기말 프로젝트 : FMCC

인천대학교 컴퓨터공학부 신소정, 김상희

Female-Male Classification Challenge

Sojeong Shin, Sanghee Kim

요 약

본 논문에서는 다양한 잡음이 섞인 여성과 남성 음성 신호에 대해 지도 학습 방법으로 성별을 분류하는 알고리즘을 소개한다. 전체 음성 파일은 한 단어에 대한 목소리가 녹음된 신호를 활용한다. 본 연구에서는 음성 신호에 대해 스펙트로그램(spectrogram) 이미지를 생성하고 생성된 이미지를 CNN 모델에 학습하여 신호의 성별을 구분하였다. 원본 신호에 포함된 발성의 주파수 범위 밖의 영역은 별도의 필터링이 아닌 스펙트로그램 이미지의 주파수 표현 범위를 조절하여 대체하였으며 최종적으로 1000개의 데이터셋에서 98.8%의 인식률을 확인하였다.

1. 서론

산업적으로 음성 신호에 대한 서비스들이 증가하는 추세를 보인다. 대부분의 서비스들은 모바일 기기에서 접근이 가능하며 따라서 공간적 제약이 없다는 특징을 갖는다. 지하철, 인파가 밀집된 장소 등, 음성 서비스를 이용할 경우에 입력되는 음성 신호는 여러가지의 잡음들을 포함하는데, 이러한 잡음에 의해 서비스에 제약이 발생할 수 있다.

Chao Li 외 8명은 “Deep Speaker: an End -to-End Neural Speaker Embedding system”에서[2] 발화자를 구분하기 위해 CNN과 LSTM 기법을 활용하였으며 성별 분류에서도 높은 성능을 나타낸 바 있다. 해당 논문에서 제안한 모델은 컨볼루션 층 5개, LSTM 층 신경망 층 12개로 이루어져 있다.

본 논문은 음성 신호의 잡음 제거 성능을 향상시킬 수 있도록 신호의 성별을 인식하는 기법을 제안한다. 제안된 알고리즘은 음성 신호의 스펙트로그램 이미지를 활용한 CNN 모델을 사용한다. 학습 모델의 성능을 증강시키기 위해 음성 신호를 3kHz로 샘플링하여 사람 목소리에 해당하지 않는 주파수 대역을 제거하였고 학습 모델의 사이즈를 축소하였다. CNN 모델은 4개의 컨볼루션 층와 1개의 신경망 층을 갖도록 설계하였다.

2. 본론

2.1. 데이터셋

본 연구에서 주어진 데이터는 다음과 같다:

(1)학습 데이터(남녀 각각 20명, 화자 별 200 발음으로 총 8,000 샘플, 5dB SNR 미만의 생활 잡음(TV잡음)에 오염된 상태

(2)테스트 데이터(남 500 샘플, 여 500 샘플로 총 1,000 샘플, 학습 데이터와 동일한 종류의 생활 잡음에 오염된 상태)

(3)음성 데이터 포맷: PCM 파일 포맷, WAV 파일에서 헤더 없는 형태. 16kHz, 16bit linear encoding, mono channel, little endian

본 연구의 실험 설계는 다음과 같다:

(4)실험환경: 본 연구에는 2.3 GHz 워드 코어 intel Core i7 프로세서와 16GB 3733 MHz가 장착된 MacBook Pro와 Python 3 Google Compute Engine 백엔드 (GPU-T4) 사용.

(5)인공신경망 모델: 총 4가지 모델을 제안하며, 그 중 DeepSpeaker CNN 모델을 최종 모델로 선택. 배치 정규화 및 드롭아웃 기법, 데이터 증강을 활용하여 모델의 일반화 성능 증가. 전처리 과정에서 주파수 대역 필터링을 통해 잡음 제거.

2.2. Optuna와 LSTM 레이어를 사용한 모델

LSTM 모델을 사용하여 음성 데이터의 시계열 특성을 학습하고, Optuna를 통해 하이퍼파라미터를 최적화하였다.

Accuracy: 86.50%

Hit: 865, Total: 1000

2.3. 스펙트로그램 DeepSpeaker CNN 모델

음성 데이터의 주파수 대역을 필터링하여

스펙트로그램 이미지를 생성하고, 이를 DeepSpeaker CNN 모델의 입력으로 사용하였다. 전처리 과정에서 TV잡음은 회절현상에 의해 고주파 성분의 소리인 경우가 있기에 STFT의 sr=16000Hz, n_fft=1024로 파라미터를 조정하여 0Hz~1600Hz까지의 주파수 성분을 포함하는 스펙트로그램으로 모델을 학습한다.

Accuracy: 96.90%

Hit: 969, Total: 1000

2.4. 모델 3: 데이터 증강 및 DeepSpeaker CNN 모델

모델 2에 데이터 증강 기법을 적용하여 Train 데이터 셋을 확장하였다. 모델 4에서는 DeepSpeaker 모델을 구현하였고, 배치 정규화, 드롭아웃을 사용하여 모델의 일반화 성능을 향상시켰다.

Accuracy: 98.80%

Hit: 988, Total: 1000

Total params: 60276993 (229.94 MB)

Trainable params: 60274049 (229.93 MB)

Non-trainable params: 2944 (11.50 KB)

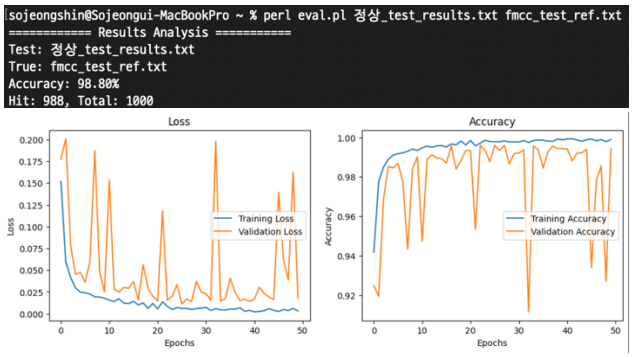

Figure 1. 전처리 포함된 스펙토그램 Deep Speaker 모델 결과, 학습곡선

3. 결 론

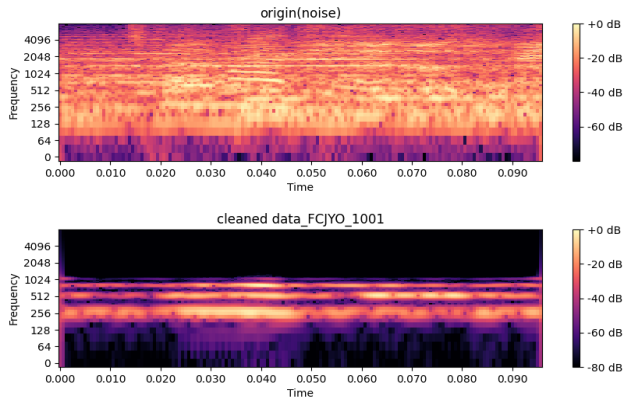

Figure 2. 가우시안 필터링 음성 신호

.raw 데이터는 오디오로 들어볼 수 없기에 노이즈의 정도와 분석하고자 하는 음성 데이터의 상태가 어떠한지 .wav 파일로 변환하여 확인하였다. 잡음의 종류가 TV잡음으로 자동차 잡음 등과는 달리 인물의 목소리가 들려 분석하고자 하는 남녀의 음성 주파수 대역과 겹치는 부분이 많아, 주파수 대역으로 무결성 데이터를 만들기는 어렵다. 하지만 사람의 음성은 기본 주파수에 정수배의 고조파가 더해져 이루어진다는 <정의1> 에 의거하여 정수배의 가우시안 필터를 정의하여 최대한 해당 인물의 목소리만 남도록 한다.

Fig.2는 ~5db 정도의 TV 잡음이 함께 들어간 FCJYO_1001.raw 원본 데이터의 스펙트로그램과 FCJYO_1001.raw 데이터의 가우시안 필터를 사용하여 정제한 스펙트로그램이다. 공통된 부분이 clean 데이터이고, 띠의 형태로 이루어져 있으며 띠의 주파수 대역은 약 156~356Hz, 500~700Hz, 1000~1200Hz 정수배의 주파수 대역으로 이루어져 정수배로 이루어져 있음을 확인할 수 있다. 따라서 잡음이 섞인 원본 데이터의 스펙트로그램을 CNN 모델을 사용하여 학습하면 가우시안 필터를 적용하여 원본 데이터를 얻어낸 방식과 동일하게 학습할 수 있다.

주파수대역으로 음성이 분리가 가능하나 모든 인물들의 목소리 별 가우시안 필터를 만들어야 분리가 되므로, 정수배의 주파수로 데이터의 음성이 이루어진다는 사실에 착안하여 16000Hz 음성 데이터를 3000Hz로 샘플링하고, DeepSpeaker CNN 모델을 이용하여 필터를 적용한 것과 같은 효과가 나도록 하였다.

모델3이 최종 FCMM 신경망 모델이며, pitch 조절로 반음크기로 데이터를 증강시켜서 학습하였다.

추후 모델의 데이터를 추가한 학습으로 성능을 증강시켜보고자 한다.

참 고 문 헌

[1]https://svantek.com/ko/academy/sound-frequency/

[2] Deep Speaker: an End -to-End Neural Speaker Embedding system - Chao Li 외 8명

[3]https://kr.mathworks.com/discovery/bandpass-filter.html

[4]코드-ChatGPT