


Date: 2025 년 8 월 27 일 

Prepared by: Tiong-Sik Ng 

English 

1. Summary of Accomplishments


- **Warmer starts:** Added a call-start warmup so the very first call after a change plays audio smoothly.
- **Barge-in enabled:** Callers can speak while audio is playing; the bot stops and handles the new request right away.
- **Faster first audio:** First clip now plays directly from S3, reducing “silent first play” cases.
- **Polish & stability:** Small flow cleanups, brief retries, and clearer docs—overall smoother demo behavior.

2. Issues & Risks

- If S3 access is restricted, first clip playback can fail—keep the current bucket access or presigned URL path.
- Removing the warmup could bring back slower first calls after deployments.
- Extremely rapid interruptions may clip a first word; observed rarely.

3. Next Steps

- Add lightweight metrics: time-to-first-audio, number of barge-ins, error counts.
- (Optional) Keep 1–2 provisioned instances for the demo window to avoid cold starts entirely.

한국어 

1. 주요 성과

- **초기 지연 개선:** 통화 시작 시 워밍업을 추가해 변경/배포 직후 첫 통화도 매끄럽게 재생됩니다.
- **바지-인 활성화:** 재생 중에도 고객 발화를 즉시 인식하여 기존 오디오를 멈추고 새 요청을 처리합니다.
- **첫 오디오 속도 향상:** 첫 클립을 S3 에서 바로 재생하여 “처음에 아무 소리 안 남” 상황을 줄였습니다.
- **안정화 작업:** 플로우 정리, 짧은 재시도, 문서 보완 등으로 전체 동작이 더 부드러워졌습니다.

2. 문제/리스크

- S3 접근이 제한되면 첫 클립 재생이 실패할 수 있어, 현재 접근 설정 또는 프리사인드 URL 방식을 유지해야 합니다.
- 워밍업을 제거하면 배포 직후 첫 통화가 다시 느려질 수 있습니다.
- 아주 빠른 중간 발화의 경우 첫 단어가 짧게 잘릴 가능성이 드물게 있습니다.

3. 다음 단계

- 간단한 지표 추가: 첫 오디오까지 시간, 바지-인 횟수, 오류 건수.
- (선택) 데모 시간대에 프로비저닝 인스턴스 1–2 개로 콜드 스타트 제거.