Date: 2025-09-03 (Wed) 📅
Prepared by: Tiong-Sik Ng 📝

English 🇺🇸

1. Summary of Accomplishments ✅
   • Tried three local TTS options end-to-end today: **XTTS**, **F5**, and **OpenVoice**; decided to proceed with **OpenVoice** for the app because it offers the best balance of **low latency** and **good speech quality** among the three (still less natural than FishSpeech, but more human-sounding than XTTS).
   • Measured OpenVoice latency (excluding model init/ckpt load): **Avg TTS 4775.3 ms**, **Avg Convert 1685.0 ms**, **Avg Total 6460.3 ms**.
   • Tuned speaking rate: confirmed **speed=1.25** sounds best (1.35 felt rushed).
   • Built a **drop-in FastAPI/UVicorn app** that swaps FishSpeech for OpenVoice while keeping all function names, arguments, utilities, and endpoints intact; preserved sentence-chunked streaming to S3 as μ-law WAV.
   • Fixed runtime/setup issues encountered during testing:

- NLTK tagger resource error (averaged_perceptron_tagger_eng) → documented the download fix.
- Python 3.9 typing crash (`np.ndarray | torch.Tensor`) → replaced with `Union[...]`.
- FastAPI 400s from Lambda calls → identified cause (**missing `speaker_wav`**) and documented request/aliasing fixes.
  • Korean comms: translated an evaluation note about OpenVoice vs FishSpeech/XTTS for sharing.

2. Issues & Risks 🔄
   • **Voice reference requirement**: OpenVoice conversion currently **requires a speaker reference**; 500s occurred when none was provided.
   • **Quality trade-off**: OpenVoice < FishSpeech in naturalness for KR, though it sounds more human than XTTS; continued text hygiene and punctuation still important.
   • **Integration nuance**: Lambda payloads previously relied on cached embeddings with Fish; OpenVoice expects either a reference WAV or a cached OpenVoice **target SE (.pth)**.

3. Next Steps 📌
   • **Server-side default speaker**: Preload a default **target SE** at app startup (from a WAV or precomputed `.pth`) and **fallback to it** when `speaker_wav` isn't provided; also add on-disk caching for any new refs.
   • **Optional request simplification**: Keep accepting per-request overrides, but make `speaker_wav` optional when a default is present; relax 400 checks accordingly.
   • **Quality pass**: Re-evaluate KR prosody with speed ~1.20–1.28 and light punctuation adjustments; compare against FishSpeech on a small benchmark script.
   • **Stability & DX**: Harden error messages for missing refs, add health/readiness checks, and document example Lambda payloads for both WAV and `.pth` flows.

날짜: 2025-09-03 (수) 📅
작성자: Tiong-Sik Ng 📝

1. 금일 수행 사항 ✅
   • 오늘 **XTTS**, **F5**, **OpenVoice** 를 순차적으로 테스트하고, 앱에는 **OpenVoice** 를 적용하기로 결정 (세 모델 중 **지연/품질 밸런스**가 가장 좋음). 자연스러움은 FishSpeech 보다 낮지만, XTTS 보다는 사람처럼 들림.
   • OpenVoice 지연(모델 로드 제외): **TTS 평균 4775.3 ms**, **Convert 평균 1685.0 ms**, **총 6460.3 ms**.
   • 발화 속도 튜닝: **speed=1.25** 가 최적, **1.35** 는 다소 급함.
   • FishSpeech → OpenVoice 로 교체한 **FastAPI/UVicorn 앱** 작성(함수명/인자/유틸/엔드포인트 유지). 문장 단위 스트리밍 → S3(μ-law WAV) 파이프 보전.
   • 테스트 중 이슈 해결:

- NLTK 태거 리소스 오류 → 다운로드 가이드 정리.
- Python 3.9 타입 힌트 충돌(`np.ndarray | torch.Tensor`) → `Union[...]`로 교체.
- FastAPI 400 문제 → **`speaker_wav` 누락** 원인 파악 및 요청/별칭 처리 방법 문서화.
  • 한국어 커뮤니케이션: OpenVoice vs FishSpeech/XTTS 비교 메모 번역 및 공유.

2. 이슈 & 리스크 🔄
   • **화자 참조 필요**: OpenVoice 변환은 기본적으로 **레퍼런스 음성**이 필요; 없는 경우 500 발생.
   • **품질 트레이드오프**: KR 자연스러움은 FishSpeech 보다 낮으나 XTTS 보다는 자연스러움; 구두점/문장 분할 등 텍스트 전처리가 여전히 중요.
   • **연동 차이**: 기존 Fish 에서의 캐시 임베딩 의존 → OpenVoice 는 WAV 또는 **타깃 SE(.pth)** 필요.

3. 다음 단계 📌
   • **서버 기본 화자**: 앱 시작 시 기본 **타깃 SE** 를 선 로드(레퍼런스 WAV 또는 .pth)하고, 요청에 `speaker_wav` 가 없으면 자동 사용; 신규 레퍼런스는 디스크에 캐시.
   • **요청 간소화(선택)**: 기본값이 있을 때 `speaker_wav` 를 선택 항목으로 허용; 400 체크 완화.
   • **품질 재점검**: 속도 1.20–1.28 및 간단한 구두점 조정으로 KR 운율 재평가; 소규모 스크립트로 FishSpeech 대비 재비교.
   • **안정화/DX**: 참조 누락 오류 메시지 개선, health/readiness 보강, WAV/`.pth` 예시 페이로드 문서화.