📅 Date: 2025 년 8 월 12 일
📝 Prepared by: Tiong-Sik Ng

English 🇺🇸

1. ✅ Summary of Accomplishments

1. FishSpeech on NIPA (GPU)
   • Deployed FastAPI service on NIPA.
   • Inbound ports blocked on provider; switched to a Cloudflare tunnel for public ingress.
   • Retired Cloud9 TTS path — NIPA is now the single TTS origin.

2. Connectivity (Local & Lambda)
   • Local PC → Cloudflare URL → NIPA works (health + synth).
   • AWS Lambda → Cloudflare → NIPA works; returns a playable URL.

3. Model Serving & Runtime
   • FishSpeech runs synchronously with one-time initialization; subsequent requests synthesize on-the-fly.
   • Verified stable request/response path and measured working latencies.

4. Chat → TTS → S3 E2E (local harness)
   • Local text → Cloud9 Chatbot reply → NIPA FishSpeech synthesis → upload to **my** S3 → presigned URL retrieval.
   • Saved outputs locally and tested multiple inputs; results are audible and correct.

2. 🔄 Ongoing / Next Steps
   • **Amazon Connect integration:**
   – Contact flow: Set `user_text` → Invoke Lambda → Play prompt (External URL via `$.audio_url`).
   • **S3 keying & sessioning:**
   – Adopt `sessions/{session_id}/{yyyyMMddHHmmss}_{uuid}.wav`; return `{session_id, key, answer, seed, latency_ms}`.
   • **Reproducibility:**
   – Fix `seed`, set `use_memory_cache=off`, log parameters; optional reference-voice pinning.
   • **Ingress plan (post-Cloudflare):**
   – Either named Cloudflare Tunnel (stable hostname) or open port 80 with Nginx → 127.0.0.1:8000 once provider security allows.
   • **Operational hardening (as time permits):**
   – Run uvicorn & cloudflared via `systemd`; add basic Bearer auth; increase presign TTL (≥600s) for Connect; S3 lifecycle for `sessions/*`.
   – (Telephony) Add 8 kHz mono PCM downsampling before upload if needed by your Connect prompts.

— — —

한국어 🇰🇷

1. ✅ 주요 작업 요약

1. NIPA(GPU) 환경에 FishSpeech 배포
   • FastAPI 서비스 구동 완료.
   • 인바운드 포트 제한으로 **Cloudflare 터널**로 임시 공개 접속 구성.
   • Cloud9 TTS 경로는 중단하고, NIPA 를 단일 TTS 기원으로 전환.

2. 연결 검증(로컬 & Lambda)
   • 로컬 PC → Cloudflare URL → NIPA(헬스/합성) 정상.
   • AWS Lambda → Cloudflare → NIPA 정상, 재생 가능한 URL 반환.

3. 모델 서빙/런타임
   • 1 회 초기화 후 동기 추론으로 온디맨드 합성 가능.
   • 요청/응답 플로우 안정성 및 지연 시간 확인.

4. Chat → TTS → S3 종단 테스트(로컬 하네스)
   • 로컬 텍스트 → Cloud9 챗봇 응답 → NIPA FishSpeech 합성 → **내 S3** 업로드 → 프리사인 URL 획득.
   • 로컬에 파일 저장 및 다양한 문장으로 청취 검증 완료.

2. 🔄 향후 진행 계획
   • **Amazon Connect 연동:**
   – 컨택트 플로우에서 `user_text` 설정 → Lambda 호출 → 외부 프롬프트(`$.audio_url`) 재생.
   • **S3 키/세션 구조 정리:**
   – `sessions/{session_id}/{yyyyMMddHHmmss}_{uuid}.wav` 채택, `{session_id, key, answer, seed, latency_ms}` 반환.
   • **재현성 확보:**
   – `seed` 고정, `use_memory_cache=off`, 파라미터 로깅; 필요 시 참조 음성 고정.
   • **Ingress 안정화(Cloudflare 이후):**
   – 네임드 Cloudflare 터널(고정 호스트네임) 또는 보안 설정 오픈 후 80 포트 + Nginx 역프록시.
   • **운영 보강(여유 시):**
   – uvicorn/ cloudflared `systemd` 등록, Bearer 인증, 프리사인 TTL(≥600s) 상향, `sessions/*` 라이프사이클 정책.
   – (텔레포니) 필요 시 업로드 전 8 kHz mono PCM 다운샘플링 추가.