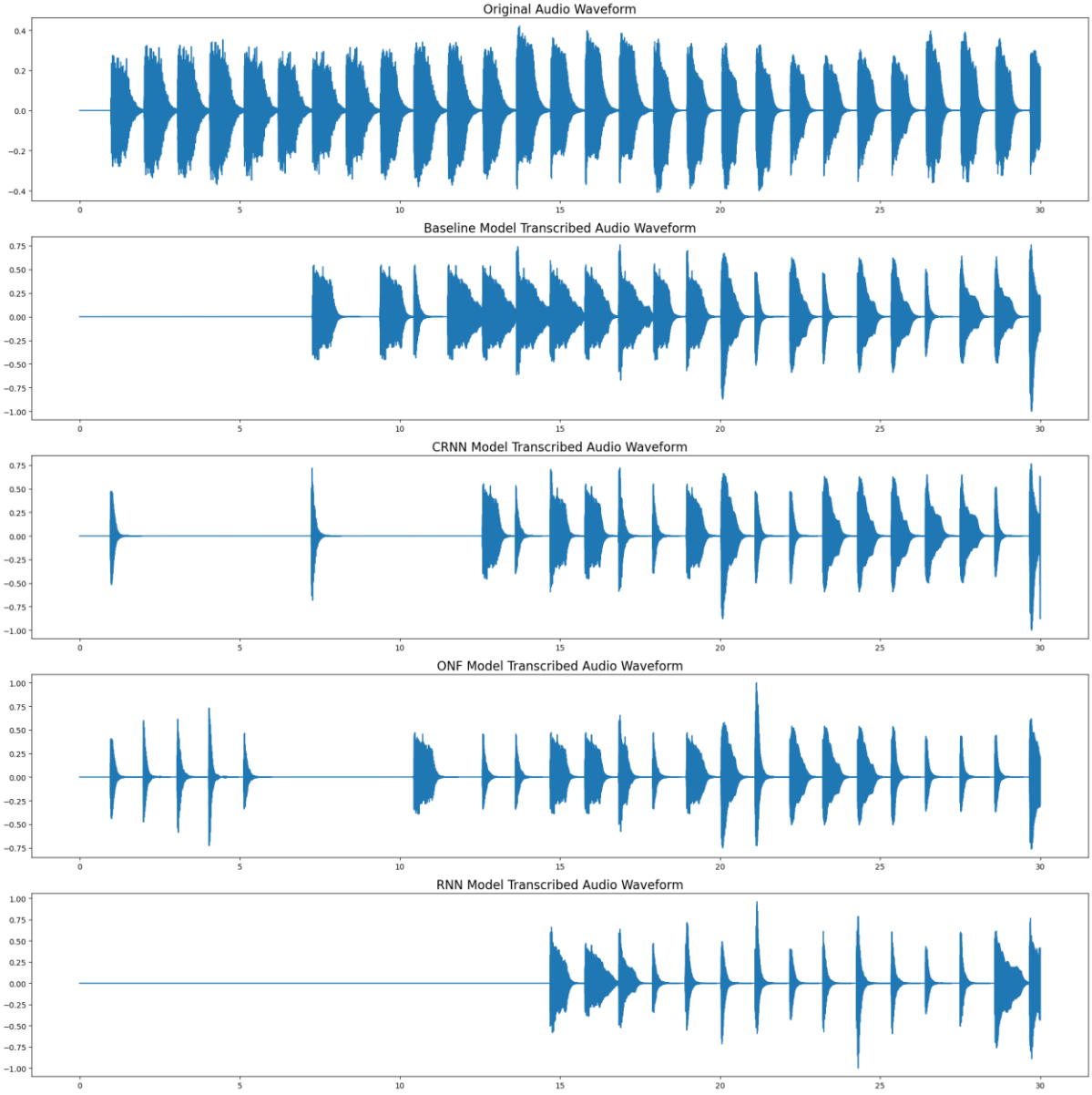
MIR Assignment 4 Report

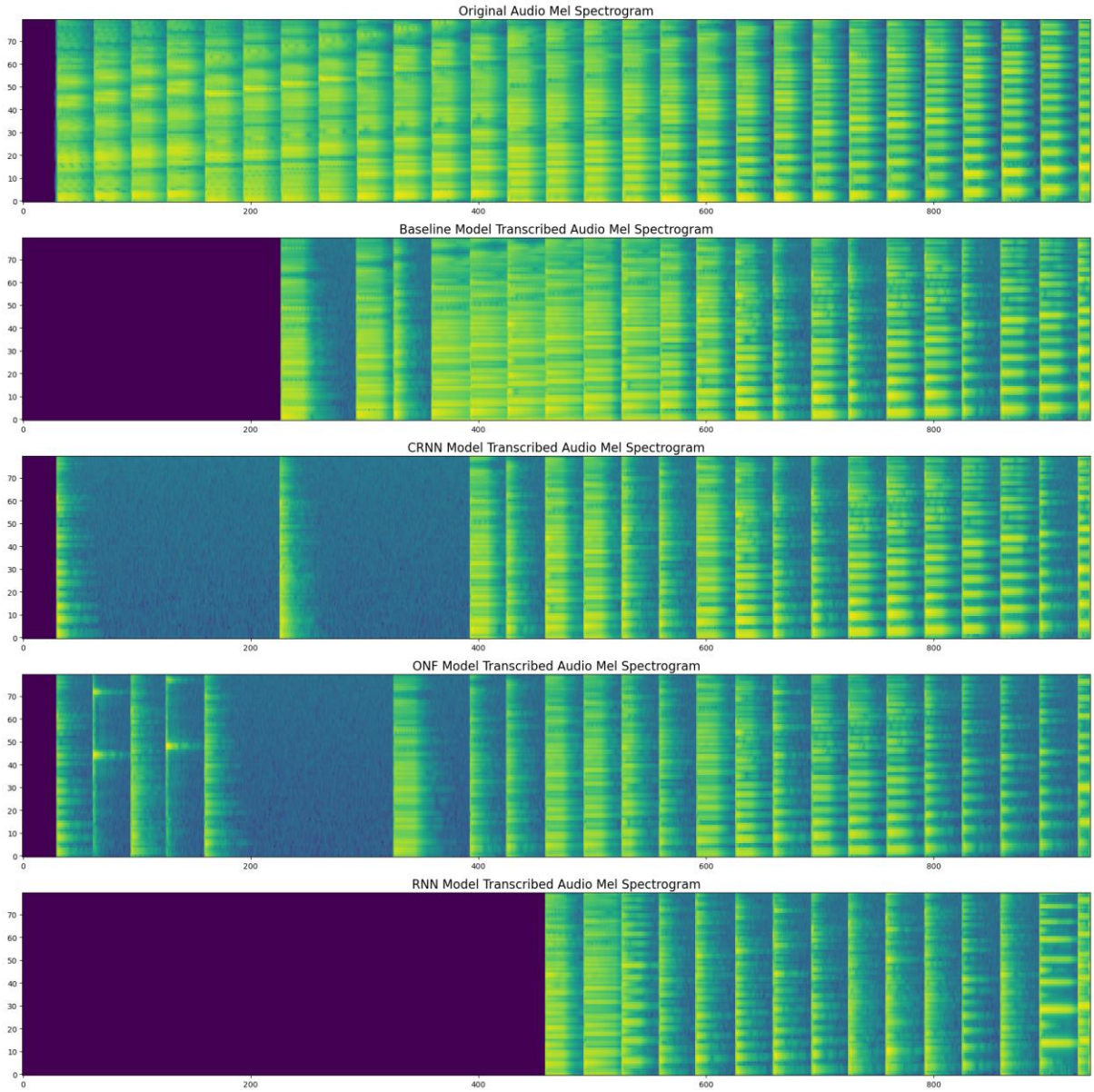
20191048 김도솔

1. FMP_C3_F03.mp3

1) 오디오 파형 (Waveform)

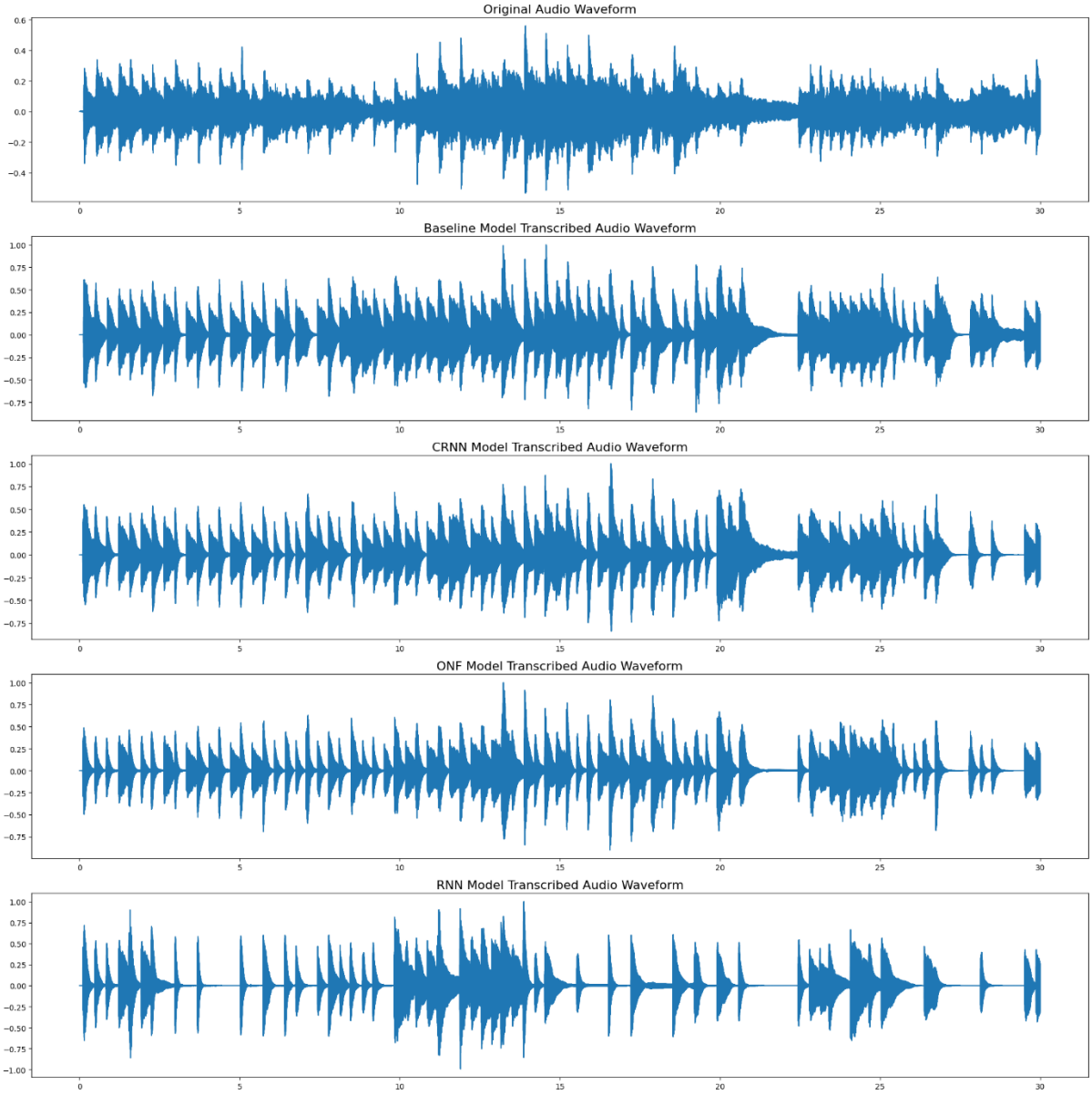


2) 멜 스펙트로그램 (Mel Spectrogram) 비교 분석

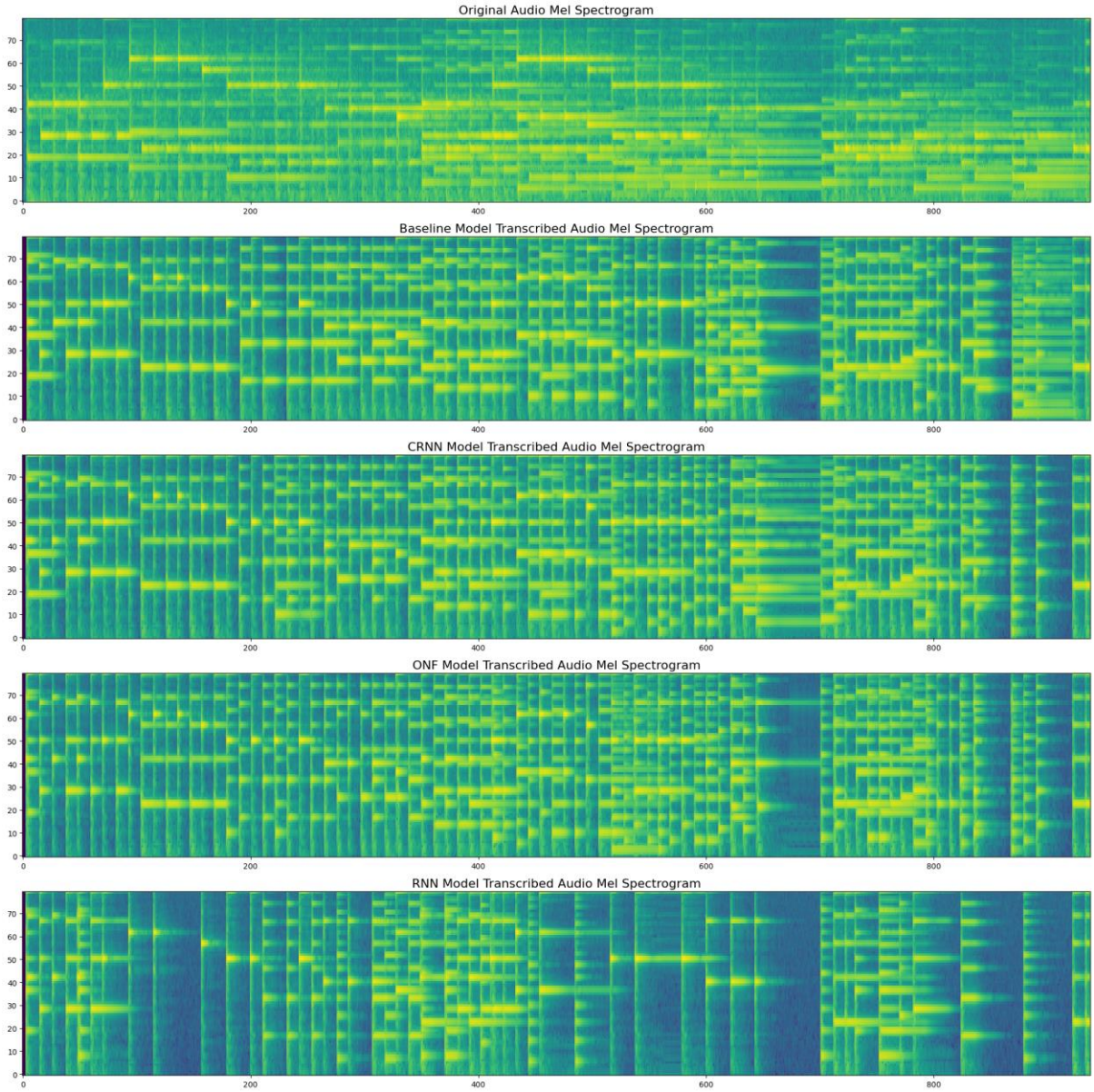


2. faded.mp3

1) 오디오 파형 (Waveform)

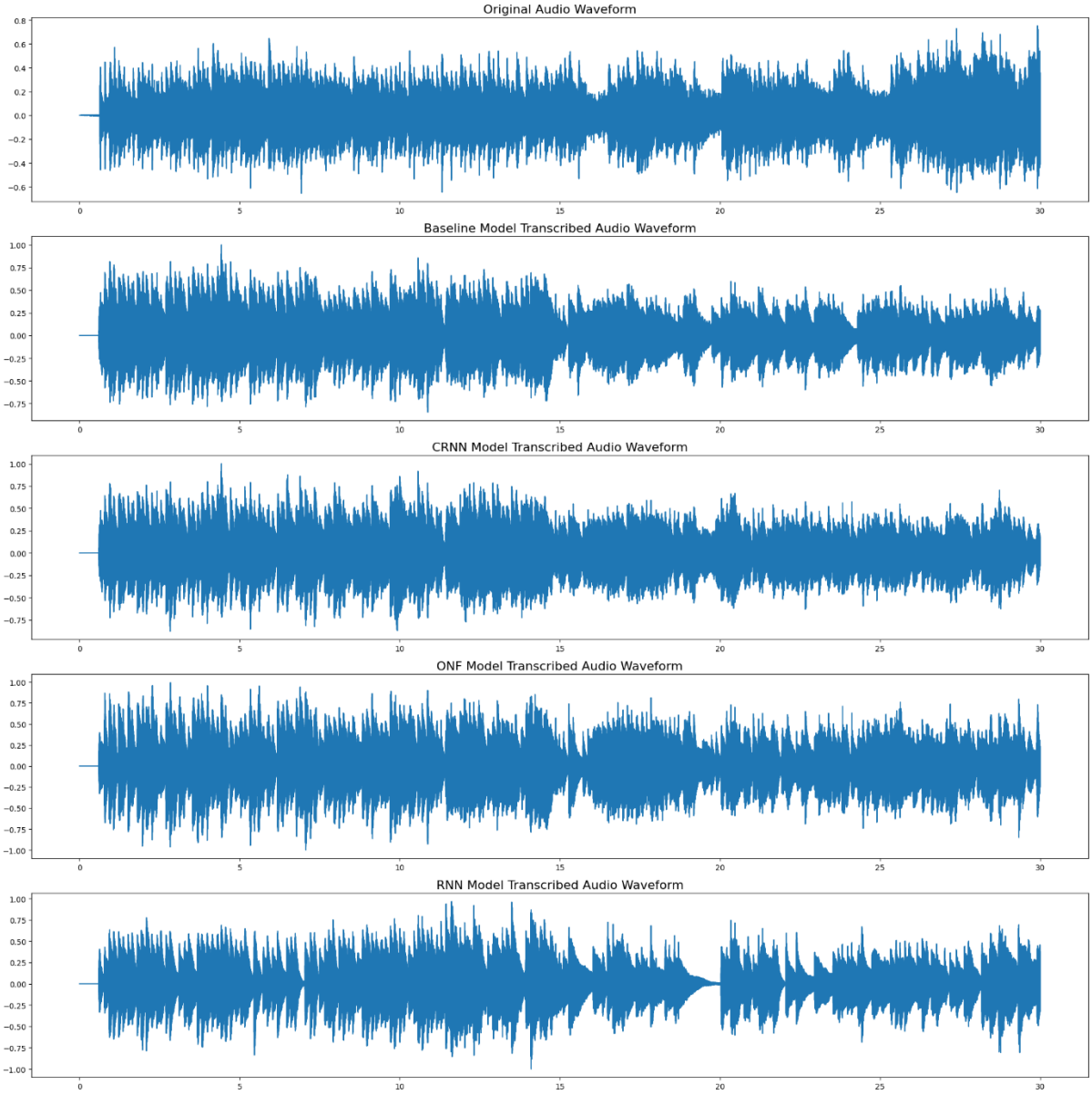


2) 멜 스펙트로그램 (Mel Spectrogram)

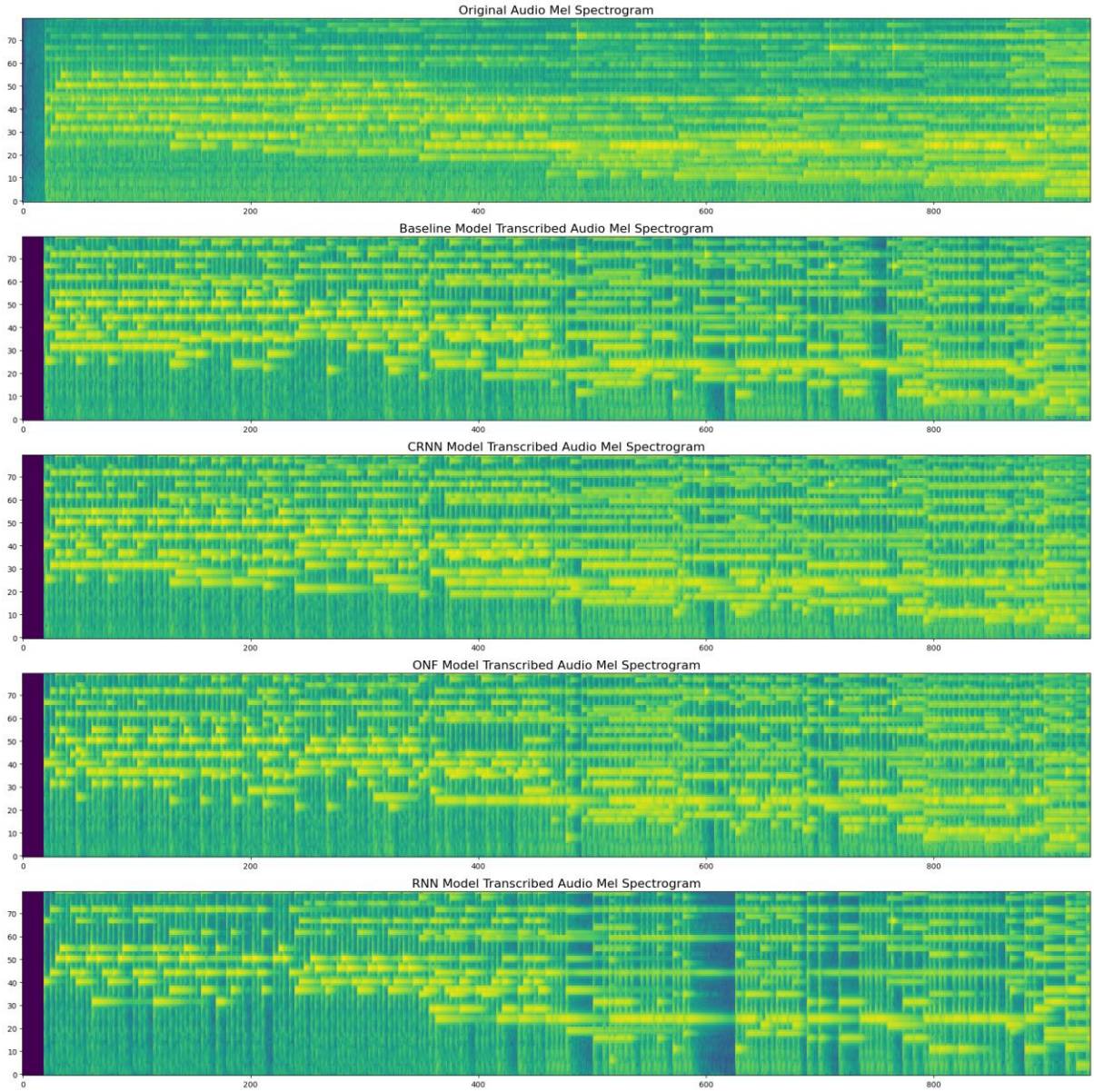


3. golden_hour.mp3

1) 오디오 파형 (Waveform)

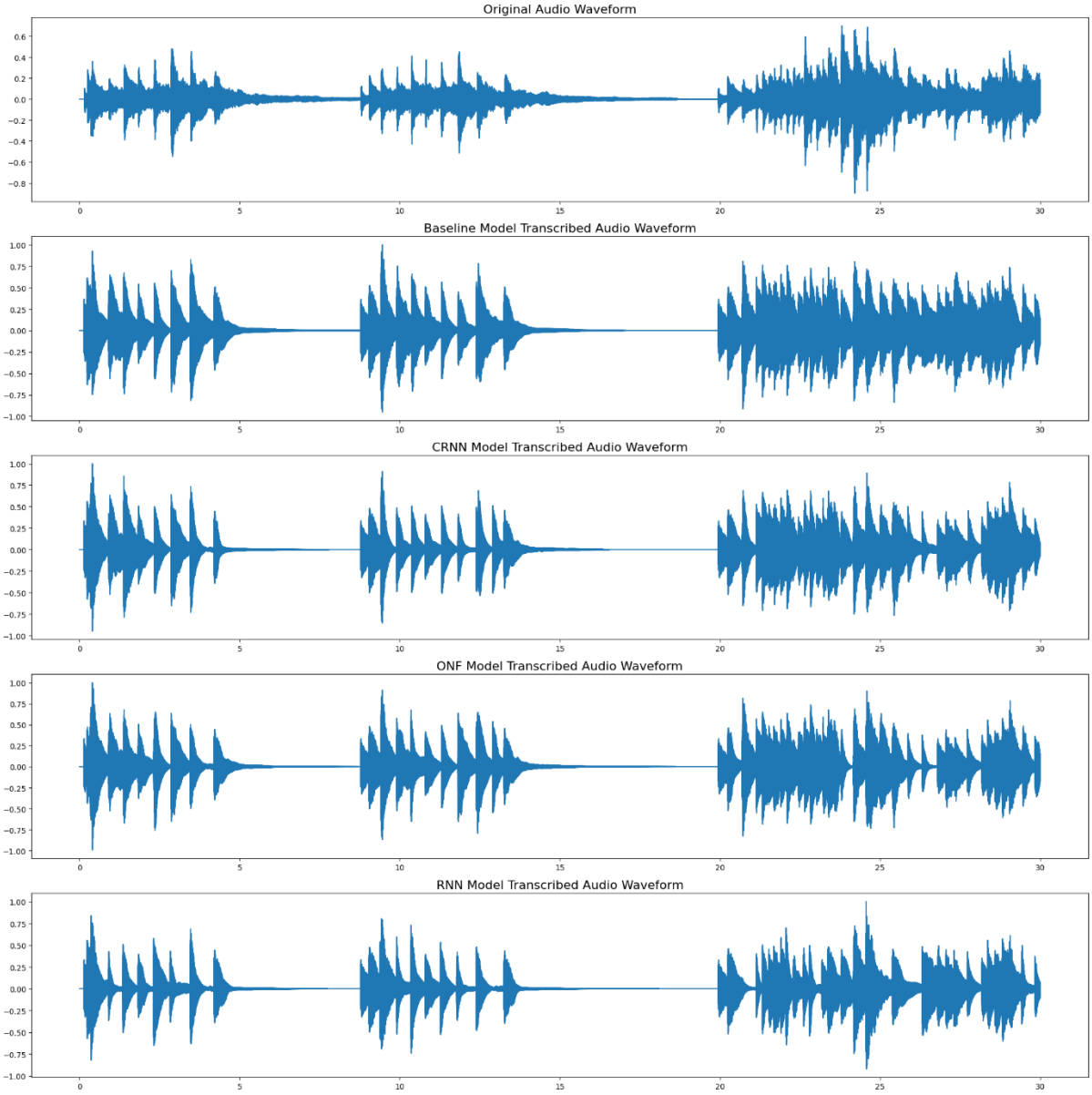


2) 멜 스펙트로그램 (Mel Spectrogram)

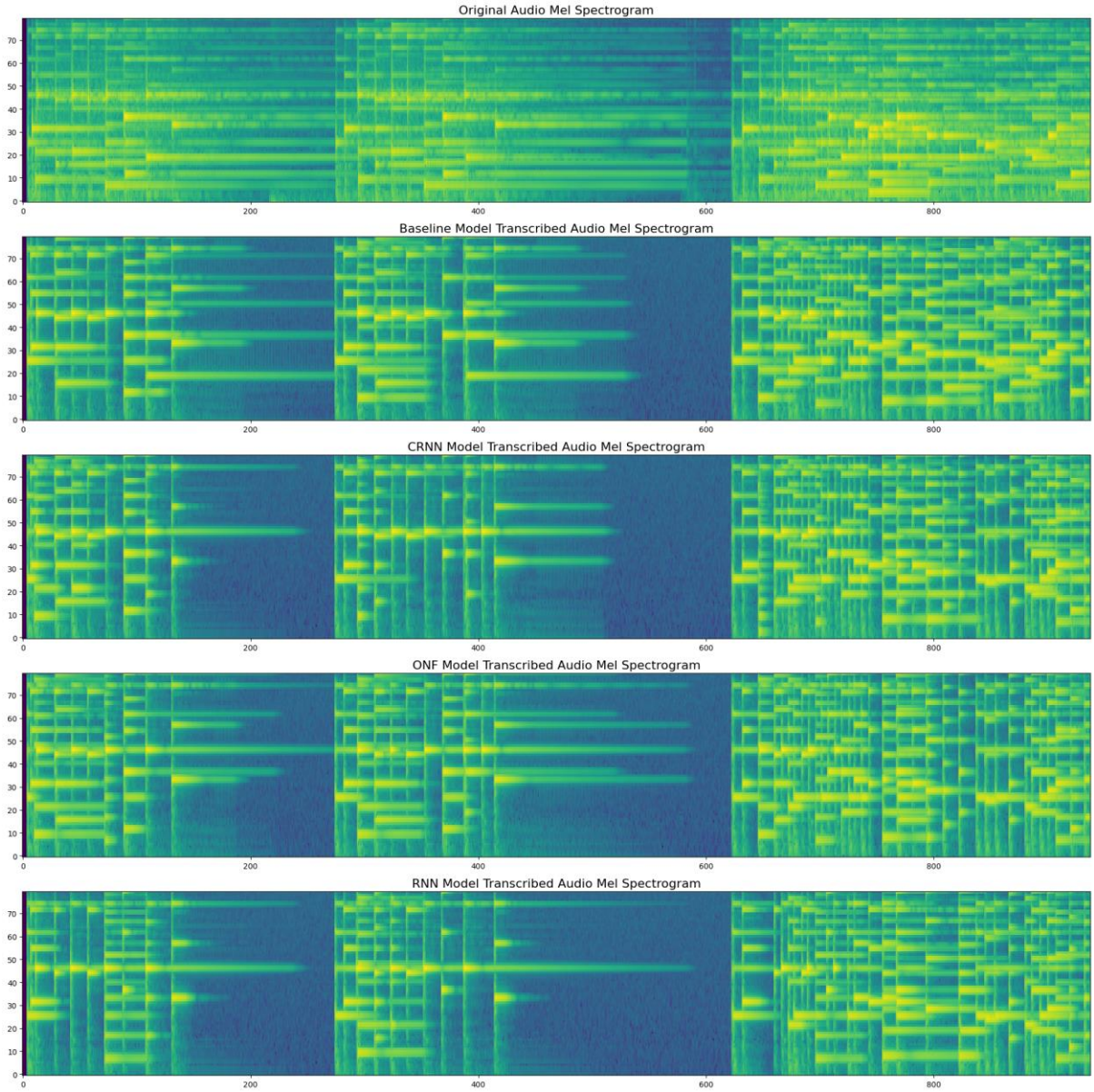


4. river_flows_in_you.mp3

1) 오디오 파형 (Waveform)



2) 멜 스펙트로그램 (Mel Spectrogram)



5. 분석

1) 모델 별 분석

Baseline Model의 경우, 원본 오디오와 비교했을 때 낮은 음이 제대로 인식 되지 않는 것을 확인할 수 있다. 음의 길이는 다른 모델들에 비해 잘 인식하는 편이다. 또한 멜 스펙트로그램에서 나타나는 주파수 대역이 다양한 축에 속해 화음이 잘 들리는 편이다.

CRNN Model도 마찬가지로 낮은 음이 잘 인식되지 않지만 CNN과 RNN 모델의 하이브리드 버전이라 둘에 비해서는 낮은 음 채보가 잘 이루어진 모습을 확인할 수 있다. 멜 스펙트로그램에서 나타나는 주파수 대역은 다양하지 않은 편이다.

ONF Model은 다른 모델들과 비교해 초반부(낮은 음)를 잘 인식하는 편이다. 하지만 낮은 음의 길이를 잘 인지하지 못해 실제 파일을 들어보면 스타카토처럼 짧게 끊겨서 들리는 단점이 있다. 높은 음의 경우는 음의 길이가 긴 경우(sustain pedal을 밟은 경우)도 다른 모델들에 비해 잘 인식하여 채보하는 편이다.

RNN Model 또한 낮은 음을 인식하는데 어려움이 있는 것으로 보인다. 또한 원본과 비교했을 때, 복잡한 피아노 연주의 주파수 패턴을 충분히 반영하지 못하고 단순한 패턴만 학습하는 모습을 보인다. 또한 다른 모델들보다 주파수 대역이 제한적인 모습을 보인다.

2) 종합 분석

전반적으로는 모든 모델이 FMP_C3_F03.mp3의 초반 부분을 잘 채보하지 못하는 모습을 보였다. 이는 모델들이 낮은 음을 잘 인식하지 못해 발생하는 결과로 예상된다. 이러한 문제가 발생하는 이유는 Maestro 데이터셋이 낮은 음보다는 높은 음의 피아노 연주나 스타일에 집중되어 있기 때문으로 추측된다.

또한 원본 오디오의 복잡한 음향 정보를 제대로 반영하지 못해 단순화된 패턴을 나타내는 양상을 보였다. 특히 주파수 대역이 원본 오디오에 비해 제한적으로 나타나거나, 음의 길이를 끝까지 채보하지 못하고 중간에 끊겨서 스타카토처럼 들리는 한계가 있었다.

그래프 상으로 비교하고, wav file도 직접 들어본 결과 RNN 모델이 가장 단순화된 모습을 보였고, ONF 모델이 그나마 원본 오디오에 가까운 결과를 보였다. RNN 모델은 시간적 종속성을 잘 학습하지만, 긴 시퀀스를 처리하는 데 한계가 있어 음의 길이를 끝까지 반영하지 못하고 중간에 끊기는 현상이 발생한다. 또한 단순한 구조로 인해 음향 신호의 복잡한 특성을 충분히 반영하지 못하는 것으로 보인다. 반면 ONF 모델은 CNN과 RNN을 결합하여 주파수 및 시간적 특성을 모두 학습할 수 있어, 피아노 연주의 복잡한 특성을 더 잘 반영할 수 있는 것으로 보인다. 이러한 장점에 추가적으로 온셋과 프레임을 분리하여 학습함으로써 음의 시작 지점과 지속 시간을 정확히 반영하여, 다른 모델들에 비해 더 정확한 채보가 가능한 것으로 보인다.