드론으로 취득된 음성정보에서 구조요청 소리 감지 및 방향 측정 딥러닝 모델 개발

Developed a deep-learning model for sound detection and direction measurement in speech information acquired by drone

요 약

대형복합재난이 지속적으로 발생하고 있는 현재 사회에서 인명을 구조하기 위한 방법으로 드론 사용이 떠오르고 있다. 재난이나 구조요청 상황에서는 환경소음과 노이즈가 크기 때문에 구조 요청 음성이 어디서 들려오는지 그 방향을 추정하기가 어렵다. 이에 대해 본 논문에서는 소음과 노이즈로부터 음성의 발원 방향을 추정하는 딥러닝 모델을 개발한다. 구조 요청 음성 오디오와 환경 소음을 합성하여서 데이터를 생성하고 해당 데이터를 전처리 한 후, 필요한 특징을 추출하여 설계한 딥러닝 모델을 통해 해당 구조 요청의 발원 방향을 추정한다.

1. 서 론

현재 사회에서는 대형복합재난이 지속적으로 발생하고 있다. 대형복합재난은 동시 또는 순차적으로 두 가지 이상의 자연, 사회 재난이 발생하고, 그 영향이 복합화되어 인명, 재산, 기반시설 마비 등 그 피해가 극심하여 국가적 위협이 되어 범부처의 통합적 대응이 필요한 재난이다. 이러한 대형복합재난에서 위험에 처한 인명을 구조하기 위한 방법으로 드론 사용이 떠오르고 있다.

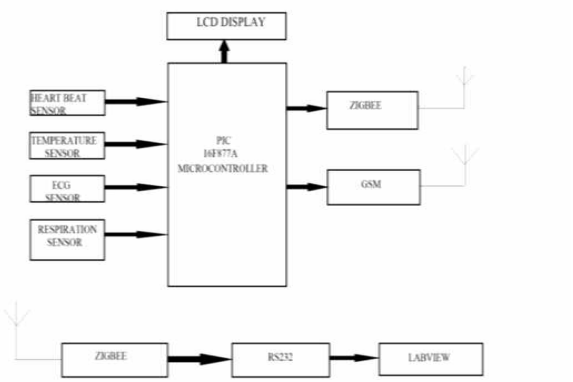
재난현장이나 구조현장에서 드론으로 습득할 수 있는 인명 구조요청 소리는 환경 소음과 노이즈가 클 확률이 높다. 이러한 환경에서 정확한 인명 구조 요청소리를 판별하고 이의 발원 방향을 측정하는 것이 해당 연구의 목적이다. 이를 위해 데이터 전처리와 딥러닝 모델을 개발한다.

본 논문에는 음성 오디오와 환경 소음을 임의로 합성하여 딥러닝 학습에 필요한 데이터 생성, 전처리 및 특징 추출, 학습 모델 구조와 성능 평가 등의 내용을 담고 있다.

2. 기존연구

2.1 기존 연구 1

Quadcopter based technology for an emergency healthcare : 교통 체증이나 도시의 정체로 인해 종종 구급차가 응급 상황 현장에 늦게 도착하게 된다. 이런 경우를 방지하고자 드론을 사용하여 신속하게 현장에 도착할 수 있게 하고, 환자의 상황을 체크 할 수 있는 것을 목표로 한다. 긴급 번호로 연락이 온다면 위치를 추적하고, GPS를 사용하여서 드론을 현장에 보낸다. 무인 항공기는 환자 몸에 편리하게 부착될 수 있는 변형 센서로 구성된 환자 실시간 모니터링 시스템을 갖추고 있다. 이를 이용해 환자의 상태를 체크하고 인근 병원과 구급대원에게 알린다.

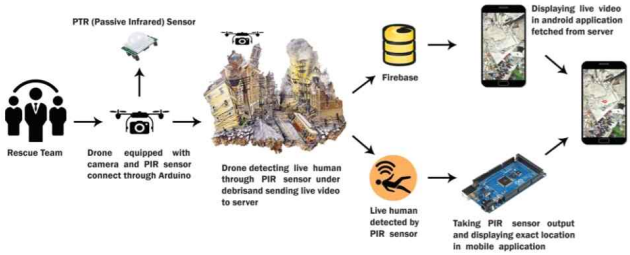


다음 그림과 같이 heartbeat sensor, temperature sensor, ecg sensor, respiration sensor로 환자의 건강상태를 체크할 수 있고 이를 ZIGBEE, GSM을 사용하여 인근 병원과 통신하는 방식을 채용한다. 이 연구를 통해 구급대원이 현장에 도착하는 시간이 지체될 때 드론을 사용해 먼저 환자의 상태를 보고, 이에 대응할 수 있어 인명구조에 도움을 줄 수 있을 것이다.

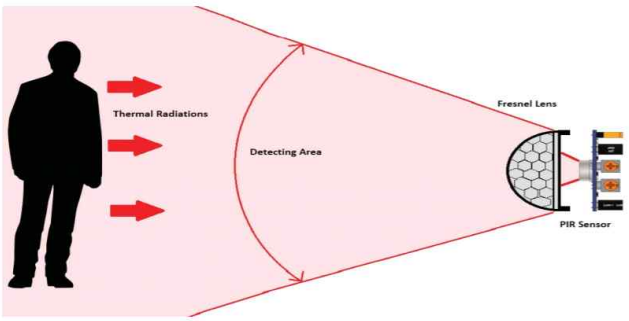
위와 같이 센서가 부착된 드론을 사고 현장에 보내 신속하게 환자의 상태를 검진 할 수 있다.

2.2 기존 연구 2

사람을 탐지할 수 있는 실시간 자율 무인 기술 시스템인 'DronAID'에 관한 연구로서, 가장 빠른 시점에 생존자의 위치를 정확히 파악하고 구조할 수 있게 하는 것을 목표로 한다. 무인 항공기 기반 시스템이기 때문에 쉽게 동원과 제어가 가능할 뿐더러, 카메라 모듈 및 센서장치가 포함되어 있어 잔해 아래에 묻혀있는 인간의 존재 역시 확인이 가능하다.



위와 같은 방식으로 작동되며, 발견한 사람의 위치를 구조대에게 전송해 신속하고 효율적으로 구조가 이루어질 수 있도록 한다. 또한 실시간으로 비디오를 서버에 전송해 희생자가 있을 것으로 예상되는 위치의 영상을 모바일로 보여준다.



PIR 센서를 탑재해 해당 센서가 탐지가 가능한 반경 안에 사람이 있는 경우, 사람이 방출하는 방사선을 감지하고, 그 위치를 파악할 수가 있다. 드론의 시야를 통해 사람을 탐색하는 것이 아닌 방사선을 감지하는 방식이기 때문에 건물 아래에 사람이 깔렸거나 사람이 외부에서 보이지 않는 경우에도 탐지가 가능하다.

2.3 기존 연구의 문제점

첫 번째 기존 연구 같은 경우, 긴급번호로 연락이 온 후에 드론이 출동하기 때문에, 만일 희생자나 피해자가 연락이 불가능한 상황인 경우는 출동이 불가능하다. 자연재해 상황 같은 경우 구조를 바라는 피해자가 연락이 가능한 상황일 가능성, 또한 연락을 할 수 있는 기기가 제대로 작동이 가능한 상황일 가능성이 크지 않으므로 매우 큰 문제점이라고 볼 수 있다. 또한 드론에 따로 자체적인 카메라가 부착되어있지 않아 부상자의 정보를 오로지 센서를 통해서만 파악해야 한다.

두 번째 기존 연구의 경우는 첫 번째 기존 연구에서는 포함되지 않았던 희생자 탐지 기능을 가지고 있으나, 희생자를 탐지할 때 사용하는 PIR 센서는 일정한 적외선을 가진 물체가 움직이는 것을 감지하는 것이므로 사람이 움직임이 없는 경우는 감지가 불가능하다. 또한 미세한 정도의 움직임 역시 감지를 하지 못하기 때문에 자연재해 현장에서 희생자가 움직이지 못하는 경우에는 성능을 기대하기가 어렵다. 더 나아가 센서 자체가 급격한 온도변화와 같은 주변환경의 영향을 받기 때문에 야외에서 사용할 때는 정확한 탐지가 어렵다.

3. 프로세스

3.1 전체 프로세스

본 연구는 <그림 1> 과 같이 데이터 확대, 데이터 전처리와 특징 추출, CNN 모델 학습, 결과 출력의 총 네 가지 과정으로 이루어져있다.

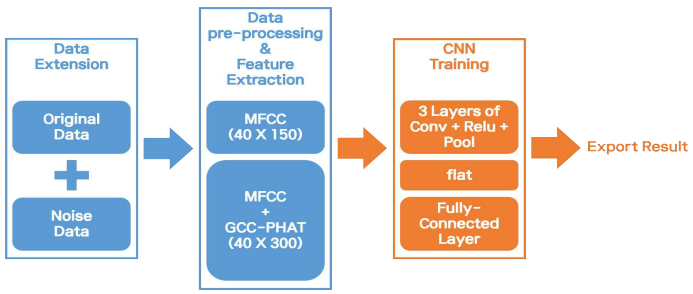


그림 5. 전체 프로세스

3.2 학습 데이터 생성 [1]

LINE ENGINEERING의 Kunihiko Sato 개발자가 올린 파이썬 코드아 쉘 스크립트를 이용하여 스테레오 타입을 가진 음성 오디오와 스테레오 타입의 환경 소음 오디오를 합성하여 학습 데이터를 생성했다. 음성 오디오와 환경 오디오는 모두 wav파일이며, 이렇게 만들어진 학습 데이터의 SNR(Signal-to-Noise ratio)는 모두 -20으로 고정하였다.

$$SNR_{db} = 20\log_{10}\frac{A_{signal}}{A_{noise}}$$

수식 1. SNR 계산 방법

A_{signal}과 A_{noise}는 각각 음성과 잡음의 '크기' 혹은 '세기'를 나타낸다. '세기'의 정의에는 몇 가지가 있는데, 본 논문에서는 진폭값의 평균 제곱근(Root Mean Square, RMS)을 각 소리의 세기로 정의한다.

음성 오디오와 환경 소음 오디오, 두 데이터 모두 양자화 bit수와 샘플링 레이트를 16bit와 44.1kHz로 통일하였다. 이 후, 음성 오디오 파일과 환경 소음 오디오 파일을 읽어들여 byte 객체로 반환하여 오디오 파일 음성 파형의 진폭값을 취득한다. 취득한 진폭값의 편균 제곱근(RMS)과 SNR을 이용하여 임의의 크기로 파형을 합성한다.

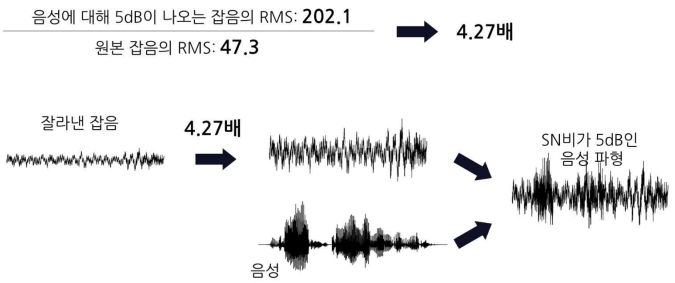


그림 6. 파형 합성 예시

두 진폭을 더한 진폭 값이 양자화 bit수인 16bit의 최대값을 넘을 수 있으므로 정규화 시켜준다.

위 과정으로 각 음성 오디오 데이터와 환경 소음 데이터에 적용시켜 데이터를 생성할 시, 음성 데이터의 수와 소음 데이터의 수를 곱한 만큼의 데이터를 생성할 수 있다.

3.2 데이터 전처리 및 특징 추출

각 채널의 차이점을 통해서 모델을 학습시켜야 하므로 채널 분리가 필수적인 전처리 요소이다. 학습과 테스트에 사용되는 데이터가 스테레오 타입의 wav파일로 총 2개의 채널을 가지고 있기 때문에 이를 각각 다른 두 개의 wav 파일로 분리하였다.

위의 과정을 통해 두 개로 나누어진 데이터는 MFCC(Mel-Frequency Cepstral Coefficient)[2], GCC-PHAT(Generalized Cross Correlation-Phase Transform)와 같은 두 개의 특징 추출 단계를 거친다.

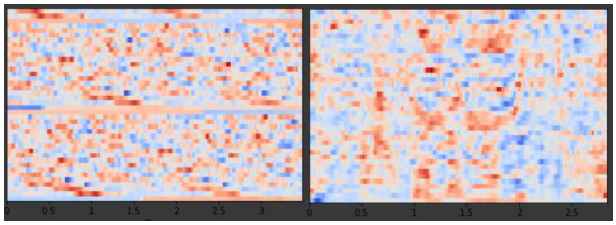


그림 7. 두 채널에서 각각 추출된 MFCC의 스펙트로그램

MFCC는 음정의 높낮이에 영향을 받지 않기 때문에 음성처리에서의 특징으로 대표적이다. 본 논문에서는 MFCC의 개수를 40개로 고정하여 추출하였으며 한 개의 데이터 당 40 X 150의 일정한 크기를 가진 벡터로 나타내었다. 또한 두 채널에서 각각 추출되는 MFCC가 다르므로, 두 벡터의 차를 특징 벡터로 사용하였다.

GCC-PHAT는 음원이 서로 다른 두 마이크로 도달하는 시간차이(Time Difference of Arrival, TDOA)를 구하는 기법 중 하나[3][4]이다.

$$\hat{G}_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{\left|X_i(f)[X_j(f)]^*\right|}$$

수식 2. GCC-PHAT의 정의

$$\hat{d}_{PHAT}(i,j) = \arg\max_d(\hat{R}_{PHAT}(d))$$

수식 3. TDOA 추정

두 입력 신호 X_i와 X_j에 대해 GCC-PHAT를 구하는 방법은 수식 2와 같다. 이렇게 구해진 GCC-PHAT는 수식 3을 거쳐 최종적으로 음원 도달시간 차이를 추정할 수 있다. 본 논문에서는 이렇게 추출된 GCC-PHAT의 스펙트럼 성분에 IFFT(Inverse Fast Fourier Transform)의 연산을 수행한 벡터를 MFCC와 동일하게 40 X 150의 크기를 가진 벡터로 나타내어 먼저 추출한 MFCC와 함께 사용하였다.

또한 더욱 원활한 학습을 위해 모델에 입력으로 사용되는 데이터는 모두 정규화를 거친 후 학습에 사용되었다.

3.3 모델 구조

본 논문에서 제시하는 모델은 CNN(Convolutional Neural Network) 아키텍처를 사용하여 설계되었다. CNN 모델이란 합성곱 연산을 사용한 신경만 네트워크 모델로, 주로 이미지를 인식하는 데에 쓰이고 있지만 본 연구에서는 MFCC처리한 음성파일에서 특징을 뽑아내어 전

처리하였기 때문에 이를 인식할 수 있는 CNN 모델을 사용하였다. 사용된 CNN 아키텍처는 컨볼루션 계층, 풀링 계층, 활성화 함수 등으로 구성되어있다.

$$C(x,y) = \sum_{a=0}^{k-1}\sum_{b=0}^{k-1} I(x-a, y-b)F(a,b)$$

수식 4. convolution 연산

수식 4는 컨볼루션 계층에서 합성곱(Convolution) 연산을 나타내는 식이다. 다음 식에서 I는 2차원 입력배열, F는 필터, k는 필터의 크기를 나타낸다.

풀링 계층(Pooling Layer)는 컨볼루션 계층에서 생성된 결과로 중복된 특징 값이나, 상호 연관성이 높은 요소들을 제거하고 특징을 강화하기 위해 사용된다. Strides와 Kernel Size, Padding 여부에 따라 데이터의 크기가 바뀌게 된다.

$$f(x) = \max(0, x)$$

수식 5. ReLu 함수

활성화 함수로는 은닉층에서 ReLu를 사용하였다. 기존에 사용되었던 Sigmoid 함수 혹은 tanh함수 보다 CNN에서 데이터를 학습시키고 출력값을 수렴하는 데에 ReLu가 더 높은 성능을 보이기 때문이다.

본 연구에서는 컨볼루션 계층 3개와 풀링 계층을 쌓았고 그 다음 ReLu를 거쳐 해당 결과 값을 Fully-Connected Layer로 연결하여 이를 0°, 60°, 120°, 180° 4개의 방향을 가진 클래스로 구분하였다.

4. 결과 및 향후 연구 방향

4.1 결과

설계된 모델에 전처리 되어진 데이터를 입력값으로 넣고 이를 학습시켜 4개의 클래스로 구분하게 된다. 사용된 데이터는 모두 8948개이며, Train:Validation:Test는 각각 6:2:2의 비율로 나누어 사용하였다.

	MFCC	MFCC+GCC
Accuracy	70%	76%

표 1. 모델 학습 결과

표 1은 MFCC로 전처리를 진행한 데이터와 MFCC와 GCC기법을 사용하여 전처리를 진행한 데이터를 입력값으로 모델에 주어 학습을 진행한 결과이다. 실험 결과, MFCC 하나의 특징값을 사용하였을 때 보다, GCC까지 특징값으로 사용하였을 때 모델이 더욱 더 신뢰성 있는 정보를 제공함을 볼 수 있다.

3.2 향후 연구 방향

이 후에는 모델의 정확도를 높이기 위해 모델의 hyperparameter와 레이어의 수, 입력 벡터의 크기 등을 바꾸어 가며 모델을 새롭게 설계할 것이다. 뿐만 아니라 노이즈 캔슬링을 위한 새로운 모델을 설계하여 더욱 확실하게 환경 소음이 제거된 음성데이터를 통해 모델이 지금보다 정확하게 발원방향을 추정할 수 있도록 하는 방향으로 연구를 진행할 것이다.

본 연구에서 진행한 모델을 이용한다면, 재난 상황에 드론을 이용하여 구조 요청이 어느 방향에서 생성되었는지 빠르게 파악하여 인명 구조에 도움이 될 수 있을 거라 기대한다.

4. 참고문헌

[1] “*딥 러닝 음성 인식에 필요한 훈련 데이터를 직접 만들어보자*”, LINE Engineering 블로그, 2018년10월01일 수정, 2019년4월14일 접속, https://engineering.linecorp.com/ko/blog/voice-waveform-arbitrary-signal-to-noise-ratio-python/

[2] Sahidullah, Md and Saha Goutam. *Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition*, 2012.

[3] 권오현, 장준혁. *심화 신경망을 사용한 다채널 마이크 구조에서의 음원방향추정*, 한국통신학회 하계종합학술발표회, 2017.

[4] Knapp, C. H. and Carter, G. C. *The generalized correlation method for estimation of time delay*, IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-24, 1976.