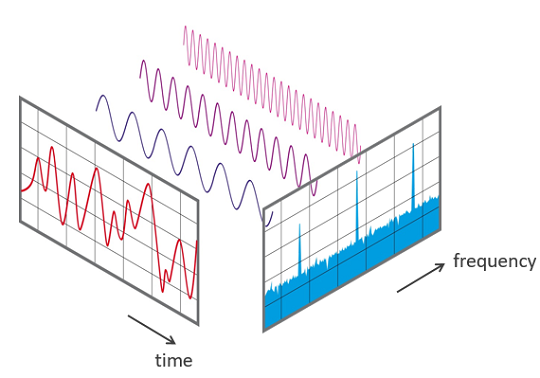
**1. 오디오 데이터**

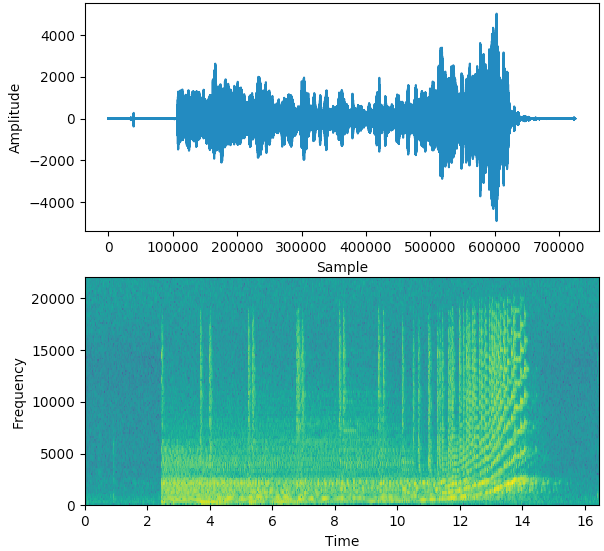
소리를 디지털 데이터로 표현하기 위해 특정 시간마다 진폭의 크기를 수치화 하여 배열 형태의 데이터로 저장한다. 이 때 소리 정보를 sampling하는 빈도수를 sample rate라고 하며, 만약 sample rate = 40000이면 2초의 음성 데이터는 80000의 길이를 갖는 배열로 전환할 수 있다. 아날로그 소리 정보를 보다 잘 표현하기 위해서는 높은 sample rate이 필요하고, 이는 데이터의 크기가 굉장히 커지는 이유가 된다. 따라서 딥러닝 모델의 입력으로는 소리를 샘플링한 데이터가 바로 들어가는 것이 아니라, Fourier transform을 거친 spectrogram 형태의 이미지 데이터로 전환하여 사용된다.

**2. Spectrogram**

실제 소리는 하나의 주파수만을 가지는 것이 아니라, 여러 주파수의 파형이 합친 형태로 존재한다. 특정 주파수가 얼만큼 존재하는지를 알기 위해서는 Fourier transform을 이용하면 된다. 기존의 소리 정보는 시간에 따른 진폭의 크기 변화이다. 이는 곧 (시간 vs 진폭) 형태의 데이터로 볼 수 있다. Fourier transform을 거치면 이를 (주파수 vs 크기) 형태의 데이터로 전환할 수 있다. 즉, 입력 소리에 특정 주파수가 얼마만큼 강하게 존재하는지 알 수 있다.



한 단계 더 나아가서, 소리는 여러 주파수의 합으로 구성됨과 동시에 이러한 주파수의 구성은 시간에 따라 변한다. 시간에 따라 어느 주파수로 구성 되어있는지를 표현한 것을 spectrogram이라고 한다. 이를 구하기 위해서는 특정 time window마다 Fourier transform을 계산하고, 이 window를 sliding하면서 나온 결과를 합치면 된다. Spectrogram은 y축을 주파수, x축을 시간으로 하는 데이터로 볼 수 있다. 여기서 강도를 색으로 표현한다면 아래와 같은 이미지로 확인할 수 있다. Spectrogram은 소리를 표현하는 이미지로 볼 수 있다. 전체 데이터를 한 번에 보는 것이 아니라, 이미지와 같이 local feature가 중요하다. 따라서 이미지 데이터와 같이 CNN 구조의 모델을 사용할 수 있으며, image augmentation과 같은 방법을 사용하여 학습 데이터를 늘릴 수도 있다.



**3. Augmentation**

Augmentation을 적용하는 단계에 따라 두 가지고 나눌 수 있다. 우선, 실제 음성 파일을 sampling 하여 얻어지는 raw data에 적용하여 다양한 데이터를 얻을 수 있다. 두 번째로는 spectrogram에 augmentation을 적용하여 다양한 데이터를 얻을 수도 있다.

**3.1 Raw data**

음성 파일에서 sampling만 거친 raw data에 백색 소음(noise)을 추가한다. Noise가 추가되어도 원본의 억양에는 변화 없으므로, 모델이 보다 더 본질적인 특징을 학습하여 억양을 분류할 수 있을 거라 예상할 수 있다.

Raw data의 전체 크기를 조절하여 마치 음성의 재생 속도를 조절하는 듯한 변화도 가능하다. 랜덤하게 음성 데이터를 느리거나 빠르게 할 수 있다. 하지만 너무 과한 속도 조절은 원본 정보를 잃게 할 수도 있다.

Sampling을 통해 raw data를 형성할 때, sampling이 시작되는 위치를 조절하여 다양한 학습 데이터를 얻을 수 있다. 특히 본 프로젝트에 주어진 학습용 음성 파일은 모두 길이가 달라, 편의성을 위해 가장 작은 길이로 자르는 과정이 존재한다. 이 때 처음부터 원하는 길이만큼 자른다. 하지만 음성 파일의 중간에서 시작해도 모두 같은 억양을 가지고 있으므로, 시작 위치를 여러 곳으로 하여 하나의 음성 파일에서 보다 많은 raw data를 얻을 수 있다.

**3.2 Spectrogram**

앞서 spectrogram은 소리를 표현하는 이미지와 같다고 말하였지만, image augmentation의 대표적인 방법인 flip이나 rotate를 적용할 수는 없다. 이들은 spectrogram의 의미를 훼손하기 때문이다.

본 의미를 훼손하지 않는 대표적인 방법에는 masking이 있다. 아래 그림과 같이 특정 주파수나 시간 대를 상수 값으로 처리한다. 이런 mask가 추가된 데이터까지 학습 데이터에 포함시켜 모델이 더 본질적인 특징을 학습할 수 있게 해준다.

텍스트, 시계, 디스플레이, 장치이(가) 표시된 사진

자동 생성된 설명

기타 방법으로는 spectrogram을 만들 때 들어가는 hyper parameter를 변경하여 하나의 raw data로부터 다양한 spectrogram을 얻는 방법이 있다. 예를 들어 n\_mels를 변경하면 spectrogram의 height이 바뀐다. 또한 n\_fft와 hop\_length를 변경하면 윈도우 크기가 달라져 spectrogram의 width가 달라진다. Spectrogram의 height와 width는 그 음성 파일을 바라보는 해상도에 해당하므로, 다양한 관점에서의 해상도로 학습하면 모델의 성능이 향상될 것이라 예측할 수 있다.