과보다 성능이 떨어지는 것이 확인된다. 이는 정규화에 따른 오디오 레벨의 증폭이 데이터 증강의 중요한 요소로 동작함을 알 수 있다.

5.2 데이터 증강 방법 별 결합의 성능

데이터 증강 방법을 조합하는 경우의 성능 향상 정도를 확인하기 위해 크게 네 가지 기준으로 구분하여 확인하였다. 첫 번째로 동일한 방법 간 각 증강 방법에서 가장 좋은 성능을 보였던 첫 번째와 두 번째 파라미터들만 사용하여 조합하는 경우이다. 두 번째로 블록 혼합을 제외한 세 가지 증강 방법 중 두 가지 방법에 대하여 가장 좋은 성능을 내었던 파라미터 간의 조합, 또는 한 방법에서는 가장 좋은 성능을 내었던 파라미터와 다른 방법에서는 가장 좋은 성능을 내었던 첫 번째와 두 번째 파라미터 간의 조합, 그 반대의 경우에 대한 조합하는 경우이다. 세 번째로 블록 혼합을 제외한 각 데이터 증강 방법의 가장 좋은 성능을 내었던 파라미터 간의 조합 또는 가장 좋은 성능을 내었던 첫 번째와 두 번째 파라미터 간의 조합이다. 네 번째로 두 번째와 세 번째 방법 간 최고의 성능을 내는 조합에 대하여 블록 혼합을 적용한 경우이다.

Table 3은 데이터 증강 방법 별 결합에 따른 성능을 나타낸다. Fig. 4와 Fig. 5(a), Table 3의 No.1 ~ No.3에서 확인할 수 있듯이, 첫 번째로 동일한 방법 간 가장 좋은 성능을 보였던 첫 번째와 두 번째 파라미터들만 사용하여 조합한 성능이 가장 좋은 성능을 보인 파라미터만 사용한 결과보다 성능이 떨어지는 것이 확인된다. 훈련 데이터를 증강함으로서 데이터의 다양성을 얻었지만, 데이터의 수가 적지만 최적화된 파

Table 3. Performance per data augmentation method and its parameters.

No.	Aug.	Param.	Vali_ f-score	Test_ f-score
1	TS	TS (0.8) TS (1.2)	69.30 %	68.19 %
2	PS	PS (-3.5) PS (-3)	70.16 %	68.60 %
3	DRC	DRC (curve1) DRC (curve3)	71.92 %	69.92 %
4	TS DRC	TS (0.8) DRC (curve3)	67.38 %	66.32 %
		TS (0.8) DRC (curve1, curve3)	70.14 %	65.20 %
		TS (0.8,1.2) DRC (curve3)	69.80 %	66.62 %
5	PS DRC	PS (-3.5) DRC (curve3)	70.83 %	67.65 %
		PS (-3.5) DRC (curve1, curve3)	69.44 %	71.48 %
		PS (-3.5, -3) DRC (curve3)	67.37 %	71.30 %
6	TS PS	TS (0.8) PS (-3.5)	71.25 %	72.90 %
		TS (0.8, 1.2) PS (-3.5)	71.36 %	71.26 %
		TS (0.8) PS (-3.5, -3)	73.66 %	75.28 %
7	TS PS DRC	TS (0.8) PS (-3.5) DRC (curve3)	72.50 %	70.52 %
		TS (0.8, 1.2) PS (-3.5, -3) DRC (curve1,curve3)	68.75 %	67.88 %
8	TS DRC BM	TS (0.8, 1.2) DRC (curve3) BM(cond1)	77.77 %	76.83 %
9	PS DRC BM	PS (-3.5) DRC (curve1,curve3) BM (cond1)	75.39 %	79.16 %
10	TS PS BM	TS (0.8) PS (-3.5, -3) BM (cond1)	77.55 %	77.32 %
11	TS PS DRC BM	TS (0.8) PS (-3.5) DRC (curve3) BM (cond1)	77.94 %	77.37 %

라미터에 대한 데이터 증강을 사용한 것이 더 좋은 성능을 보인다.

두 번째로 No.4 ~ No.6에 해당하는 블록 혼합을 제외한 세 가지 증강 방법 중 두 가지 방법에 대하여 가장 좋은 성능을 내었던 파라미터 간의 조합, 또는 한 방법에서는 가장 좋은 성능을 내었던 파라미터와 다른 방법에서는 가장 좋은 성능을 내었던 첫 번째와 두 번째 파라미터 간의 조합, 그 반대의 경우에 대한 조합하는 경우에 대한 성능을 확인한다. TS를 조합에 사용할 경우 전체적으로 성능이 떨어지는 것이 확인된다. 특히 DRC와 함께 사용 될 경우 66 % 전후의 성능을 보이며 데이터 증강을 적용하지 않았을 때의 성능과 큰 차이가 없는 것이 확인된다. PS를 함께 적용할 경우 대체로 70 % 이상의 성능을 보인다. DRC는 사용 유/무에 따른 경향이 확인되지 않는다.

세 번째로 No.7에 해당하는 블록 혼합을 제외한 각 데이터 증강 방법의 가장 좋은 성능을 내었던 파라미터 간의 조합 또는 가장 좋은 성능을 내었던 첫 번째와 두 번째 파라미터 간의 조합에 따른 성능을 확인한다. 이는 가장 좋은 성능을 내는 파라미터 간 조합의 성능이 좋은지, 두 번째 성능을 내는 파라미터에 대한 데이터 증강을 적용함으로서 보다 훈련 데이터를 증강한 결과가 좋은지를 확인할 수 있다. 첫 번째 방법에서 얻은 결과와 동일하게 가장 좋은 성능을 내는 파라미터만을 사용한 경우가 2.64 % 향상된 성능을 보인다. No.4, No.5, No.7의 결과를 비교함으로서 두 번째 방법에서 얻은 TS에 따른 성능의 감소, PS에 따른 성능의 개선도 함께 확인된다.

네 번째로 No.8 ~ No.11에 해당하는 두 번째와 세 번째 방법 간 최고의 성능을 내는 조합에 대하여 블록 혼합을 적용했을 때의 성능을 확인한다. 이는 블록 혼합에 따른 성능을 확인할 수 있다. 블록 혼합을 적용할 경우 전반적인 성능 향상이 일어난 것이 확인된다. 사용된 네 가지 방법에 대하여 평균 6.695 % 의 성능 향상을 보였다. 가장 성능이 낮게 나온, DRC+TS에 대해 10.21 %, 성능이 좋은 TS+PS에 대해 2.04 %만큼 성능이 향상되었다. 이는 DRC+TS가 일반화되지 않는 문제를 보여줌과 동시에 TS+PS가 일반화되었음을 보여준다. 혼합 과정을 통해 훈련 데이터가 보다 다양해지며, DNN이 일반화된 것을 나타낸다.

VI. 결 론

본 논문에서는 DNN 기반 오디오 표식을 위한 데이터 증강 방법의 연구를 하였다. 이를 위해 네 가지 데이터 증강 방법을 제시하며, 데이터 증강 방법과 파라미터, 방법의 조합에 따른 결과를 확인하였다.

파라미터에 따른 성능을 확인함으로서 각 방법에 따른 최적 파라미터를 확인하였다. 서로 다른 데이터 증강 방법 간의 조합에 따라 PS를 함께 적용한 것이 효율적임을 확인하였고, 데이터를 증강하는 것보다 최적화된 파라미터에 대한 데이터 증강을 적용하는 것이 성능 개선에 도움이 되는 것을 보인다. 블록 혼합을 적용하여 전반적인 성능이 향상됨을 확인할 수 있다.

특정 데이터 및 클래스에 대한 평가이며, 하나의 DNN 구조만을 사용한 점을 들어 본 논문에서 얻은 결과를 일반화하는 데에 한계가 있겠지만, 본 논문에서 사용한 데이터와 유사한 데이터의 경우에는 적용 가능할 것이다.

감사의 글

본 논문은 2018년 정보통신기술진흥센터 지원을 받아 수행된 연구(No.1711055381)의 연구 결과 중 일부이다.

References

1. E. Wold, T. Blum, D. Keislar, and J. Wheaten, “Content-based classification, search, and retrieval of audio,” IEEE Multimedia, **3**, 27-36 (1996).
2. D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: an IEEE AASP challenge,” Proc. of IEEE WASPAA, 1-4, (2013).
3. P. Cano, M. Koppenberger, and N. Wack, “Content-based music audio recommendation,” Proc. ACM 13th, 211-212 (2005).
4. P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, “CHiME-home: A dataset for sound source recognition in a domestic environment,” Proc. of IEEE WASPAA, 15, 2015.

5. J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” in. IEEE Signal Process. Lett., **24**, 279-283(2016).
6. S. Mum, S. Park, D. K. Han, and H. Ko, “Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane,“ Proc. DCASE, 93-97 (2017).
7. R. Seizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, “Large-scale weakly labeled semi-supervised sound event detection,” arXiv preprint arXiv:1807.10501, July (2018).
8. M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Trans. Signal Process., **45**, 2673-2681(1997).
9. G. E. Dahl, T. N. Sainath, and G. E. Hinton “Improving DNNs for LVCSR using rectified linear units and dropout,” Proc. IEEE ICASSP, 8609-8613 (2013).
10. M. Hilsamer and S. Herzog, “A statistical approach to automated offline dynamic processing in the audio mastering process,” In. DAFx, 35-40 (2014).
11. Dolby E, “Standards and practices for authoring Dolby Digital and Dolby E bitstreams,” Dolby Labortories, Inc. 2002.
12. J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” Proc. IEEE ICASSP, 776-780 (2017).
13. S. M. Beitzel, *On Understanding And Classifying Web Queries*, (Ph.D. thesis, Illinois Institute of Technology, Chicago, IL, CiteSeerX 10.1.1.127.634, 2006).

저자 약력

▸ 김 범 준 (Bum-Jun Kim)

2017년 2월: 연세대학교 컴퓨터공학 학사
2017년 3월 ~ 현재: 연세대학교 전산학과 석사과정

<관심분야> 디지털 신호처리, 음질 개선, 음성 신호처리, 적응 신호처리

▸ 문 현 기 (Hyeongi Moon)

2013년 2월: 연세대학교 전기전자공학부 학사
2013년 3월 ~ 현재: 연세대학교 전기전자공학부 석박 통합과정

<관심분야> 오디오 신호처리, 3D 오디오, 오디오 부호화

▸ 박 성 욱 (Sung-Wook Park)

1993년 2월: 연세대학교 전자공학과 학사
1995년 2월: 연세대학교 신호처리 석사
1998년 8월: 연세대학교 신호처리 박사
2009년 3월 ~ 현재: 국립강릉원주대학교 전자공학과 부교수

<관심분야> VLSI 신호처리, 멀티미디어 시스템

▸ 박 영 철 (Young cheol Park)

1986년 2월: 연세대학교 전자공학과 학사
1988년 2월: 연세대학교 전자공학과 석사
1988년 2월: 연세대학교 전자공학과 박사
2002년 3월 ~ 현재: 연세대학교 컴퓨터정보통신공학부 교수

<관심분야> 디지털 신호처리, 오디오 신호처리, 음성 신호처리, 적응 신호처리