|  |
| --- |
| **스테레오 타입으로 구성된 음성데이터의 성별, 나이 및 발원 방향 추정 딥러닝 모델 개발** |
|  |
| Development of a deep-learning model for estimation of gender age and origin of audio data composed of stereo type |
|  |
| **요 약**  현대 사회에서는 대형복합재난이 지속적으로 발생하고 있지만, 모든 재난 현장이 인력 투입이 가능한 장소라고확신하기는 어렵다. 그런 문제점을 해결하기 위하여 본 연구에서는 스테레오 타입으로 구성되어 있는 음성 데이터를 적합한 필터들을 사용하여 노이즈를 제거한 뒤, 실제 사람의 음성 데이터만을 추출해낸다. 그 후 해당 음성의 남녀노소를 판별한 뒤 어느 방향으로부터 발원되어 왔는지까지를 추정하는 딥러닝 모델을 개발한다. | | |

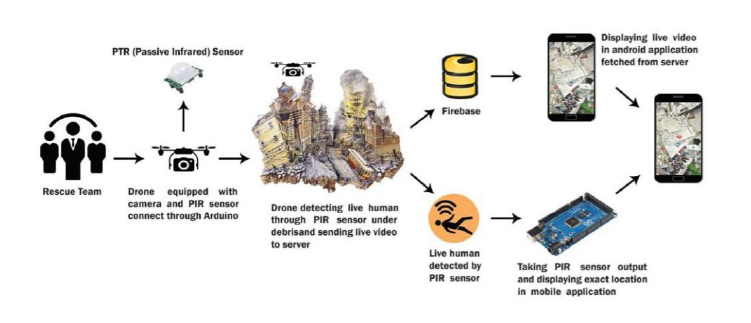
1. **서 론**
   1. 연구 배경

현대 사회에서는 대형복합재난이 지속적으로 발생하고 있다. 대형 복합재난은 동시 또는 순차적으로 두 가지 이상의 자연, 또는 사회적인 재난이 발생하는 것이며 그 영향이 복합화되어 피해가 극심하기 때문에 국가적 위협으로 분류하여 범부처의 통합적 대응이 필요한 재난이다. 이러한 대형 복합 재난의 모든 현장이 사람이 접근 가능한 현장이라고 확신하기는 어렵다. 본 연구는 그러한 상황에서 더욱 더 신속한 인명 구조가 이루어질 수 있도록 노이즈가 극심한 음성데이터에서 사람의 음성만을 추출하여 해당 발원지를 추정하고자 한다.

* 1. 연구 목표

연구에 사용되는 데이터는 드론에서 취득된 스테레오 음성으로 사람의 구조요청 소리와 극심한 노이즈 소리로 이루어져있다. 본 연구의 목표는 해당 음성 데이터로부터 필요한 사람의 구조 요청 소리만을 분리해내고, 해당 소리를 낸 사람을 남녀노소 4개의 클래스로 구분한다. 또한 해당 소리가 어디로부터 발생되었는지 그 방향을 20도 간격으로 추정하는 것을 목표로 한다.

1. **기존 연구**
   1. **DronAID : A Smart Human Detection Drone for Rescue [1]**

해당 논문은 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 센서 장치가 포함되어 있어 잔해 아래에 묻혀있는 인간의 존재 역시 확인이 가능하다. 

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.

희생자를 탐지할 때 사용하는 PIR 센서는 일정한 적외선을 가진 물체가 움직이는 것을 감지하는 것이므로, 사람의 움직임이 없는 경우는 감지가 불가능하다는 단점이 있다. 또한 미세한 정도의 움직임 역시 감지를 하지 못하기 때문에 희생자가 움직이기가 여의치 않은 경우에는 성능을 기대하기가 어렵다. 더 나아가 센서 자체가 급격한 온도 변화와 같은 주변 환경의 영향을 민감하게 받아들이기 때문에 야외에서 사용할 때는 정확한 탐지가 어렵다.

* 1. **드론으로 취득된 음성 정보에서 구조 요청 소리 및 방향 감지 딥러닝 모델 개발 [2]**

해당 논문은 본 연구 이전에 선행으로 진행되었던 연구에 관한 논문으로써, 드론으로 추출된 스테레오 타입의 음성데이터를 사용하여 해당 음원의 발원지를 60도 간격으로 추정하는 딥러닝 모델을 개발하는 것을 목표로 하였다.

결과적으로 사람의 목소리를 추출하여 해당 목소리의 발원 방향을 추정하는 것까지는 성공하였지만, 해당 추정의 범위가 60도 정도로 실제 현장에 해당 각도를 적용시켜 본다면 꽤나 넓은 범위가 선택되어 희생자의 정확한 위치파악이 어렵다. 게다가 스테레오 타입의 음성데이터에서 노이즈를 제거하는 부분이 구현되지 않았기 때문에 실제 다른 데이터들이 실시간으로 적용되었을 때 사람의 목소리만을 추출해내기가 어렵고, 마찬가지로 남녀노소의 클래스 구분을 할 수 없는 모델이라는 단점이 있다.

1. **시스템 모델**
   1. **기존 연구와 차이점 및 해결방안**

데이터는 기존 연구와 동일한 데이터를 이용해 연구를 진행하지만, 해당 데이터로부터 추출할 특징의 개수가 달라질 것이다. 기존에는 MFCC와 GCC-PHAT 두 가지 특징을 사용하여 연구를 진행하였지만, 이번에는 그것보다 특징의 개수를 더 늘려서 연구를 진행할 예정이다. 기존에 사용하였던 MFCC와 GCC-PHAT외에도 LPC Cepstrum, FTT Cepstrum, PLP Spectrum, STFT 등을 음성 데이터로부터 추출하여 더욱 더 풍성한 특징들을 사용할 계획이다.

또한 기존 연구에서는 아주 기본적이고 간단한 CNN 모델을 사용하여 음성의 발원방향을 추정하였지만, 이번에는 더욱 더 성능을 향상시키기 위하여 더욱 깊고 복잡한 신경망 모델을 사용할 예정이다.

1. **프로젝트 내용**

이 프로젝트는 노이즈가 포함되어있는 스테레오 타입의 음성데이터로부터 노이즈를 제거하여 사람의 구조요청 소리만을 추출해내고, 해당 음성의 발원 방향을 추정함과 동시에 남녀노소 클래스를 구분할 수 있는 딥 러닝 모델을 개발하는 프로젝트이다.

현재는 노이즈 제거 단계가 완료되고, 특징 추출 단계를 진행하고 있는 상황이다.

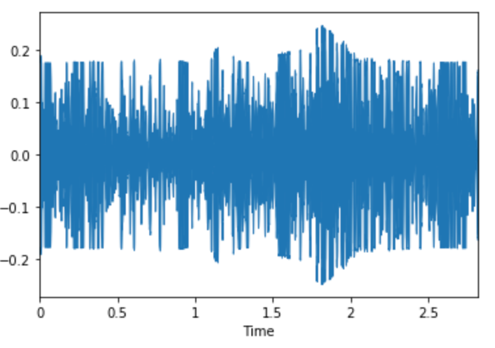
* 1. **노이즈 제거**

노이즈가 포함된 음성 데이터에서 사람의 음성만을 남겨놓고 나머지 노이즈들을 제거하기 위해 다음과 같이 세 가지 필터를 사용하였다.

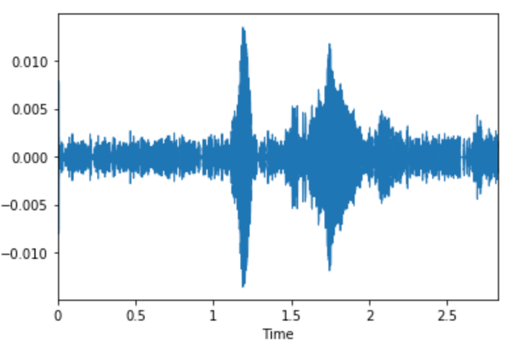
첫 번째는 Band Pass Filter로 높은 주파수와, 낮은 주파수의 범위를 정해주어 원하는 범위의 주파수만을 남기고 나머지를 제거하는 필터이다. 해당 단계에서 제일 많은 노이즈들을 제거하였다.

그 다음으로는 자잘한 노이즈들을 제거하는 단계가 남아있었는데, 코드를 사용해서 해당 노이즈들을 제거하기가 쉽지 않아 방향을 바꾸어 노이즈를 제거하는 대신에 사람의 목소리를 강화하는 쪽으로 진행하였다.

MFCC, Centroid, Power/Energy 이렇게 세 단계를 진행하였고, 그 결과 아래의 그림처럼 노이즈가 상당히 제거되고 사람의 목소리가 추출되었음을 확인할 수가 있다.



[노이즈 제거 전]



[노이즈 제거 후]

* 1. **데이터 전처리 및 특징 추출**

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

위의 과정을 통해 두 개로 나누어진 데이터는 먼저 방향 추정 모델의 특징 벡터를 위해 MFCC, GCC-PHAT, STFT, PLP Cepstrum, SFTF와 같은 특징 추출 단계를 거쳤다.

MFCC는 음정의 높낮이에 영향을 받지 않기 때문에 음성처리에서의 특징으로 대표적이며, GCC-PHAT은 음원이 두 마이크로 도달하는 시간 차이(TDOA)를 구하는 기법 중 하나이다. [3][4] 채널 간의 시간 차를 나타내는 특징으로 시간적 변화에 민감한 변화인 PLP Cepstrum을 추출하였고, 마지막으로 전체적인 주파수 정보를 포함하고 있는 STFT 역시 특징 벡터에 포함시켰다.

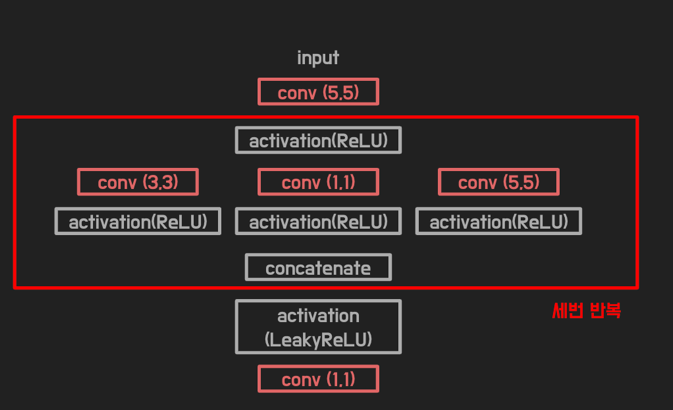
처음에는 데이터 손실을 최대한 줄여보고자 (2065, 150) 형태의 벡터로 특징을 추출하였는데, 필요없는 정보도 함께 포함되었기 때문인지 정확도도 낮고 학습 시간 역시 굉장히 오래 걸리는 관계로 최종적으로 (120, 150) 크기의 벡터로 특징을 추출하였다.

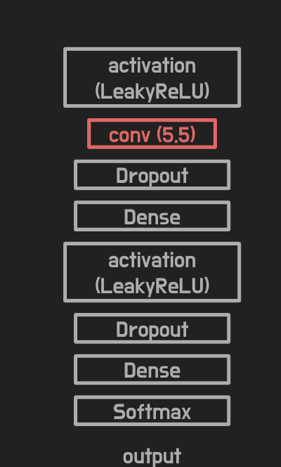
성별/나이 추정을 위한 특징은 다음과 같이 10차/12차 LPC Cepstrum 계수, 8차/12차 FFT Cepstrum 계수를 사용하였다. 처음에는 이렇게 4개의 특징만 추출하여 사용할 예정이었으나, 정확도 향상을 위해 MFCC 특징 역시 포함하여 추출해주었다.

* 1. **방향 추정 모델 설계**

방향 추정 모델은 다음과 같이 설계하였다.

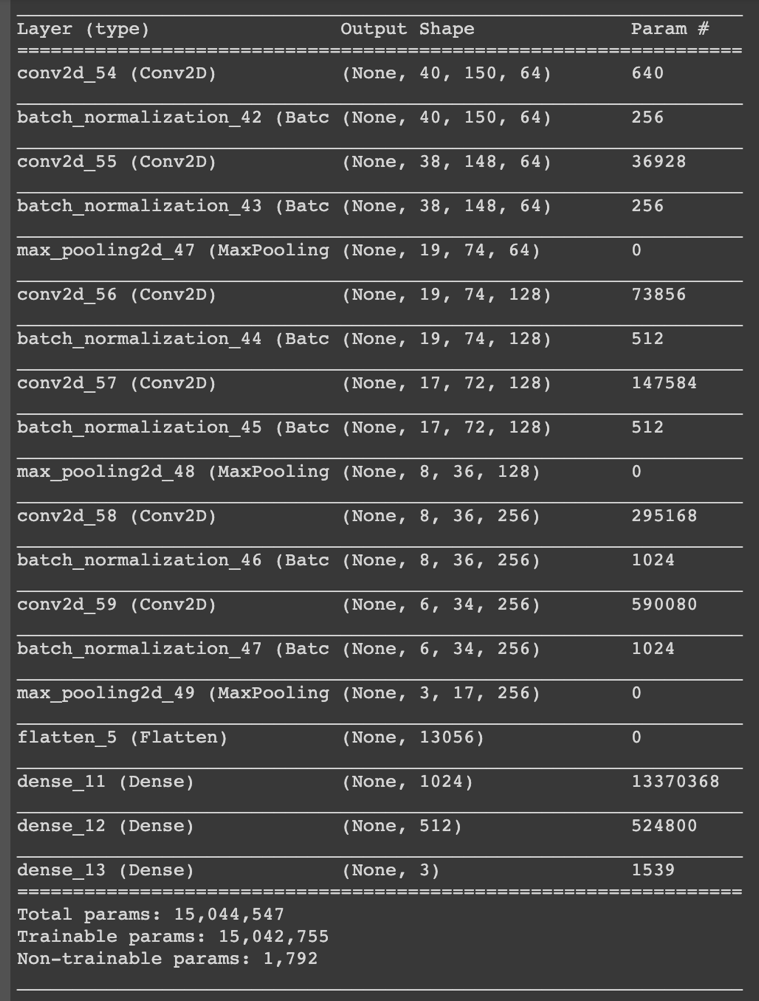
(설계도가 너무 복잡한 관계로 블록 구조도 첨부)





* 1. **성별/나이 추정 모델 설계**

성별/나이 추정 모델은 다음과 같이 설계하였다.



* 1. **모델 설정**

DeadReLU 문제를 제어하고자 ReLU뿐만아니라 LeakyReLU를 혼합하여 사용하였고, Activation 함수 역시 여러 Activation함수들의 Loss를 비교하여 가장 최적의 함수라고 생각되었던 Nadam을 사용하였다.

배치사이즈는 256으로 학습시켰으며 손실함수는 Categorical Cross Entropy를 사용하였다.

이 외에도 오버피팅을 문제를 해결하기 위해 배치 정규화 층과 드롭아웃 층을 추가하였고, L2 Norm을 사용해주었다. 그럼에도 오버피팅이 완전히 잡혀지지 않아, Early Stopping 기능을 사용해 30개의 에폭동안 Validation loss가 더 이상 줄어들지 않으면 모델 학습을 멈추도록 설정하였다.

* 1. **모델 평가**

모델을 검증하는 데에는 K-Fold Validation, 즉 K겹 교차 검증 방법을 이용하여 모델을 검증하였으며, 이 때의 K는 5로 설정하여 총 5번의 검증을 진행하도록 하였다.

그 결과 방향 추정 모델의 정확도는 91.89%가 나왔으며 성별/연령 추정 모델의 정확도는 97.98%로 상당히 높은 정확도를 기록하였다.

1. **결론 및 향후 연구**

현재 모델의 정확도도 낮지 않은 편이지만, 조금 더 전처리와 특징 추출 단계를 최적화 한다면 더욱 더 높은 정확도의 모델을 학습시킬 수 있을 것이라고 생각한다. 또한, 20도 간격으로 발화 방향을 추정하는 것보다 더욱 더 세밀한 10도 간격의 발화 방향을 추정한다던지, 방향 뿐만아니라 드론으로부터의 거리가 어느 정도 되는지를 추정한다면 더욱 더 대형 복합 재난의 경우에 도움이 될 수 있을 것이라고 생각한다.

**참 고 문 헌**

**[1]** Rameesha Tariq, Maham Rahim, Nimra Aslam, Narmeen Bawany, Ummay Faseeha, "DronAID : A Smart Human Detection Drone for Rescue", IEEE, 2018.

**[2]** 이수인, 이민혜, 박준혁, “드론으로 취득된 음성 정보에서 구조요청 소리 및 방향 감지 딥러닝 모델 개발”, KCC, 2019

[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