키워드 음성인식을 위한 음성합성 기반 자동 학습 기법

임재봉, 이종수, 조용훈, 백윤주
부산대학교 컴퓨터공학과
jaebonglim@pusan.ac.kr, hl2daujs@pusan.ac.kr, kchoyh95@pusan.ac.kr,
yunju@pusan.ac.kr

A Automated Method for Training Keyword Spotter based on Speech Synthesis

Jaebong Lim, Jongsoo Lee, Yonghun Cho, Yunju Baek
School of Computer Science and Engineering, Pusan National University

요 약

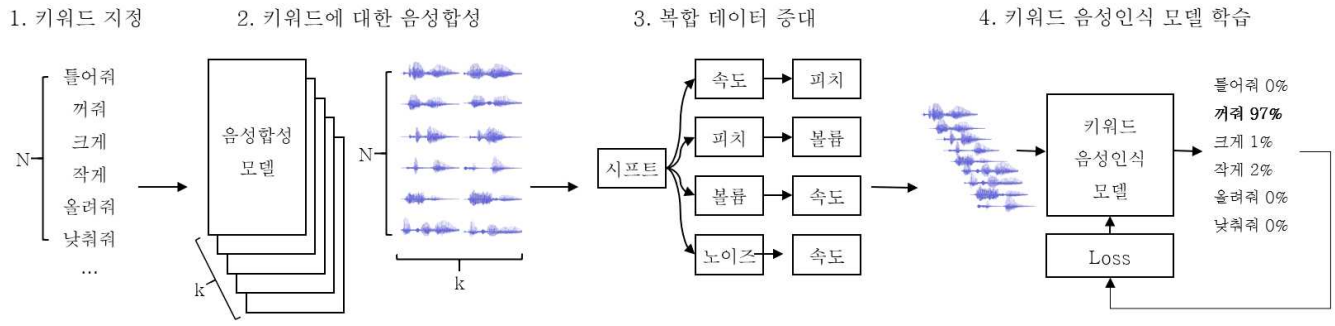
최근 경량 딥러닝 기반 키워드 음성인식은 가전, 완구, 키오스크 등 다양한 응용에 음성 인터페이스를 쉽게 적용할 수 있는 기술로서 주목받고 있다. 키워드 음성인식은 일부 키워드만 인식 가능한 음성인식 기술로서 저성능 디바이스에서 활용 가능한 장점이 있다. 그러나 응용에 따라 필요한 키워드에 대하여 다시 음성데이터를 수집해야하고 이를 학습하여 모델을 새로 준비해야하는 단점이 있다. 따라서 본 연구에서는 음성데이터 수집 없이 음성합성을 통해 생성한 음성으로만 키워드 음성인식 모델을 학습하는 음성합성 기반 자동 학습 기법을 제안하였다. 생성한 음성데이터를 활용하고자 하는 시도가 활발히 이루어지고 있으나, 기존 연구에서는 정확도를 유지하기 위하여 수집한 실제 음성데이터가 필요한 한계가 있다. 제안한 자동 학습 기법은 생성한 음성데이터에 대해 복합 데이터 증대 기법을 적용하여 실제 음성데이터 없이 키워드 음성인식의 정확도를 높였다. 제안한 기법에 대하여 상용 음성합성 서비스를 기반으로 수집한 한국어 키워드 데이터세트를 활용하여 성능평가를 진행하였다. 20개 한국어 키워드에 대해 실험한 결과, 제안한 기법을 적용하여 학습시킨 키워드 음성인식 모델의 정확도는 86.44%임을 확인하였다.

1. 서론

딥러닝 기반 음성인식 기술의 발전으로 음성 인터페이스는 스마트 스피커, 스마트 가전, 차량 내비게이션 등 다양한 응용에서 널리 활용되고 있다. 딥러닝 기반 음성인식 기술은 음성신호를 음소 또는 문자와 같은 단위 텍스트로 변환하는 음향 모델과 단위 텍스트를 언어 정보와 결합하여 문장으로 변환하는 언어 모델로 구성된다. 음향 모델과 언어 모델의 결합을 통한 음성인식은 단어나 문장 구조와 무관하게 높은 정확도를 보이는 장점이 있다. 그러나 모델의 큰 용량과 복잡성으로 인하여 서버 또는 고성능 디바이스가 필요한 단점이 있다. 따라서 음성인식 서버와의 통신을 통한 클라우드 컴퓨팅을 활용하는 방법과 고성능 디바이스 기반의 온디바이스 음성인식을 활용하는 방법은 비용이 큰 한계가 있다.

이러한 한계를 극복하기 위하여 저성능 디바이스에서 활용할 수 있는 경량 딥러닝 기반 음성인식 기술에 대한 연구가 활발하다[1]. 경량 딥러닝 기반 키워드 음성인식 기술은 음성신호 중 학습한 키워드만을 인식하는 기술로 모델의 크기와 복잡성이 낮아 성능이 제약적인 디바이스에서 활용 가능하다. 이러한 특성으로 가전, 완구, 키오스크 등 간단한 키워드 인식이 필요한 응용에서 널리 활용 가능하다. 예를 들어 가전제품을 위한 음성 인터페이스는 “켜줘”, “꺼줘”를 학습한 키워드 음성인식 모델을 활용할 수 있다. 그리고 완구는 “안녕”, “놀아줘” 등의 키워드를 학습하여 활용할 수 있다. 그러나 응용에 따라 필요한 키워드에 대하여 다시 음성데이터를 수집해야하고 이를 학습하여 모델을 새로 준비해야하는 단점이 있다.

이러한 단점을 극복하기 위해 최근 음성합성 모델을 활용하여 음성데이터를 생성하고 이를 학습에 활용하고자 하는 시도가 이루어지고 있다[3-5]. 음성합성 모델은 딥러닝 모델을 활용하여 다양한 목소리의 음성을 자연스럽게 생성할 수 있다. 그러나 생성한 음성데이터만을 학습 시 수집한 실제 음성데이터를 학습하였을 때보다 정확도가 떨어지는 한계가 있다.



(그림 1) 키워드 음성인식을 위한 자동 학습 기법의 다이어그램

기존 키워드 음성인식 연구에서는 생성한 음성데이터를 학습하여 필요한 수집데이터를 최소화하였다[2]. Lin의 연구에서는 생성한 음성으로 학습 시 정확도가 떨어지는 문제를 개선하기 위하여, 먼저 수집한 음성데이터를 활용하여 다양한 키워드에 대해 Multi-task Pretraining을 적용하였다[3]. 따라서 기존 연구에서는 수집한 실제 음성데이터가 일부 필요한 한계점이 있다.

본 연구에서는 음성데이터 수집 없이 음성합성을 통해 생성한 음성으로만 키워드 음성인식 모델을 학습하는 자동 학습 기법을 제안하였다. 높은 정확도의 생성데이터 기반 키워드 음성인식 모델을 위하여 생성한 음성데이터에 대해 복합 데이터 증대 기법을 적용하였다. 제안한 기법을 통해 인식하고자 하는 키워드를 텍스트로 지정하면 자동으로 높은 정확도의 키워드 음성인식 모델을 얻을 수 있다. 텍스트 입력에 대해 음성합성으로 학습데이터를 생성하므로 별도의 데이터 수집 및 데이터 레이블링이 불필요한 장점이 있다. 따라서 제안한 기법을 활용하여 음성인터페이스를 다양한 응용에 손쉽게 적용할 수 있다.

2. 키워드 음성인식을 위한 자동 학습 기법

본 연구에서는 상용 음성합성 모델과 복합 데이터 증대 기법을 활용하여 키워드 음성인식을 위한 자동 학습 기법을 제안하였다. 각 화자에 대해 학습된 다수의 음성합성 모델을 활용하여 지정한 키워드에 대한 음성데이터를 생성하고 다양한 음성 변환을 혼합한 복합 데이터 증대 기법을 적용하여 학습데이터를 생성하였다. 그리고 생성한 학습데이터를 활용하여 경량 딥러닝 기반의 음성인식 모델을 학습시켰다. 이렇게 자동 학습한 음성인식 모델은 높은 정확도로 키워드를 인식할 수 있음을 확인하였다.

자동 학습 기법은 그림 1과 같이 4단계로 구성하였다. 첫 번째 키워드 지정 단계는 응용에 따라 음성인식 할 키워드를 지정한다. 자동 학습된 음성인식 모델은 현 단계에서 지정한 N개의 키워드를 인식한다. 두 번째 음성합성 단계는 지정한 키워드의 음성데이터를 합성한다. 다양한 목소리를 학습한 k개의 음성데이터 합성 모델을 활용하여 지정한 키워드 텍스트에 대해 다양한 음성데이터를 생성한다. 세 번째 단계는 합성한 음성데이터에 대해 복합 데이터 증대 기법을 적용하여 생성데이터를 다량 확보한다. 네 번째 키워드 음성인식 모델 학습 단계는 생성 및 증대시킨 음성데이터를 활용하여 지정한 N개의 키워드 중 하나로 분류하여 출력하는 키워드 음성인식 모델을 학습한다.

상용 음성합성 모델을 통해 생성한 음성데이터만을 학습한 키워드 음성인식 모델은 정확도가 떨어지므로 복합 데이터 증대 기법을 활용하여 이를 개선시켰다. 복합 데이터 증대 기법은 음성데이터에 대하여 다양한 음성 변환을 적용하되 자연스러운 음성을 유지하도록 최적화시켰다. 복합 데이터 증대 기법은 시프트, 속도, 피치, 볼륨, 노이즈 삽입 총 5개의 변환을 랜덤하게 조합하도록 구성하였다. 시프트는 음성데이터의 시간 축을 변환시키고, 속도는 음성의 길이를 변환시킨다. 피치는 음의 높낮이를 변환시키고 볼륨은 음의 크기를 변환시킨다. 노이즈 삽입은 다양한 노이즈를 포함한 공개 데이터세트를 활용하여 다양한 음성데이터를 생성시켰다.

이렇게 생성한 음성데이터를 활용하여 경량 딥러닝 기반의 키워드 음성인식 모델을 학습시켰다. 키워드 음성인식 모델은 전처리 단계와 인식 단계로 구성하였다. 먼저 전처리 단계는 STFT (Short Time Fourier Transform)를 적용하여 시간, 주파수별 음성 정보를 포함하는 MFCC (Mel Frequency

Cepstrum) 피처를 획득하였다. MFCC는 음성의 중요한 정보만을 추출한 피처로서 음성인식 및 음성인식 모델 학습을 용이하도록 하기 위해 채택하였다. 그리고 인식 단계에서는 획득한 피처를 CNN (Convolutional Neural Network) 기반 모델에 입력하여 키워드 음성인식을 수행하도록 구성하였다. CNN 기반 모델은 저성능 디바이스에 탑재 가능하도록 6개 층과 50k의 파라미터를 가지는 구조로 최적화하였다. 6층의 모델 구조는 64개 채널의 3x3 Convolution 1층, 64개 채널의 3x3 Depth-Separable Convolution 4층, 20 채널의 Dense 층 1개로 구성하였다.

3. 실험 및 성능평가

실험 및 성능 평가를 위하여 생성한 음성데이터를 기반으로 학습데이터를 구성하였고, 수집한 음성데이터를 기반으로 테스트 데이터를 구성하였다. 이를 통해 제안한 기법으로 학습한 키워드 음성인식 모델이 수집한 실제 음성데이터에 대해서 잘 동작하는지 평가하였다. 학습데이터는 상용 음성학습 모델을 활용하여 구성하였다. 상용 음성합성 모델은 한 명의 화자에 대해 학습 된 Tacotron[4] 기반 모델로, 500개의 서로 다른 상용 모델을 활용하여 지정한 키워드 별로 500개 목소리의 음성데이터를 생성하였다. 40개 키워드에 대해 각각 500개의 음성데이터를 생성하였고, 이 중 절반인 20개를 키워드로 지정하였고 나머지 20개는 Negative 데이터로 지정하였다. 테스트 데이터는 11명의 실험자를 대상으로 총 4400개의 음성데이터를 수집하여 활용하였다.

구성한 데이터세트를 활용하여 성능평가를 진행하였다. 먼저 복합 데이터 증대 기법을 적용하지 않았을 때 키워드 음성인식의 정확도가 62.8%임을 확인하였다. 복합 데이터 증대 기법을 포함한 제안한 기법을 적용하였을 때 키워드 음성인식을 정확도가 86.44%로 대폭 향상됨을 확인하였다.

4. 결론

본 연구에서는 키워드 음성인식 모델을 위한 음성합성 기반 자동 학습 기법을 제안하였다. 키워드 음성인식에서 음성데이터 수집 없이 생성한 음성데이터로 높은 정확도의 모델을 얻을 수 있음을 확인하였다. 본 연구를 통해 스마트 스피커, 스마트 가전, 차량 내비게이션 등 다양한 응용에 음성 인터페이스를 용이하게 적용할 수 있을 것으로 기대된다.

Acknowledgement

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-01117, 로컬 음성인식을 위한 머신러닝 기반 초소형 모듈 개발)

참고문헌

[1] Y. Zhang, N. Suda, L. Lai and V. Chandra, “Hello Edge: Keyword Spotting on Microcontrollers,” arXiv preprint arXiv:1711.07128, Fab. 2018

[2] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu and Z. Wu, "Speech Recognition with Augmented Synthesized Speech," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Dec. 2019.

[3] J. Lin, K. Kilgour, D. Roblek and M. Sharifi, "Training Keyword Spotters with Limited and Synthesized Speech Data," 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020.

[4] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2018.