딥러닝을 이용한 오디오 콘텐츠 분석 기반의 자동 음량 제어 기술 개발

*이영한 **조충상 ***김제우

전자부품연구원

{*yhlee, **ideafisher, ***jwkim}@keti.re.kr

Development of Automative Loudness Control Technique based on Audio Contents Analysis using Deep Learning

*Young Han Lee **Choongsang Cho ***Je Woo Kim

Korea Electronics Technology Institute (KETI)

요약

국내 디지털 방송 프로그램은 2016년 방송법 개정 이후, ITU-R / EBU에서 제안한 측정 방식을 활용하여 채널 및 프로그램 간의 음량을 맞추어 제공되고 있다. 일반적으로 뉴스나 중계와 같이 실시간으로 음량을 맞춰야 하는 분야를 제외하고는 평균 음량을 규정에 맞춰 송출하고 있다. 본 논문에서는 일괄적으로 평균 음량을 맞출 경우 발생하는 저음량의 명료도를 높이기 위한 기술을 제안한다. 즉, 방송 음량을 조절하는 기술 중의 하나로 오디오 콘텐츠를 분석하여 구간별 음량 조절 정도를 달리함으로써 저음량에서의 음성은 상대적으로 높은 음량을 가지고 배경음악 등을 상대적으로 낮음 음량을 가지도록 생성함으로써 명료도를 높이는 방식을 제안한다. 제안한 방식의 성능을 확인하기 위해 오디오 콘텐츠 분석 정확도 측정과 오디오 파형 분석을 실시하였으며 이를 통해 기존의 음량 제어 기술과 비교하여 음성 구간에 대해 음량을 증폭시키는 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 '디지털 텔레비전 방송프로그램 음량 등에 관한 고시'의 방송법을 개정하였다. 이 고시에서는 방송음량 측정 방법인 BS.1770-3 [1]를 이용하여 2016년 5월부터 국제 권고 수준인 평균 음량 -24 LKFS 수준에 맞춰 방송 프로그램을 송출하도록 규제하고 있다 [2].

현재 국내를 비롯한 국외에서는 방송 프로그램의 음량 기준을 준수하기 위해 다음의 세 가지 방법으로 음량 기준을 준수하고 있다. 첫 번째는 방송 프로그램의 제작 단계에서 음량 기준에 맞춰서 오디오의 음량을 조절하여 방송 프로그램을 제작하는 방법이고, 두 번째는 방송 송출 바로 전단계인 송출용 오디오 인코더(일반적으로 비디오 인코더 내에 포함) 앞에서 '라우드니스 조절 장비'를 설치하여 실시간으로 송출되는 방송 프로그램의 음량을 실시간으로 자동 조절하는 방법이다. 마지막으로 기존 제작된 방송 프로그램을 '라우드니스 변환 장비'를 사용하여 파일기반으로 제작된 방송 프로그램의 음량을 기준에 맞춰 재생성하는 방법이다. 라우드니스 조절 장비를 이용하는 방법은 실시간으로 입력되는 오디오 데이터를 이전 입력 오디오 데이터로부터 예측하여 음량을 제어하기 때문에 원 음원의 왜곡이 불가피하게 발생하게 된다. 하지만, 라우드니스 변환 장비의 경우에는 일반적으로 프로그램의 전체 오디오 데이터를 입력받아 먼저 전체 오디오의 음량을 측정한 후에 음량 기준(-24 LKFS)에 맞춰 선형적으로 음량 이득(Gain)을 조절하므로 원 음원의 왜곡 없이 음량을 조절할 수 있는 장점을 갖는다.

그러나, 기존 제작된 방송 프로그램들은 음량 기준에 대한 고려 없이 프로그램을 제작한 경우가 많기 때문에 '라우드니스 조절 장비' 또는 '라우드니스 변환 장비'를 적용하여 출력되는 오디오 데이터들은 "사람의 목소리가 작아져서 소리의 명료도 및 의사 전달 기능이 저하"되는 경우가 다수 발생하고 있다. 즉, 최근 방송 프로그램의 오디오 음량에 대한 시청자들의 불만은 규제 이전의 '채널 간 또는 프로그램 간의 음량 불균형'에서 규제 이후 '전체적인 음량의 감소 또는 음성의 상대적 음량 감소로 인한 명료도 하락'으로 전환되었다.

본 논문에서는 이러한 시청자들의 불만을 해소하기 위해서 상대적으로 음성의 명료도를 강화할 수 있도록 오디오 콘텐츠를 분석하여 음성에는 상대적으로 높은 음량을 할당할 수 있는 알고리즘을 제안한다.

2. 오디오 콘텐츠 기반 자동 음량 제어 기술

딥러닝을 이용한 오디오 콘텐츠 분석 기반의 자동 음량 제어 기술은 그림 1과 같이 오디오 콘텐츠 분석 모듈, 음량 측정 모듈, 자동 음량 제어 모듈로 구성으로 되어 있다. 먼저 방송 음량이 오디오 콘텐츠 분석 모듈 및 음량 측정 모듈로 입력된다. 본 연구에서는 음량 측정 모듈의 단위와 콘텐츠 분석 단위를 일치시키기 위해 400 ms 기준으로 sub-frame을 구성하였고 이는 48 kHz 기준으로 19,200샘플에 해당한다. 콘텐츠 분석 모듈에서는 각 400 ms 별 오디오 분석 결과를 자동 음량 제어 모듈로 제공하며, 자동 음량 제어 모듈은 분석된 결과와 음량 측정 결과를 이용하여 입력된 신호의 음량을 조절한다. 음량 측정 모듈은 BS. 1770-3을 따르며 나머지 두 모듈에 대한 상세 설명은 다음과 같다.

그림 1. 콘텐츠 판단 기반 적응적 오디오 음량 제어 알고리즘

1. 오디오 콘텐츠 분석 모듈

오디오 콘텐츠 분석 모듈의 입력 신호는 Mel-spectrogram 으로, 각 채널별 독립적으로 변환하며 차수는 128로 설정한다. 최종적으로 2x128x38의 3차원 tensor를 생성한 후 CNN 기반의 딥러닝 네트워크에 전달하여 분석 결과를 얻는다. 제안한 방식에서는 Convolutional layer - Batch Normalization - ReLU (Rectified Linear Unit) - Pooling 으로 구성된 Conv. Block을 정의하였으며 [3] 이를 4회 반복, 연결하였다. 최종적으로는 Fully-connected layer를 삽입한 후 4개의 음원 종류에 맞도록 결과를 도출하였다.

2. 자동 음량 제어 모듈

오디오 음량 자동 제어 모듈은 목표 음량, 딥러닝을 통해 얻어진 분석 결과와 오디오 신호를 입력으로 받으며 목표 음량에 맞추어 자동 제어된 오디오 신호를 실시간 출력한다. 특히, 제안된 자동 제어 모듈은 컨트롤된 오디오 음량을 피드백 하여 적응적으로 오디오 음량을 컨트롤하는 구조를 갖는다. 입력된 오디오 신호는 목표 오디오 음량과 컨트롤된 신호의 음량을 비교하여 전체적인 컨트롤을 위한 이득 연산과정을 수행하며, LKFS 단위의 목표 음량과 피드백된 음량의 비율을 선형적 이득 값으로 변환하여 오디오 프레임 단위의 정밀 오디오 음량 컨트롤을 과정을 수행하게 된다. 분석된 오디오 콘텐츠 특성은 정밀 컨트롤 단계에서 적용되며, 이를 위해 분석 결과가 음성일 경우 정밀 컨트롤 이득을 정해진 범위 내에서 적응적으로 증가시키는 시키며 프레임간의 이득 스무딩 과정을 통해 컨트롤된 오디오 신호를 얻는다.

3. 콘텐츠 분석 성능평가 및 음질 평가

오디오 콘텐츠 분석 모듈을 개발하기 위해 Database를 구축하였다. 출력 결과는 무음, 음성, 음성/배경음악 혼합, 배경음악의 총 4개로 정의하였다. 구축에 사용한 음원은 방송 콘텐츠 중, 영화,뉴스, 스포츠 중계 장르에 대해서 진행하였으며 정답지 작업을 위해 음원을 청취 후 클래스 레이블 작업을 진행하였다. 총 약 50,000여개의 데이터를 수집하였으며, 훈련/검증/테스트 데이터셋의 비율을 7:2:1로 구성하였다.

딥러닝 프레임워크로는 PyTorch v0.4 [4]을 사용하였으며 오디오 전처리를 위해서는 librosa 라이브러리[5]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 100 epochs로 설정하였으며 mini-batch 의 크기는 32 샘플로 정의하여 학습하였다.

그림 2는 제안한 오디오 콘텐츠 분석 모델의 정확도 그래프이다. 훈련 데이터 (녹색) 및 검증 데이터(주황색)가 87% 대로 수렴하는 것을 확인할 수 있다. 최종적으로 30-epoch를 통해 획득한 콘텐츠 분류 정확도는 86.1% 이다.

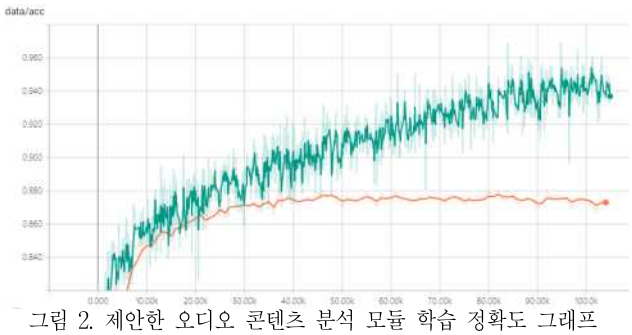

그림 2. 제안한 오디오 콘텐츠 분석 모듈 학습 정확도 그래프

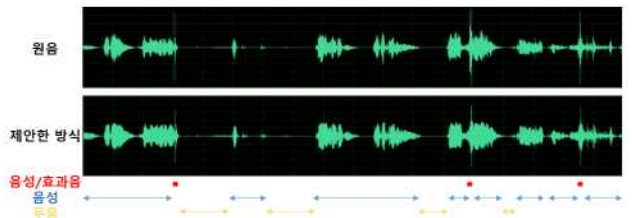

그림 3. 콘텐츠 분석 기반 음량 제어 분석을 위한 오디오 파형

파형 분석을 위해 영화 콘텐츠에 대해서 기존의 방식과 제안한 방식으로 생성한 음원을 비교하였다. 그림 3에서와 같이 음성 구간에서는 제안한 방식이 음량을 높이고 상대적으로 혼합 구간 및 무음 구간은 음량을 줄이거나 유지함으로써 명료도를 향상시키는 것을 확인할 수 있다.

4. 결론 및 향후 계획

본 논문에서는 방송 프로그램의 음량 문제를 해결하기 위해서 오디오 콘텐츠를 분석하고 이 결과에 따라 적응적으로 오디오 음량을 제어하여 음성의 명료도를 높이는 자동 음량 제어 기법을 제안하였다. 제안된 기법은 분석 정확도 및 파형 비교를 통해 목적에 맞도록 음성 신호에 대해 음량을 증가시키는 것을 확인하였다. 향후에는 성능을 향상시키기 위한 방법으로 딥러닝 모델을 개선과 콘텐츠별 음량 제어 알고리즘을 조정하고 오디오 분석 통합 프로그램을 개발하고자 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] ITU-R Rec. BS.1770-3, "Algorithms to measure audio programme loudness and true-peak audio level," Aug, 2012.
[2] 미래창조과학부고시 제2014-87호, 디지털 텔레비전 방송프로그램 음량 등에 관한 기준, 2014년 11월.
[3] 이영한, 조충상, 김제우, "오디오 음량 자동 제어를 위한 콘텐츠 분류 기술 개발," 2018 한국방송·미디어공학회 하계학술대회, 2018년 6월.
[4] Paszke, et al. "Automatic differentiation in PyTorch," In NIPS workshop, 2017.
[5] McFee, et al. "librosa: Audio and music signal analysis in python," In Proc. of 14th python in science conference, pp.18-25, 2015.