Weakly labeled 데이터 기반 음향 이벤트 인식 알고리즘 성능 분석

임우택, 서상원, 박수영, 정영호, 이태진
한국전자통신연구원 실감 AV 연구그룹
wtlim@etri.re.kr

Performance analysis of acoustic event detection algorithm using weakly labeled data

Wootaek Lim, Sangwon Suh, Sooyoung Park, Youngho Jeong, Taejin Lee
Realistic AV Research Group
Electronics and Telecommunications Research Institute (ETRI)

요 약

음향 이벤트 인식 기술은 오디오 신호에서 음향 이벤트를 예측하는 기술로, 최근 대용량 데이터베이스의 배포, 인식 알고리즘과 하드웨어의 발전, 관련 인식 대회 등에 힘입어 많은 연구가 이루어지고 있는 분야이다. 본 논문에서는 음향 장면 및 이벤트 인식 관련 대회인 DCASE 챌린지에 대하여 기술하고, 약한 레이블 기반의 데이터를 학습해 강한 레이블을 예측하는 DCASE 챌린지 과제 4 에 대하여 설명한다. 또한 DCASE 챌린지 과제 4 에 제출된 다양한 음향 이벤트 인식 알고리즘과 데이터베이스의 종류에 따른 성능을 비교하여 음향 이벤트 인식 성능을 분석한다.

1. 서론

음향 이벤트 인식 기술은 오디오 신호에서 발생하는 이벤트 종류를 찾는 문제로, 최근 많은 연구가 이루어 지고 있다 [1]. 음향 이벤트 인식을 위해 사용되는 오디오 신호는 시계열 데이터이기 때문에 음향 이벤트가 시작하는 시간과 끝나는 시간이 존재하며, 이를 시작점(onset)과 끝점(offset)이라고 한다. 따라서 음향 이벤트 데이터베이스(database)를 만들 때는 발생한 이벤트 클래스(class)를 정의하고 해당 이벤트에 대한 시작점과 끝점을 태깅(tagging)하는 것이 일반적이며, 이렇게 시작점과 끝점이 태깅 된 데이터를 강한 레이블(strong label)이 존재하는 데이터라고 한다. 그러나 강한 레이블이 존재하는 데이터베이스를 만드는 일은 사람의 시간과 노력, 비용 등이 많이 소요되는 작업이기 때문에, 음향 이벤트 데이터베이스에는 약한 레이블(weak label)만 존재하는 경우가 많다. 여기서 약한 레이블이란, 하나의 오디오 클립에 대해서 발생한 이벤트 클래스는 태깅되어 있지만, 그 이벤트의 시작점과 끝점은 태깅되어 있지 않은 경우를 말한다.

전술한 바와 같이, 음향 이벤트 인식을 위한 데이터베이스를 구성할 때 강한 레이블을 태깅하는 것은 사람의 시간과 노력이 많이 필요하기 때문에 제작 과정에 어려움이 있다. 따라서 비교적 태깅하기 쉬운 약한 레이블을 포함하는 데이터베이스가 상대적으로 많이 존재하고, 강한 레이블을 포함하는 데이터베이스는 많이 공개되어 있지 않으며 활용 가능한 데이터베이스라도 그 크기가 매우 작다 [1]. 따라서 약한 레이블만을 가지고 음향 이벤트 및 시작점과 끝점까지 예측하고자 하는 연구가 많이 수행되고 있다.

본 논문에서는 음향 이벤트의 약한 레이블을 기반으로 학습하여 강한 레이블을 추정하는 음향 이벤트 인식 대회인 DCASE(Detection and Classification of Acoustic Scenes and Event), 과제(task) 4 에 대해 기술하고, 다양한 모델 구조와 데이터베이스 종류에 따른 음향 이벤트 인식 성능을 분석한다.

2. DCASE Challenge task4

DCASE 는 IEEE AASP(Audio and Acoustic Signal Processing) 기술 위원회가 후원하며, 음향 장면 분류 및 음향 이벤트 인식 대회인 챌린지(challenge)와 관련 연구 결과를 발표하고 토론하는 워크숍(workshop)으로 구성된다. 기존 DCASE 2018 챌린지 과제 4: ‘Large-scale weakly labeled semi-supervised sound event detection in domestic environments’ 와 비교하여 DCASE 2019 챌린지 과제 4 는 ‘Sound event detection in domestic environments’ 로 제목과 세부 내용면에서 조금 변경되었으나, 큰 틀에서는 AudioSet 데이터베이스[2]를 이용하여 가정 환경에서 발생 가능한 ‘Speech’, ‘Vacuum cleaner’ 등 10 개 종류의 음향 이벤트를 인식하는 동일한 목적을 갖고 있다 [3, 4]. 좀 더 세부적으로 볼 때 과제 4 의 핵심 목표는 적은 규모의 약한 레이블 데이터와 대용량의 태깅되지 않은 데이터 등을 활용하여 주어지는 테스트 데이터에 대한 시작점과 끝점을 포함하는 강한 레이블을 예측하는 것이다. 이는 좀 더 실생활과 유사한 환경의 데이터를 가지고 학습하여 실제 서비스에 적용 가능한 기술을 개발 하고자 하는 의미가 있다.

DCASE 2018 과 비교하여 DCASE 2019 챌린지에서 제공되는 데이터베이스에는 약간의 변화가 있다. 먼저 약한 레이블만 태깅되어 있는 ‘weakly_labeled’ 데이터와, 인식해야 하는 목표 음향 이벤트 클래스를 포함하고 있지만 태그 정보는 전혀 없는 ‘unlabeled_in_domain’ 데이터는 동일하게 제공된다.

표 1. DCASE Challenge task4 데이터베이스 구성 및 변화

Database			2018	2019
Development dataset	Training set	Weakly labeled set [A]	- 1578 clips (2244 class occurrences) - w/ weak labels	- 1578 clips (2244 class occurrences) - w/ weak labels
		Unlabeled in domain set [B]	- 14412 clips - no labels	- 14412 clips - no labels
		Unlabeled out of domain set	- 39999 clips - no labels	- None
		Synthetic strongly labeled set [C]	- None	- 2045 clips (6032 events) - w/ strong labels
	Test set/Validation set		- 288 clips (906 events) - w/ strong labels	- 1168 clips (4093 events) - w/ strong labels
Evaluation dataset			- 880 clips (3187 events) - w/ strong labels	- TBA - w/ strong labels

반면에 인식해야 하는 목표 음향 이벤트 클래스 외의 음향 이벤트까지도 포함하며 태그 정보가 제공되지 않았던 'unlabeled_out_of_domain' 데이터가 사라지고, AudioSet 이 아닌 다른 데이터베이스[5, 6]를 이용하여 합성된 음향 이벤트 데이터인 'synthetic' 데이터가 새로 추가 되었다. 또한 음향 이벤트 인식 알고리즘의 성능을 검증하기 위한 'validation' 데이터는 DCASE 2018 챌린지 과제 4 의 테스트 데이터(test set)와 평가용 데이터(evaluation set)이 통합되어 제공된다. 이렇게 주어진 3 가지 종류의 데이터를 이용하여 테스트 데이터에 대해 음향 이벤트 인식 성능을 평가하며, 주어진 데이터들을 어떻게 활용하는지 여부가 챌린지에서 해결하고자 하는 중요 목표 중에 하나이다. 최종 평가용 데이터는 추후 공개될 예정이며, 표 1 과 같이 DCASE 챌린지 과제 4 의 데이터베이스의 구성 및 변화를 정리하였다.

3. 음향 이벤트 인식 알고리즘 성능 분석

본 절에서는 2 절에서 서술한 DCASE 챌린지 과제 4 의 음향 이벤트 인식 알고리즘의 성능을 분석하기 위해 3 가지 시스템을 비교하였다. 성능 분석에 사용된 첫 번째 시스템은 DCASE 2019 챌린지 과제 4 에서 제공하는 기준(baseline) 시스템으로, DCASE 2018 챌린지 과제 4 에 제출된 1 등 모델[7]을 바탕으로 네트워크 구조 등을 간략화하여 제공된다[4]. 해당 모델은 컨벌루션 회귀 신경망(CRNN) 기반의 Mean-teacher 모델을 사용하였다. 성능 분석을 위한 두 번째 시스템은 영국의 Surrey 대학교 연구팀에서 공개한 시스템이 사용되었다 [8]. 이 시스템은 기본적으로 컨벌루션 신경망(CNN)을 기반으로 레이어 구조에 변화를 주어 성능을 비교하였으며, 데이터베이스에 따른 성능 또한 비교하였다. 마지막으로 세 번째 시스템은 DCASE 2018 챌린지 과제 4 에 제출된 인셉션(Inception) 모듈을 포함하는 컨벌루션 회귀 신경망 모델 기반의 음향 이벤트 인식 시스템이다 [9]. 이 시스템은 컨벌루션 신경망에서 인셉션 모듈을 활용하여 여러 크기의 수용영역(receptive field)을 동시에 분석함으로써 음향 이벤트 인식을 수행하며, 추가적으로 인식 성능 향상을 위한 다양한 후 처리 방법을 제안하였다.

앞서 기술한 3 가지 음향 이벤트 인식 알고리즘의 성능 분석 결과는 표 2 와 같다. 각 알고리즘의 이름과 모델의 구조가 분류되어 있으며, 모델 별로 학습(training)에서 사용한 데이터베이스 구성을 정리하였다. 데이터베이스 [A], [B], [C]는 표 1 에서 기재된 바와 같이 각각 약한 레이블을 태그 정보로 갖고 있는 'weakly_labeled' 데이터와 태그 정보가 없는 'unlabeled_in_domain' 데이터, 그리고 다른 데이터베이스를 이용해 합성되어 강한 레이블을 태그 정보로 갖고 있는 'synthetic' 데이터이다. 알고리즘 별 인식 성능을 평가하기 위해 활용된 데이터는 DCASE 2019 챌린지 과제 4 의 검증용 데이터인 'validation' 데이터가 사용되었다. 인식 성능은 챌린지와 동일한 기준인 이벤트 기반(event-based)의 F-점수(F-score)를 이용하여 평가되었으며, 보다 상세한 분석을 위한 보조 지표로 세그먼트 기반(segment-based) F-점수를 함께 제시하였다.

음향 이벤트 인식 알고리즘의 성능을 분석한 결과는 다음과 같다. 먼저 DCASE 2019 챌린지 과제 4 에서 제공하는 기준 시스템은 [A, B, C] 데이터를 사용하며 23.5%의 성능을 보였다. 다음으로 Surrey 시스템을 [A] 데이터 만을 사용하여 학습한 결과는, 모델 구조 별로 편차가 있으나 최대값 풀링(max pooling)을 사용한 CNN9-Ⅱ 모델에서 가장 우수한 성능인 24.1%의 인식률을 보였다. 또한 이 모델은 세그먼트 기반 F-점수에서 다른 알고리즘 대비 높은 인식률을 보여 약한 레이블 인식에 강점을 보였다. 동일한 Surrey 시스템을 [C] 데이터 만을 사용하여 학습하고 평가한 결과는 12.4%의 낮은 인식률을 보였다. 이는 다른 환경에서 녹음된 데이터베이스로 학습하고 예측을 수행하기 때문에 데이터 특성이나 잡음의 분포 형태가 달라 발생하는 성능 하락으로 볼 수 있다. 마지막으로 세 번째 시스템인 인셉션 모듈을 포함하는 컨벌루션 회귀 신경망 모델 기반의 음향 이벤트 인식 알고리즘은 [A] 데이터만을 활용하여 학습한 경우 21.8%, [A, B] 데이터를 활용하여 학습한 경우 23.8%의 인식 성능을 보여 데이터 [B]를 활용함으로써 약 2%의 성능 향상을 얻을 수 있음을 확인했다. 추가적으로 동일한 모델과 데이터 구조에 셀프어텐션(self-attention)[10]을 회귀 신경망 다음 레이어로 추가하여 각각 2%, 1%의 성능 향상을 얻을 수 있었다.

표 2. 음향 이벤트 인식 알고리즘 성능 비교

System	Model	Database	Event-based F-score	Segment-based F-score
DCASE Baseline 2019 [4, 7]	Mean-teacher	[A, B, C]	23.5%	54.7%
Surrey [8]	CNN5	[A]	18.0%	59.9%
	CNN9- I	[A]	20.0%	58.6%
	CNN9- II	[A]	24.1%	63.0%
	CNN13	[A]	17.0%	58.3%
	CNN9- II	[C]	12.4%	41.3%
Inception CRNN [9]	CRNN w/ Inception	[A]	21.8%	55.5%
		[A, B]	23.8%	58.2%
		[C]	15.6%	40.0%
	CRNN w/ Inception + Self-attention	[A]	23.8%	56.4%
		[A, B]	24.8%	55.1%

마지막으로 [C] 데이터 만을 활용하여 학습한 결과는 15.6%의 성능을 보여 Surrey 시스템과 동일하게 데이터베이스의 변화에 따른 성능 편차를 확인하였다.

4. 결론

본 논문에서는 음향 장면 및 이벤트 인식 관련 대회인 DCASE 챌린지와 약한 레이블 기반의 데이터를 학습해 강한 레이블을 예측하는 과제 4 에 대하여 설명하였다. 또한 해당 과제에 제출 된 다양한 음향 이벤트 인식 알고리즘의 성능을 분석하였으며 과제 4 에서 제시된 데이터베이스의 활용과 모델의 구조에 따른 성능을 비교하여 음향 이벤트 인식 성능을 분석하였다.

감사의 글

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술 진흥센터의 지원을 받아 수행된 연구임 (No. 2017-0-00050, 신체기능의 이상이나 저하를 극복하기 위한 휴먼 청각 및 근력 증강 원천 기술 개발)

참고문헌

[1] 서상원, 임우택, 정영호, 이태진, 김휘용, “실생활 음향 데이터 기반 이중 CNN 구조를 특징으로 하는 음향 이벤트인식 알고리즘,” 방송공학회논문지 23(6), 2018.

[2] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio set: an ontology and human-labeled dataset for audio events,” In Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), 776-780

[3] “DCASE2018,” http://dcase.community/challenge2018/.

[4] “DCASE2019,” http://dcase.community/challenge2019/.

[5] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. “Freesound datasets: a platform for the creation of open audio datasets,” In Proc. 18th International Society for Music Information Retrieval Conference (ISMIR 2017), 486-493.

[6] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Toon van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers. “The SINS database for detection of daily activities in a home environment using an acoustic sensor network,” In Proc. Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017), 32- 36.

[7] Lu JiaKai, “Mean teacher convolution system for dcase 2018 task 4,” Technical Report, DCASE2018 Challenge

[8] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley, “Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems,” arXiv preprint arXiv:1904.03476 (2019).

[9] Wootaek Lim, Sangwon Suh, and Youngho Jeong, “Weakly labeled semi-supervised sound event detection using CRNN with inception module,” In Proc. Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE 2018), 74- 77.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention is all you need,” In Proc. Neural Information Processing Systems (NIPS 2017).