Date: 2025-09-02 (Tue) 📅
Prepared by: Tiong-Sik Ng 📝

English 🇺🇸

1. Summary of Accomplishments ✅
   • Verified Amazon Connect playback end-to-end: preset filler plays; newly generated audio also plays after fixing output format to **WAV/8 kHz/mono/pcm_mulaw** and S3 upload handling.
   • Measured XTTS v2 latency on GPU (single-sentence): ~0.7–1.0 s processing; **RTF ≈ 0.20–0.37** — feels "live."
   • Began **text hygiene + manual chunking** pipeline (punctuation, micro-pauses) and drafted **multi-reference conditioning** (slice long ref into 3–5 clips of 6–10 s) to improve Korean naturalness.
   • Explored **CosyVoice2** (0.5B): environment stood up and inference tested; observed latency comparable to FishSpeech; noted ≤30 s @16 kHz prompt limit for text+audio prompting; parked for later evaluation.
   • Prepared plan to auto-slice long reference audio + optional ASR alignment to produce XTTS-ready refs and a CosyVoice2 prompt.

2. Issues & Risks 🔄
   • **Voice quality/naturalness (KR)**: XTTS still sounds foreign/robotic on some lines; misreads occur without careful punctuation/normalization; build lacks `use_phonemes` flag → rely on text hygiene + multi-ref conditioning.
   • **Model latency**: CosyVoice2 ≈ FishSpeech in our setup; further kernel/AMP/tuning may be needed to beat current times.
   • **Connect UX**: Even with fast synthesis, telephony constraints (8 kHz µ-law, contact-flow timing) can reduce perceived naturalness/snappiness.
   • **Fallback choices**: If XTTS and CosyVoice2 remain unsatisfactory, may need a cloud model (e.g., ElevenLabs Flash) with quota/usage limits, or a small supervised fine-tune of XTTS.

3. Next Steps 📌
   • **Reference prep**: Slice `/home/work/VALL-E/audio_samples/promotional_calls.wav` into 3–5 clean KR clips (6–10 s, PCM16 mono 44.1/48 kHz); register as XTTS speaker refs and re-test with punctuation + 150–220 ms inter-sentence gaps.
   • **Quality gates**: Try light generation knobs (if available): temperature 0.35–0.45, repetition_penalty 1.05–1.10, speed 0.97–1.02; optionally add ASR self-check (CER) to select the best take for critical prompts.
   • **CosyVoice2 retry (optional)**: Create ≤20 s **prompt_16k.wav** with matching prompt text; test `<|ko|>` + `text_frontend=False` path for prosody steering; compare naturalness.
   • **Server ergonomics**: Add server-side **auto-slice & cache** for a single long reference upload; expose engine toggle (FishSpeech / XTTS / CosyVoice2) and keep **Connect seatbelt** (8 kHz µ-law to S3).
   • **Alternatives**: If quality remains below bar, evaluate cloud TTS (quota permitting) or plan a **small XTTS fine-tune** with 5–20 min of clean KR audio+text.

날짜: 2025-09-02 (화) 📅
작성자: Tiong-Sik Ng 📝

1. 금일 수행 사항 ✅
   • Amazon Connect 에서 프리필러 및 신규 합성 오디오 **재생 검증** 완료(출력 포맷을 **WAV/8 kHz/mono/pcm_mulaw** 로 고정, S3 업로드 정리).
   • **XTTS v2** GPU 기준 단문 지연 측정: 처리시간 약 0.7–1.0 초, **RTF ≈ 0.20–0.37**(체감상 실시간에 가까움).
   • 한국어 자연스러움 개선을 위해 **문장 분할/중간 무음 삽입** 및 **복수 참조 음성(6–10 초, 3–5 개)** 전략 초안 마련.
   • **CosyVoice2(0.5B)** 탐색/실행 확인: 지연이 FishSpeech 와 유사, **16 kHz 30 초 이내**의 프롬프트 제약 확인 → 후속 재평가 예정.
   • 장문 레퍼런스 음성 + 텍스트를 기반으로 **XTTS 용 참조 클립 자동 생성**(필요 시 ASR 정렬) 계획 수립.

2. 이슈 & 리스크 🔄
   • **한국어 자연스러움**: 일부 문장에서 외국인 억양/로봇톤; 띄어쓰기/구두점 미흡 시 오인식; `use_phonemes` 부재 → 텍스트 정규화와 다중 레퍼런스에 의존.
   • **모델 지연**: CosyVoice2 ≈ FishSpeech → 추가 최적화 필요 가능.
   • **Connect 제약**: 8 kHz μ-law/플로우 타이밍으로 체감 품질/속도 감소.
   • **대안 검토**: XTTS/CosyVoice2 품질 미달 시 클라우드 모델(한도 이슈) 또는 XTTS **소규모 파인튜닝** 고려.

3. 다음 단계 📌
   • **레퍼런스 준비**: `/home/work/VALL-E/audio_samples/promotional_calls.wav` 에서 3–5 개(6–10 초) 구간 추출 → XTTS 다중 참조로 등록 후 문장별 합성/간격(150–220 ms) 적용 재시험.
   • **품질 튜닝**: (지원 시) temperature 0.35–0.45, repetition_penalty 1.05–1.10, speed 0.97–1.02; 중요 문장은 **ASR-CER** 로 베스트 샷 선택.
   • **CosyVoice2 재시도(선택)**: ≤20 초 **prompt_16k.wav** + 매칭 텍스트로 `<|ko|>/text_frontend=False` 테스트.
   • **서버 개선**: 단일 장문 레퍼런스 업로드 → 서버에서 **자동 슬라이싱/캐시**; 엔진 토글 및 **8 kHz μ-law → S3** 시트벨트 유지.
   • **대안 경로**: 필요 시 클라우드 TTS 평가 또는 **XTTS 파인튜닝** 착수.