오디오 음량 자동 제어를 위한 콘텐츠 분류 기술 개발

*이영한 **조충상 ***김제우
전자부품연구원
{*yhlee, **ideafisher, ***jwkim}@keti.re.kr

Audio Contents Classification based on Deep learning for Automatic Loudness Control

*Young Han Lee **Choongsang Cho ***Je Woo Kim
Korea Electronics Technology Institute (KETI)

요약

오디오 음량을 자동으로 제어하는데 있어 음성이 있는 구간에 대해서 음량이 급격히 줄어드는 것을 막기 위해 콘텐츠에 대한 분석이 필요하다. 본 논문에서는 방송 음량을 조절을 위한 세부 기술로 딥러닝 기반의 콘텐츠 분류 기술을 제안한다. 이를 위해 오디오를 무음, 음성, 음성/오디오 혼합, 오디오의 4개로 정의하고 이를 처리하기 위한 mel-spectrogram을 이용하여 2D CNN 기반의 분류기를 정의하였다. 또한 학습을 위해 방송 오디오 데이터를 활용하여 학습/검증 데이터 셋을 구축하였다. 제안한 방식의 성능을 확인하기 위해 검증 데이터셋을 활용하여 정확도를 측정하였으며 약 81.1%의 정확도를 가지는 것을 확인하였다.

1. 서론

디지털 방송으로의 전환 후, 디지털 방송 음량에 대한 기준이 없는 상태로 방송을 하면서 TV 시청자들은 채널 간 또는 프로그램 간의 전환 시에 오디오 음량 레벨의 급격한 변화로 인해 많은 불편을 겪었다. 이를 해결하기 위해서 우리나라에서는 2014년 5월 방송법을 개정하여 '디지털 텔레비전 방송프로그램 음량 등에 관한 고시'를 통해서 2016년 5월부터 국제 권고 수준인 평균 음량 -24 LKFS 수준에 맞춰 방송 프로그램을 송출하도록 규제하고 있다 [1].

우리나라의 방송 프로그램의 음량 기준은 방송 프로그램에서 주관적으로 인지하게 되는 음량을 객관적으로 표시할 수 있는 방송 음량 측정 방법인 BS.1770-3 [2], R128[3]을 기반으로 미국, 캐나다, 일본, EU 등의 법률과 기준을 참고하여 ITU 권고 기준에 준수하고 있다.

현재 국내를 비롯한 국외에서는 방송 프로그램의 음량 기준을 준수하기 위해 제작단에서 맞추는 방법, 송출 전 라우드니스 조절 장비를 이용하여 조절하는 방법과 기존 제작된 방송을 라우드니스 변환 장비를 이용하여 재생성하는 방법의 세 가지 방식으로 처리하고 있다.

특히, 세 가지 방식 중 라우드니스 변환 장비를 이용한 재생성 방법은 실시간 방식인 라우드니스 조절 장비에 비해 원음 왜곡이 적으며 정확하게 음량을 맞출 수 있는 것이 장점이다. 또한 실시간 송출이 필요한 일부 방송을 제외하고 VOD 서비스 등 다양한 분야에 적용할 수 있기 때문에 활용성이 높다. 하지만 일반적으로 방송 음량을 일률적으로 조절하기 때문에 평균적으로 높은 음량으로 인해 작은 음량이 더욱 낮게 조절되는 단점이 존재한다 [4]. 이를 해결하기 위해 본 논문에서는 음성의 명료도를 강화할 수 있는 사람의 청각 특성을 고려하여 콘텐츠를 분류하고 이를 활용하여 오디오 음량을 적응적으로 제어하는 알고리즘의 활용할 수 있는 오디오 콘텐츠 분류 기술을 제안한다.

2. 음량 자동 조절을 위한 콘텐츠 분류기

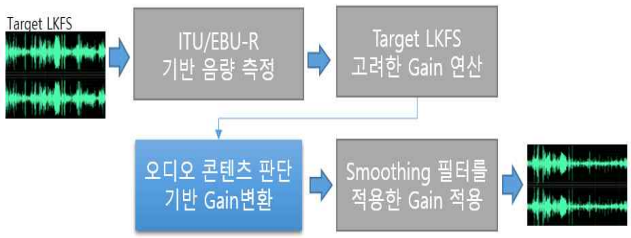

그림 1. 콘텐츠 판단 기반 적응적 오디오 음량 제어 알고리즘

오디오 콘텐츠 분류기는 그림 1에서 콘텐츠 판단 기반 gain 변환 모듈의 세부 구성으로 결과물이 오디오 음량 측정/제어 기술과 연동하여 동작하기 때문에 입력신호의 길이를 ITU-R BS.1770-3 [2]에서 음량을 측정하는 길이인, 400 ms 로 정의하였다. 이에 대해 512 sample shift의 2048-point STFT을 이용한 주파수 변환을 하였으며 청각적 특성을 고려하기 위해 128개의 mel-filterbank를 적용하여 2채널의 (128×38) 크기의 2차원 데이터인 mel-spectrogram을 생성하였다. 출력 분류 클래스는 무음, 음성, 음성/오디오 혼합, 오디오의 총 4개로 정의하였다.

표 1. 오디오 콘텐츠 분류기 구성표

Baseline 모델		계층 증가 모델	
Layer	Spec.	Layer	Spec.
Conv.	C2->C16	Conv.	C2->C16
BN	C16	BN	C16
ReLU	-	ReLU	-
Pooling	(2,2)	Pooling	(2,2)
Conv.	C16->C32	Conv.	C16->C32
BN	C32	BN	C32
ReLU	-	ReLU	-
Pooling	(2,2)	Pooling	(2,2)
-	-	Conv.	C32->C64
-	-	FC	1000
FC	4	FC	4

표 1은 baseline으로 구성한 모델과 이를 활용하여 계층을 추가한 모델에 대한 상세 정의이다. 기본 모듈은 커널 크기를 5로 하는 2D convolutional layer (Conv.)와 batch normalization (BN), rectified linear unit (ReLU)으로 구성되어 있으며 하나의 모듈을 거칠 때마다 2D max-pooling을 통해 채널별 파라미터를 축소시키는 반면에 채널 수를 2배씩 증가시켰다. 최종단에서 fully -connected layer를 이용하여 최종 클래스의 확률을 획득하였다.

3. 모델 학습 및 성능평가

모델 학습을 위해 오디오 데이터를 구축하였다. 구축에 사용한 음원은 방송 콘텐츠 중, 영화, 뉴스, 스포츠 중계의 장르에 대해서 선정하였으며 정답지 작업을 위해 음원을 청취 후 클래스 레이블 작업을 진행하였다. 구축된 데이터는 400 ms 프레임 단위 기준으로 약 38,000여 개의 데이터로 이루어 졌으며, 훈련/검증/테스트 데이터셋의 비율을 7:2:1 로 하였다.

딥러닝 프레임워크로는 PyTorch v0.3 [5]을 사용하였으며 오디오 전처리를 위해서는 librosa 라이브러리[6]를 활용하였다. 오디오 콘텐츠 분류기의 모델 학습을 위해 beta_1 0.9, beta_2 0.999 설정의 Adam Optimizer를 사용하였으며 초기 learning-rate은 0.0001로 하였다. 총 학습은 10,000 여회의 반복에 해당하는 12 epochs로 설정하였으며 mini-batch 의 크기는 32 샘플로 정의하여 학습하였다.

그림 2는 제안한 baseline 모델의 정확도 그래프이다. 훈련 데이터(주황색) 및 검증 데이터(붉은색)가 80% 대로 수렴하는 것을 확인할 수 있다. 최종적으로 12 epoch를 통해 획득한 콘텐츠 분류 정확도는 81.1% 이다.

본 논문에서는 모델 용량을 증가시켰을 때의 성능을 확인하기 위해 표 2에 언급한 계층 증가 모델에 대해 학습을 진행하였다. 학습에 사용한 하이퍼 파라미터는 이전과 동일하게 설정하였으며 학습 횟수를 30 epochs 기준으로 하였다. 그림 3에서 확인할 수 있듯이, 모델의 용량이 커짐에 따라 훈련 데이터의 분류 성능(붉은선)은 92%대로 수렴구간 없이 증가하는 반면, 검증 데이터의 성능(분홍선)은 오히려 20 epoch 이후 감소함을 나타내었다. 이러한 현상은 훈련 데이터로 모델이 과학습되는 overfitting 현상으로 데이터를 증가시키거나 정규화 기법이 필요한 것으로 나타난다.

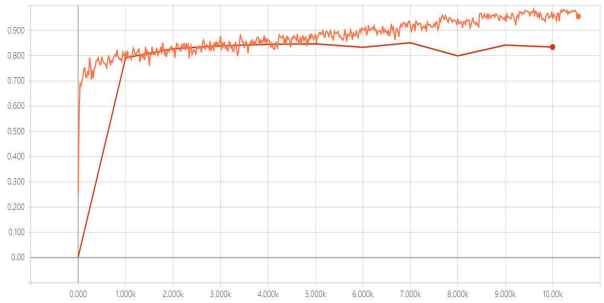

그림 2. Baseline 모델의 정확도 그래프

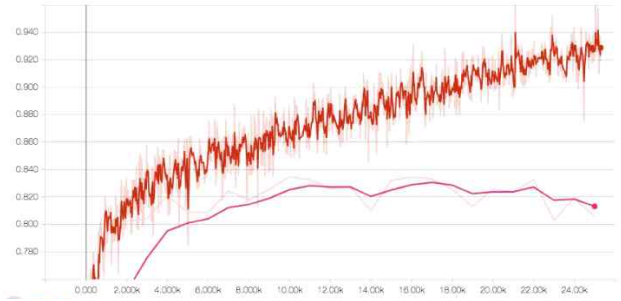

그림 3. 계층 증가 모델의 정확도 그래프

4. 결론 및 향후 계획

본 논문에서는 오디오 음량 자동 제어를 위한 콘텐츠 분류 기술을 제안하였다. 제안된 알고리즘은 2D CNN을 통해 구현되었으며, 검증 데이터셋에 대해 81.1 %의 정확도를 가지는 것을 확인하였다. 이는 오디오 음량 자동 제어에 있어 콘텐츠별 제어 정책을 분리함으로써 명료도를 확보하는 자동 제어 기술을 개발하는데 포함될 수 있다.

향후 계획으로 confusion matrix 분석을 통해 클래스 재정의를 진행할 예정이며, 모델 복잡도 증가에 따라 훈련/검증 성능 차이를 줄이기 위한 추가 데이터 구축 및 모델 최적화 연구를 진행하고자 한다.

Acknowledgement

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2017-0-00788, 딥러닝기반 지능형 오디오 분석을 통한 적응적 오디오 콘텐츠 변환 솔루션 개발)

참고문헌

[1] 미래창조과학부고시 제2014-87호, 디지털 텔레비전 방송프로그램 음량 등에 관한 기준, 2014년 11월.
[2] ITU-R Rec. BS.1770-3, "Algorithms to measure audio programme loudness and true-peak audio level," Aug, 2012.
[3] EBU R128, "Loudness Normalization and Permitted Maximum Level of Audio Signals", Jun. 2014.
[4] 이영한, 조충상, 김제우, "청각 특성을 고려한 적응적인 오디오 음량 자동 제어 기술 개발," 게재 예정.
[5] Paszke, et al. "Automatic differentiation in PyTorch," In NIPS workshop, 2017.
[6] McFee, et al. "librosa: Audio and music signal analysis in python," In Proc. of 14th python in science conference, pp.18-25, 2015.