

Deep Learning Basic Study

▼ Week01

*파란색 문장은 Gemini의 추가 및 보완 사항

1. Conversation & Human Interaction 패턴

2. 여러 사람이 있는 공간에서 자연스럽게 대화에 참여하는 LLM 모델을 개발하고 싶음.

- 현재 AI 대화 시스템은 기본적으로 1:1 대화를 가정함. 직접적으로 폴더를 구분해주지 않으면 자신이 대화하는 상대를 구분하지 않고 구조상 자신이 누구와 대화를 나눌지도 결정하지 못함.
- **화자 인식**(speaker recognition)과 **자연언어이해**(Natural Language Identification)와 **자연언어생성**(text generation) 및 **음성합성**(Text to Speech)을 통합함으로써 자신이 누구와 대화하는지를 인지해서 어떤 대화 기록을 참조할지를 결정하고, 누구에게 몇 명에게 이야기할지를 적절하게 판단해 대화에 참여하는 시스템임.
 - 사람은 2명 이상이 한 공간에 있는 상황에서 어떤 사람과 몇 명과 대화하느냐에 따라서 자연스럽게 단어를 다르게 선택하고 문장을 다르게 구성하며, 목소리의 크기나 말투 등도 달라짐. Audience Design이라는 사회언어학 이론의 내용임.
 - AGI 또한 사회적 상황에서 적절하게 대화에 참여하고 답변을 생성할 수 있는 능력을 가져야 한다고 생각함.
 - 서로 다른 참여자와의 대화 속에서도 스타일과 논리 면에서 일관성을 유지해야 한다는 점에서 하나의 주체로서의 LLM 연구에도 도움이 될 것이라고 생각함.

3. 모델에 필요한 입력과 출력

- **모델의 입력**은 우선 당연히 **실시간 오디오 데이터**여야 할 것 같음. 직접 LLM 모델에 하는 이야기가 아니더라도 LLM이 배경 참여자로서 이뤄지고 있는 대화의 내용을 이해하고 있어야 할 것 같음.
- 또다른 입력으로는 참여자 명 수, 언어 정보 등이 메타 데이터로 들어가면 좋을 것 같은데, 현재 기술 수준으로도 실시간 오디오 데이터로 자체 생성과 분석이 가능한 함. 따라서 선택사항.

- **대화 기록 버퍼(Conversation History Buffer)**도 입력에 필요함. 최근 N분간 대화 요약과 화자별 메타 데이터(성향, 지식 수준) 등이 함께 입력되어야 상대에 따라 대화를 적절하게 생성할 수 있음.
- 모델의 출력을 중간 출력과 최종 출력이 있을 것 같음. **중간 출력은 화자 식별과 화자 인지 결과임.** 이미 화자 사전(speaker dictionary)에 등록된 참여자인지 확인하고 식별, 검증하거나 새롭게 등록을 해서 별도의 화자 시스템에 보낼 출력 결과를 생성함. **최종 출력은 TTS된 적절한 답변(대화)임.**
 - **또다른 중간 출력으로 발화 여부 판단**, 즉 침묵할 것인지 맞장구를 칠 것인지 답변 혹은 주도적인 질문을 할 것인지를 결정한 사항이 필요함. 이것을 제어 모듈에 전달해야 함.
 - **수신자 추정**도 해야 함. 상대방이 말하는 상대가 나(AI)인지 다른 대화 참여자인지 판단한 확률값도 발화 여부 판단에 필요함.
 - **최종 출력도 audience design**을 가미하려면 답변으로 생성한 텍스트와 함께 **스타일 토큰**(속삭이듯, 활기차게 등)과 **타이밍 제어 토큰**(바로 끼어들기, 말 완전히 끝나고 0.5초 뒤 말하기) 등이 필요함. (만약 음성-음성 End-to-End로 설계한다면 타이밍 제어 토큰만 필요할 것임)
 - 마지막으로, 실시간 대화 시스템에서 중요한 건 반응 속도이기 때문에 가벼운 모델이 필요하리라 생각함.
 - 또한 화자 식별과 화자 인지(화자 검증이 더 정확한 용어임)를 별도의 모델로 처리하지 않고 Open-set 화자 식별 파이프라인으로 해결할 수 있음.

참고자료 및 논문

- 화자 식별과 검증에는 Bredin et al. (2019), "Pyannote.audio: Neural Building Blocks for Speaker Diarization"과 Desplanques et al. (2020), "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification" 논문 참조하면 좋음. huggingface의 diarization 모델을 베이스라인으로 활용할 수도 있음.
- 대화 모델은 Rubenstein et al. (2023), "AudioPaLM: A Large Language Model That Can Speak and Listen" 참고하면 좋을 것 같음. 음성 입력을 텍스트로 변환하지 않고 뉘앙스와 운율 정보를 보존하는 모델이고, 내가 구상하는 최종 시스템과 가장 유사함.
- 한 개의 모델이 아니라 인지 모델 + 대화 모델이라는 이중 구조로 설계해서 인지 모델은 화자 식별과 검증, 음성 인식을 담당하고 대화 모델은 말하기 전략 수립, 스타일 설정, 화자 맞춤형 답변 생성을 하는 식으로 구상함. 입력 → 중간 출력(인지 모델

의 출력, 이때 화자 DB도 이용) = 중간 입력(대화 모델 입력) → 최종 출력 flow면
될 것 같음.