Perceiver 모델을 이용한 사용자 음성 구간 축약

최연웅[1], 이재준[1], 한현택[1], 이해연[1]
[1]금오공과대학교 컴퓨터소프트웨어공학과
kqwe1859@gmail.com, owns1484@naver.com, taek4549@gmail.com,
haeyeoun.lee@kumoh.ac.kr

Voice Segment Reduction using Perceiver Model

Yeon-Ung Choi[1], Jae-Jun Lee[1], Hyeon-Taek Han[1], Hae-Yeoun Lee[1]
[1]Dept. of Computer Software Engineering, Kumoh National Institute of Technology, Korea

요 약

최근 스마트 기기에서 오디오 데이터를 이용하는 응용 기술들이 증가하면서, 오디오 데이터에서 관심 있는 구간을 찾아내는 기술의 필요성이 증가하고 있다. 본 논문에서는 Perceiver 모델을 활용하여 오디오 데이터에서 사람의 음성 구간을 검출하고 축약하는 방법을 제안한다. Perceiver 모델은 복잡한 입력 데이터에 대하여 Self-attention을 기반으로 특징을 추출하면서 이전의 특징을 다음 입력으로 다시 학습하는 특징을 갖고 있어서 연속적인 데이터인 오디오에 효율적으로 적용할 수 있다. 외부 및 자체에서 수집한 음성과 비음성 데이터셋에 대하여 실험을 진행하였고, 10초 단위 세그먼트에서 대해서 92.4%의 검출 정확도를 달성하였다.

1. 서론

오디오 데이터는 기존의 단순히 녹음하여 보관하는 단계를 넘어서, 최근에는 다양한 스마트 기기의 보급으로 일상 생활에서 다양한 응용이 가능한 중요한 데이터로 사용되고 있다.

방대한 오디오 데이터에서 특정 오디오 구간을 검출하는 필수적으로 요구되며, 사람의 목소리가 존재하는 사용자 음성 구간을 의미있는 데이터로 설정하고 해당 구간을 검출하는 것은 다양한 환경에서 유용하게 사용할 수 있다.

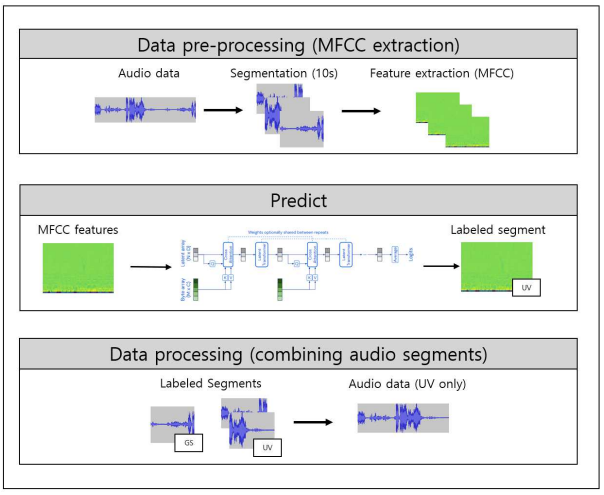
기존에는 사용자 음성 구간 검출을 위하여 통계적으로 해석한 휴리스틱 모델에 따라 사람의 목소리가 존재하는 구간을 판별하였지만, 최근엔 다양한 머신러닝 기술의 발전으로 인해 오디오 처리에서도 딥러닝 모델이 도입되어 사용되고 있다.

본 논문에서는 딥러닝 기술 중에서 Perceiver 모델을 사용하여 사용자 음성 구간을 검출하여 축약하는 방법을 제안한다. Perceiver 모델은 Transformer 모델에서 입력 데이터의 종류에 따라서 연산과 메모리를 과도하게 요구하는 문제를 해결하고자 만들어진 것으로, Cross-attention과 Self-attention 블록을 거치면서 이전 학습 결과를 다시 입력으로 받아 학습하는 특징을 가지고 있다.

따라서, Perceiver 모델은 연속적인 특징을 갖는 데이터들에 대하여 범용적으로 사용되는 CNN 모델에 기반하는 검출 방법보다 더 적합하다고 볼 수 있다[1, 2]. 음성 축약 시스템에서 처리하는 오디오 데이터가 시계열 데이터이므로 Perceiver 모델의 이전 결과를 반영하는 특징이 예측 단계에서 정확도 향상에 도움이 될 것으로 판단하여 사용하였다.

2. Perceiver 모델 기반의 사용자 음성 구간 축약

본 논문에서 제안하는 사용자 음성 구간 축약 방법의 구조는 (그림 1)과 같다.



(그림 1) 제안하는 시스템의 작동 구조도

데이터 전처리 단계에서 오디오 데이터를 입력으로 받으면, 10초 단위의 세그먼트로 입력을 구분한 후에 특징 분석을 위해 오디오 신호를 고속 푸리에 변환과 Mel Filter Bank를 사용한 Mel Spectrogram을 계산한다. 그 후에 MFCC 특징을 추출하고 Perceiver 모델의 입력으로 사용한다.

예측 단계에서 Perceiver 모델을 통해 해당 세그먼트를 비음성 오디오(General Sound, GS)와 음성이 있는 오디오(User Voice, UV)로 이진 분류를 수행하여 사용자 음성 구간을 검출한다. 참고로, 학습 단계에서는 예측 단계와 동일하게 사전에 이진 분류된 학습 오디오 세그먼트를 이용하여 Perceiver 모델의 학습을 수행한다.

마지막으로 축약 단계에서 예측 단계를 통하여 음성 오디오로 검출된 세그먼트들을 연결하여 축약을 수행한다.

Perceiver 모델은 Transformer를 기반으로 하는 모델로 기존 Transformer에서 오디오 혹은 이미지 등의 복잡한 입력이 이루어졌을 때 특징이 과도하게 추출되는 것을 억제한 방법이다. 또한, 이전 단계의 특징 영향을 받는 입력을 통하여 학습을 진행하기 때문에 시계열 연속적인 특징을 갖는 오디오 데이터에 대한 학습을 효과적으로 진행할 수 있다.

Perceiver 모델의 구성 요소 중 Cross-Attention 모듈에서 고차원인 입력 데이터를 고정된 크기의 잠재 공간에서 병목을 통과하도록 하여 입력을 제한하는 방식으로 사전 처리하고 이후에 Transformer와 유사한 Self-attention 블록을 통해 학습하여 기존 Transformer 모델에서 발생하는 2차 스케일링 문제를 방지하며 해당 모델의 블록들을 중첩 활용하여 깊은 모델을 구성한다.

또한 이 과정에서 입력 데이터의 공간 또는 시간 정보의 소실로 인하여 순차적 맥락이 사라지는 현상을 방지하기 위해 각각 입력에서 순차적 정보와 학습 양상을 합하여 학습을 진행한다. 결과적으로 기존 Transformer 모델의 Attention 기반 학습을 복잡한 입력 데이터에 대해서도 수행할 수 있다.

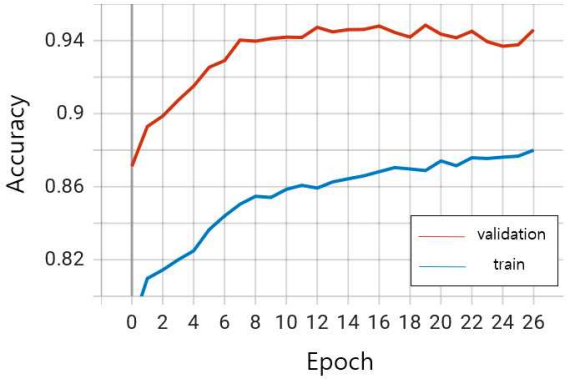
3. 실험 결과

제안하는 방법의 성능을 실험하기 위하여 Google Audio Set, Magic Hub Korean Conversational Speech Corpus, OpenSLR Seoul Corpus, AI Hub 한국인 대화 음성 등 공개된 데이터셋과 자체적으로 녹음하여 수집한 오디오에 대하여 비음성 오디오(GS)와 사람의 음성이 포함된 오디오(UV)로 구분하고 10초 단위로 나누어 세그먼트 데이터로 사용하였다. 모델의 학습에는 GS와 UV에서 각각 56,000개의 오디오 데이터를 사용하였고, 검증 과정에서는 각각 2,200개의 오디오 데이터를 사용하였다.

제안하는 방법은 Tensorflow API를 기반으로 개발하였고, 오디오 분류 정확도 향상을 위해 사전에 Perceiver 모델의 하이퍼 파라미터들에 대한 조정과 최적화를 하였다.

입력된 10초 단위 세그먼트로 구분된 오디오 데이터는 학습을 진행하기 전에 Z-Score Normalization을 적용하여 정규화하였다.

(그림 2)는 사용자 음성 검출을 위한 Perceiver 모델의 학습 과정에서의 정확도 추이를 표시한 그래프이다.



(그림 2) Perceiver의 학습 epoch 당 정확도 추세

Perceiver 모델의 학습 후에 검증 데이터를 통하여 제안하는 방법에 대한 사용자 음성 검출에 대한 정확도를 분석하였다.

검증 과정을 위한 학습에서는 각 모델에 대하여 Epoch 100, Early Stopping 7을 적용하여 학습을 진행하였으며, 정확도 비교에는 예측된 결과에 대하여 Sigmoid를 적용하고 세그먼트의 실제 분류 라벨과 모델에서 예측된 결과의 ROC 커브에서 최대의 임계값을 산출하였다.

<표 1>에는 실험을 통한 정확도와 손실값을 요약하여 나타내었다. 또한, 비교를 위하여 MobileNet 기반의 기존 연구[1] 정확도도 포함하였다. 비교군에 사용한 MobileNet 모델은 CNN 기반의 딥러닝 모델로 리소스가 제한되는 컴퓨팅 상황에서 사용을 위하여 컨볼루션 연산 단계에서 최적화를 수행한 모델이다.

<표 1> 실험 결과

	Accuracy (%)	Loss
Perceiver	92.4	0.183
MobileNet	95.8	0.128

제안하는 방법의 Perceiver 모델이 상대적으로 낮은 것은 확인하였다. 차후 연구에서 정확도 향상을 위한 모델의 재설계나 하이퍼 파라미터에 대한 조정을 수행할 예정이다.

4. 결론

본 논문에서는 Perceiver 모델을 사용하여 오디오 데이터에서 사용자 음성 구간을 검출하여 축약하는 방법을 제안하였다. 다양한 출처에서의 GS와 UV를 대상으로 음성 구간을 검출한 결과 92.4%의 정확도를 보였다.

Perceiver 모델은 순차적인 오디오에 대한 학습에 적합하지만, GS와 UV의 이진 분류 수행은 특징의 순차성에 의존하지 않아 기존 CNN 기반의 모델보다 향상된 결과가 나오지 못한 것으로 추정된다. 차후에 지속적인 연구를 수행할 계획이다.

Acknowledgement

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1F1A1057742).

참고문헌

[1] 이재준, 한현택, 최연웅, 이해연, “MobileNet을 이용한 사람 음성 구간의 오디오 축약 방법”, Proc. of KIIT Conference, 2021, pp. 523-525.

[2] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, “Perceiver: General Perception with Iterative Attention”, Proc. of the 38th International Conference on Machine Learning, 2021, pp.4651-4664.