딥러닝을 활용한 종단형 자동 이득 제어 알고리즘

전용현, 김석민, 김정훈, 김남수
서울대학교
{yhjeon, smkim, jhkim}@hi.snu.ac.kr, nkim@snu.ac.kr

End-to-end Automatic Gain Control Using Deep Learning

Yong Hyeon Jeon, Seok Min Kim, Jeung Hun kim, and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 적절한 데이터셋 전처리를 통하여 end-to-end로 경량화된 자동 이득 제어 알고리즘 연구이다. 기존의 rule-based 자동 이득 제어 기술은 직전 신호 세기의 이동평균에 기반해 서서히 이득을 조절하는 기법을 사용하였기에 급격한 신호 세기 변화에 빠르게 대처하지 못하는 경향이 있었다. 반면 해당 기법은 단순 이동평균이 아닌 인공 신경망으로 이득을 예측하여 급격한 신호 세기 변화 환경에서 더 자연스러운 이득 제어를 보여주었다.

Ⅰ. 서 론

자동 이득 제어란 입력 음량 변동에도 출력 음량이 상대적으로 일정한 값을 유지하게 하는 자동 제어 기술이다. 이 기술은 입력 음량이 지나치게 요동치지 않도록 자동적인 가변 증폭 제어를 하는 것으로, 음량 감소가 나타나면 이득을 증가시키고, 음량의 급격한 증가가 출력에 반영되지 않도록 이득을 감소시킨다.

전통적인 자동 이득 제어 알고리즘은 신호 세기의 이동평균에 기반해 이득을 제어한다. 이런 알고리즘은 갑작스러운 음량 변화에 실시간으로 대응하기 어렵고, 불필요한 신호를 그 크기가 작다는 이유만으로 과도하게 증폭하기도 한다.

최근 딥러닝의 발전으로 인하여 잡음 제거, 반향 감소, 에코 제거 등은 전통적인 음성 향상 기술에서 딥러닝 기술로 대체되어왔다. 자동 이득 제어 또한 음성 향상 기법의 하나지만, 딥러닝을 이용한 연구가 거의 이루어지지 않고 있던 분야이다.

본 논문에서는 기존의 자동 이득 제어 알고리즘을 대체할 수 있는 종단형 학습 기반 모델을 제시한다. 해당 모델의 코드는 https://github.com/CARNIVAL-IITP/Automatic_gain_control 에서 확인할 수 있다.

Ⅱ. 본론

1) 종단형 자동 이득 제어 학습 데이터 구축

음성 신호의 크기가 변형되지 않은 clean data와 변형된 distorted data의 쌍을 제작하였다.

각 음성의 초기 3초를 20ms 단위 프레임으로 나누어 프레임별 신호의 세기를 계산한 후, 가장 세기가 큰 프레임을 기준으로 신호 크기를 조절하는 방식으로 데이터를 전처리했다. 음성 신호를 $x(t)$, 음성 신호의 총 길이를 T, $split(f(t),\tau)$는 $f(t)$를 시간 τ(=0.02s) 단위로 쪼개어 $\tilde{f}_i(t)|_{i=1}^{n}$로 반환하는 함수라 정의하고

$$\tilde{f}_i(t)|_{i=1}^{n} = split(x(t),\tau)$$

$$P_i = \frac{\int_0^{\tau}(f_i(t))^2 dt}{\tau}, i_{\max} = \arg\max_i P_i$$

$$P_{\max} = P_{i_{\max}}, x_{target}(t) = x(t)/5P_{\max}$$

해당 수식으로 음성 신호의 목표 세기를 설정했다.

전처리가 끝난 데이터를 이용해 clean data(향상 목표)와 distorted data(이득 조절을 수행해야 하는, gain이 왜곡된 신호)의 쌍을 제작했다. 각 데이터는 음성 하나로만 이루어진 I형 데이터와 공백을 사이에 둔 두 음성으로 이루어진 II형 데이터로 이루어졌다.

I형 clean 데이터는 제로 패딩을 통해 10초 길이로 변환하였고, II형 clean 데이터는 공백 길이에 따라 총 길이가 10초에 맞도록 음성 신호를 잘라내어 제작했다. I형 distorted 데이터는 음성 신호가 존재하는 부분(padding 부분을 제외) 중 일부 구간의 세기를 변형해 제작하고, II형 distorted 데이터는 두 음성의 세기를 각각 다른 값으로 변형하여 제작했다.

Distorted 데이터에서 패딩 또는 공백 구간은 전부 음성보다 세기가 현저히 작은 백색 소음으로 대체하였다. 백색 소음 구간의 향상 목표는 무음이기 때문에 모델이 작은 세기의 백색 소음을 입력받았을 때 소리를 증폭하지 않고 오히려 제거하도록 학습한다.

2) 종단형 자동 이득 제어 모델 설계

자동 이득 제어 모델은 프레임별로 하나의 스칼라를 출력하는 모델로 입출력 구조가 매우 간단하다. 따라서 복잡한 모델 구조가 오히려 과적합 또는 불필요한 연산 증가를 불러올 것으로 판단하여 최대한 경량화함과 동시에 시계열적으로 순차 처리가 가능한 모델 구조인 GRU[3]를 선택했다. 입력으로는 STFT(Short Time Fourier Transform)을 통한 음성 스펙트로그램의 세기 값을 이용했고, 프레임 단위로 실시간 처리할 수 있도록 구현했다. GRU+MLP 구조로 이루어진 주 모델은 프레임 단위로 이득을 계산하고, 계산한 이득을 입력 신호에 곱해서 최종 출력을 얻는다. 얻은 최종 출력을 목적 신호와 대조하여 손실 함수를 계산하는데, 이때 출력과 목적 신호 각각에 로그 함수를 취한 후 평균 제곱 오차로 손실을 결정한다. 로그 함수를 취하는 이유는 신호 세기의 절대적인 차이를 기준으로 학습하는 경우 모델이 신호 증폭보다 감쇄에 더 비중을 두는 경향이 생기기 때문에, 절대적인 차이보다 그 비율에 기준을 두도록 하기 위함이다.

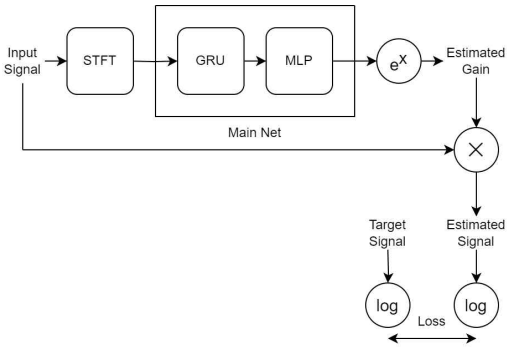

그림 1. 종단형 자동 이득 제어 모델 구조

3) 실험 및 결과

16kHz 한국어 음성 데이터베이스 SITEC DB[4]를 활용하여 학습하였다. 총 400명의 화자로부터 얻은 8,000개의 음성 표본 중 360명 화자의 7,200개 음성 표본을 train data로, 40명 화자의 800개 음성 표본을 test data로 사용하였다. I형 데이터의 음성 세기 변형 구간은 1.5~3초에서 시작하여 1~2초간 이어지도록 했고, II형 데이터의 두 음성 길이가 각각 3~4초 사이로 공백 길이는 2~4초가 되도록 설정했다. 각 길이는 균등 분포로 결정하도록 했다.

I형 데이터와 II형 데이터 모두에서 음성 세기의 변형 정도는 $4^X, X \sim N(0,1)$로 설정하여, 표본마다 다른 값으로 변형했다.

Distorted 데이터에서 공백 구간을 대체하는 백색 소음은 표준 가우시안 백색 소음에 1e-4를 곱하여 설정했다. I형 데이터와 II형 데이터 제작 예시는 표 1과 같다.

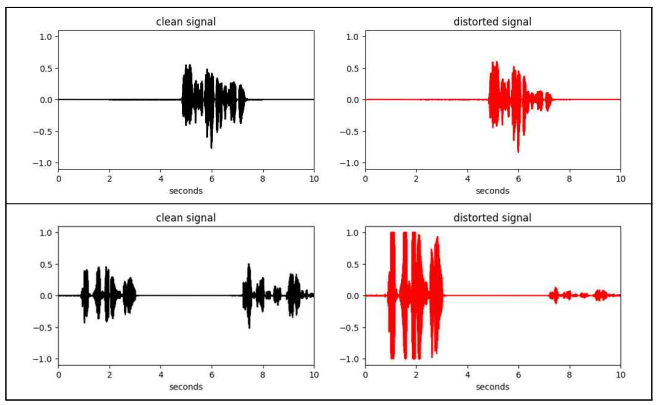

표 1. I형 데이터(상), II형 데이터(하) 예시. 각 clean signal은 $P_{\max}$ 값을 통해 적절한 세기로 조절되었다. 예시의 I형 distorted signal은 2.84초 구간부터 1.44초간 신호가 2.24배 증폭되었고, II형 distorted signal은 공백을 기준으로 앞 신호와 뒤 신호가 각각 3.27배, 0.25배 증폭되었다.

위와 같이 데이터를 구성하여 모델을 학습했다. STFT의 window length는 20ms, hop length는 10ms, frequency bin은 총 160개로 설정했다. Window는 Hann window를 사용했다. 해당 설정으로 learning rate를 1e-2로, 총 200 epoch 학습했다.

본 모델의 성능을 검증하기 위해 음성의 전반적인 품질을 측정하는 주관적 지표인 MOS test를 진행했다. MOS는 주관적인 방법을 통하여 음성 품질을 평가하는 것으로, 사람이 직접 음성을 청취해 1~5점 사이로 점수를 측정한다.

본 모델로 향상한 음성 표본 30개와 기존 이득 조절(AGC) 모델로 향상한 음성 표본 30개를 각각 추출하여 10명의 전문가가 음성 품질을 평가하였다.

	기존 AGC 모델	본 모델
MOS	3.297	3.537

표 2. MOS 측정 결과

표 2의 결과로 본 모델이 기존 AGC 모델보다 이득 제어 성능이 향상되었음을 보았다. 특히 기존 AGC 모델은 신호의 크기에만 기반해 이득을 조절하므로 크기가 작은 백색 소음을 지나치게 증폭하여 잡음을 생성하는 경향이 있었지만, 본 모델은 백색 소음 구간을 증폭하지 않고 음성 신호만을 잘 증폭하였다. 아울러 본 모델은 기존 모델과는 달리 이동평균 개념을 사용하지 않기 때문에 급격한 신호 세기 변화에도 더욱 강인한 모습을 보였다.

Ⅲ. 결론

본 논문에서는 경량화된 딥러닝 모델로 기존의 신호 세기에 기반한 자동 이득 제어 알고리즘을 대체할 수 있음을 확인하였다. 기존 모델과 비교해 신호 세기 변화에 더 빠르게 대응하는 등 성능 또한 증가함을 볼 수 있었다. 딥러닝 모델 특성상 전통적인 향상 모델보다 inductive bias가 작아 이어지는 프레임들 사이에서 이득 추정값이 크게 변하는 경우가 종종 생기는데, 해당 문제를 조사하여 손실 함수를 보완하거나 모델 구조를 변경하는 등의 방법으로 개선할 여지가 있다.

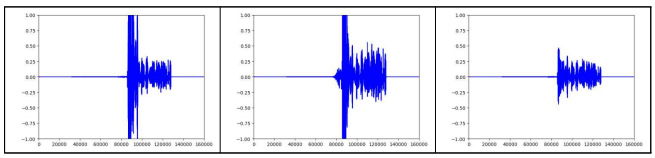

표 3. 자동 이득 제어를 적용하기 전 신호(좌), 기존 자동 이득 제어 모델로 향상한 신호(중앙), 본 모델로 향상한 신호(우)

ACKNOWLEDGMENT

이 논문은 2023년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Pérez, Juan Pablo Alegre, Santiago Celma Pueyo, and Belén Calvo López. Automatic gain control. Springer Fachmedien, 2011.

[2] Yang, Jun, et al. "Deep learning based automatic volume control and limiter system." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[3] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[4] 김봉완, et al. "SiTEC 의 공동 이용을 위한 음성 코퍼스 구축 현황 및 계획." (2003).