

**SK네트웍스 Family AI 과정 14기  
 데이터 전처리 인공지능 학습 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 학습 결과서 |
| **제출 일자** | 2025.08.29 |
| **깃허브 경로** | [4팀\_깃허브](https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-4Team) |
| **작성 팀원** | 공지환 |

1. **모델 비교 및 선정 이유**

* 비교 대상 모델:

| **모델명** | **종류** | **선정 이유** |
| --- | --- | --- |
| MeloTTS | VITS2 기반 TTS | 한국어 자연 발화 지원, 학습/추론 코드와 config 구조가 명확하여 파인튜닝에 용이 |
| mms-tts | 다국어 지원 TTS(Meta) | 한국어 파인튜닝 시 uroman 전처리 필수 |

* 실험 모델 수: 총 2종
* 최종 선정 모델: MeloTTS

1. **모델 구조 및 아키텍처**

* 모델 아키텍처 도식
  + 입력 텍스트 → 임베딩 → 인코더블록(Conformer) → Variational Latent Space → 출력 오디오
  + (Vit2 구조 기반)
* 구성 요소 설명:

| **계층명** | **역할** | **구성 요소** |
| --- | --- | --- |
| 임베딩 | 텍스트 벡터화 | 글자 단위 임베딩 |
| 인코더 | 발화 단위 텍스트 추출 | Conformer 블록 N개 |
| 변분 인코더 | 텍스트 발화 다양성 학습 | Gaussian Posterior Network |
| 디코더 | 오디오 파형 생성 | Flow 기반 디코더 + HifiGAN |
| 출력층 | 오디오 출력 | Waveform (16kHz, mono) |

1. **학습 설정 및 하이퍼파라미터**

| **항목** | **값** |
| --- | --- |
| 학습 데이터 수 | 41 |
| 검증 데이터 수 | 4 |
| 에폭(Epoch) 수 | 10 |
| 배치 크기 (Batch Size) | 8 |
| Gradient Accumulation | 2 |
| 학습률 (Learning Rate) | 1e-4 |
| 옵티마이저 | AdamW |
| Sampling Rate | 44,100Hz |
| STFT 파라미터 | filter\_length=2048,  hop\_length=512,  win\_length=2048 |
| Segment Size | 32프레임 |
| 화자 | n\_speakers=256,  spk2id={KR:0, Hong:1}  (Hong 사용) |

1. **학습 결과 및 성능 평가**

* 학습 결과 요약 : 학습 실패  
  + 실패 이유:
    - 모델 train시 순환참조 문제 해결 실패
    - 마지막 에러:
      * MemoryError
      * 메모리 에러 임의 해결 시 RuntimeError 발생
      * Checkpoint 로딩 코드의 무한 재귀가 문제로 추정

1. **향후 계획**

* 다른 모델 탐색 및 학습 재시도
* conqui-ai/TTS
* MeloTTS 학습실패 원인 파악 후 재시도