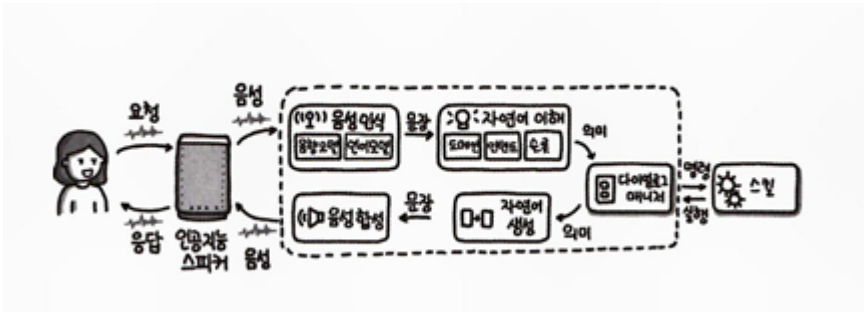
제 5장. 스마트 스피커

스마트 스피커는 어떻게 말을 알아들을까?

- 실제 사람의 말을 알아듣는 과정
    - 음성 녹음 → 서버 전송 → 서버에서 분석
- 사람에게 말을 하는 기능
    - 녹음된 음성을 서버에서 받아와 재생
- 스마트 스피커에는 웨이크업 단어만 알아 들을 수 있는 음성인식 엔진이 스피커에 내장되어 있음

분석 과정



이해 영역

- 음성인식, 자연어 이해

실행 영역

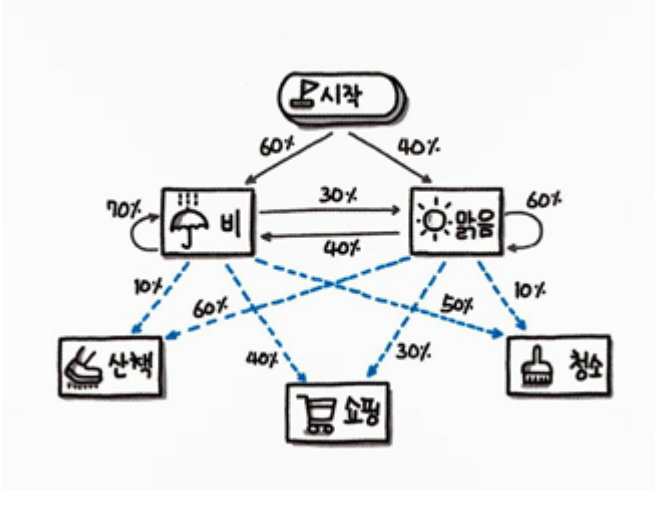
- 다이얼로그 매니저, 스킬

생성 영역

- 자연어 생성, 음성 합성

음성 인식

- 과거
    - 소리의 최소 단위인 음소가 결합하여 단어와 구문을 만드는 규칙을 언어학자들이 분석해 if-then 규칙으로 프로그래밍
    - 한계
        - 사람마다 음소를 다르게 발음, 음소의 패턴은 가까이 있는 음소에 영향을 받고, 생략도 많이 됨
- 은닉 마르코프 모델
    - 은닉된 상태와 관찰 가능한 결과로 구성된 통계적 모델
    - 예시)



- 은닉된 상태 : 날씨, 동거인은 매일 날씨 확인 및 그날 일정을 결정
  - 동거인의 행동을 보면 날씨를 확률로 예측할 수 있음, 그 결과가 위 모습의 형태
- 위 행동을 관찰하면 날씨를 예측할 수 있음
- 이러한 원리를 음성인식에 적용
  - 어떤 특정 파형의 결과가 '에이-비-시' 라면 결과는 'A-B-C'일 가능성이 높다고 예측하는 식
    - 은닉 마르코프 모델을 적용한 통계 기반은 규칙 기반보다 훨씬 더 좋은 성능을 보여줬음
- 딥러닝과 빅데이터
  - 현재는 딥러닝과 빅데이터를 이용해 뛰어난 음성인식 모델을 만들 수 있게 됨
    - 음향 모델

언어 모델, 오인식 단어 보정

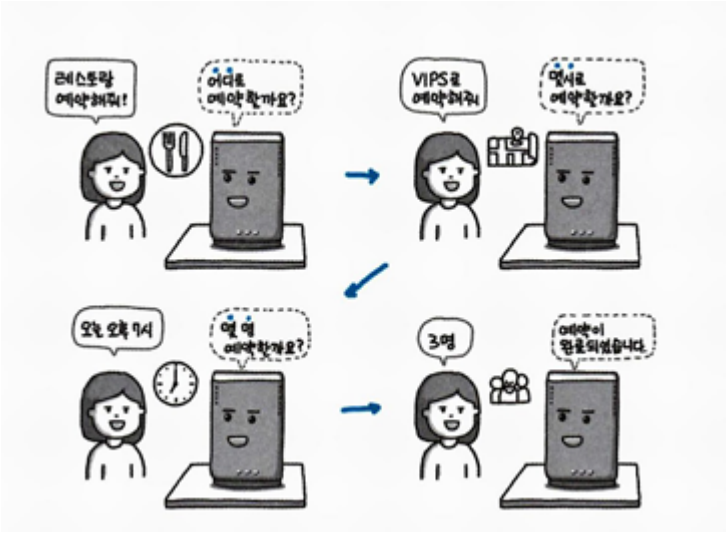
- 인간의 인지능력은 상식에 기반에 잘못 들은 단어도 보정해서 이해 함
  - 규칙 기반의 음성인식이 좋은 성과를 내지 못했던 이유가 바로 인간 처럼 보정할 수 있는 두뇌의 역할이 빠져있었기 때문
- 언어 모델은 관련 문장을 수 없이 많이 보았으며 학습했기 때문에 사전 지식이 있어서 사용자가 "오늘 날씨가 엿돼" 라는 말로 잘못 인식해도 "오늘 날씨 어때" 로 보정하여 알아들음

자연어 이해

- 스마트 스피커가 텍스트 문장을 감지하면 먼저 도메인을 구분 → 도메인 분류

발화	도메인	인텐트
"오늘 날씨 어때?"	날씨	조회
"최신 가요가 듣고 싶어."	음악	재생
"레스토랑 예약해 줘."	예약	진행

- 위 내용에서 중요한 부분이 빠졌는데 어느 지역의 날씨를 알려줘야 되는지에 대한 정보가 빠짐
  - 누락된 정보를 채워주는 과정을 슬롯 필링 이라고 함



- 위 그림처럼 추가로 질문을 해야되는 경우 대화를 이어나가야 함 → 멀티 턴이라 함

발화	도메인	인텐트	슬롯	슬롯 필링
"오늘 날씨 어때?"	날씨	조회	위치: 현재 위치	o
"최신 가요가 듣고싶어"	음악	재생	대상: 최신 가요	x
"레스토랑 예약해 줘"	예약	진행	장소: 빕스 시간: 오늘 오후 7시 인원: 3명	o

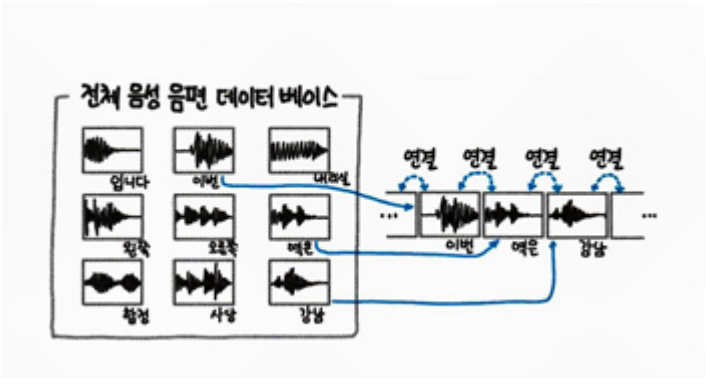
다이얼로그 매니저, 명령 실행

- 여러 도메인에 맞춰 적절한 행동을 수행하도록 명령을 내리는 역할

자연어 생성, 대화 디자인

- 스마트 스피커는 챗GPT 와는 조금 다름
  - 스마트 스피커는 문제 해결용 대화시스템임
  - 목적이 분명한 대화만을 주로 함
    - 자유로운 대화생성을 안함, 정해진 템플릿에 정보를 채워 문장을 생성하는 방법을 주로 사용
      - "현재 날씨 는 OO고, 온도는 OO도입니다."

연결 합성



- 스마트 스피커는 템플릿 기반으로 문장을 생성하기 때문에 성우가 녹음한 소리가 아직까지 품질이 가장 좋음

- 템플릿과 단어들을 따로 녹음해놓고 조합하면 여러 문장을 만들어낼 수 있는데 이를 연결 합성 혹은 USS (unit selection synthesis, 음편 선택 합성) 이라 함

정리

1. 사람이 질문하면 음성을 텍스트 문장으로 변환하고,
2. 문장을 이해한 다음에는 명 령을 생성 합니다.
3. 명령으로 스킬을 실행한 다음에는 다시 문장을 만들어내고,
4. 마지막으로 음성을 합성하여 문장을 소리내어 읽습니다.