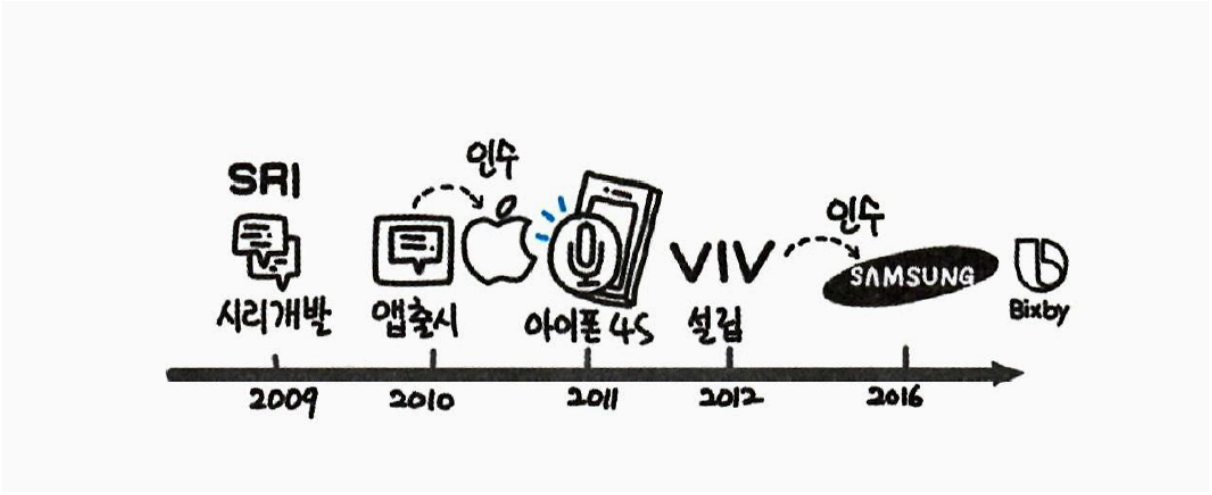
스마트 스피커

인공지능 비서의 탄생

항목	핵심 내용
세계 최초	2011 Apple Siri → 2014 Amazon Echo → 2016 Google Assistant 등
한국 시장 진입	2016 SKT NUGU → 2017 네이버 클로바 → 2017 카카오미니 (6개월 만에 출시)
카카오의 기술 토대	- 다음커뮤니케이션(고품질 검색·한글 NLP) - 다이알로이드(음성 인식) 인수 - 모듈형 에이전트 아키텍처 + 온톨로지
6개월 단축의 비결	핵심 음성·NLP 기술을 이미 보유해 '일사천리' 개발 속도
향후 과제	대화 품질 강화, 다중 언어 지원, 새로운 서비스 영역 진출

애플 시리, 음성인식 비서의 시대를 열다

- 음성 인식 도입이 Siri를 텍스트 챗봇에서 진정한 "음성 비서"로 바꾸었음.
- Apple 인수 직후 잡스 사망과 내부 갈등이 Siri의 발전을 지연시킴.
- 핵심 인력 이탈 → Viv Labs가 독자 AI 비서를 개발하고, 삼성 인수로 Bixby로 이어짐.



아마존 알렉사 스마트 스피커의 시대를 열다

아마존이 2014년에 첫 스마트 스피커를 공개 → Echo.

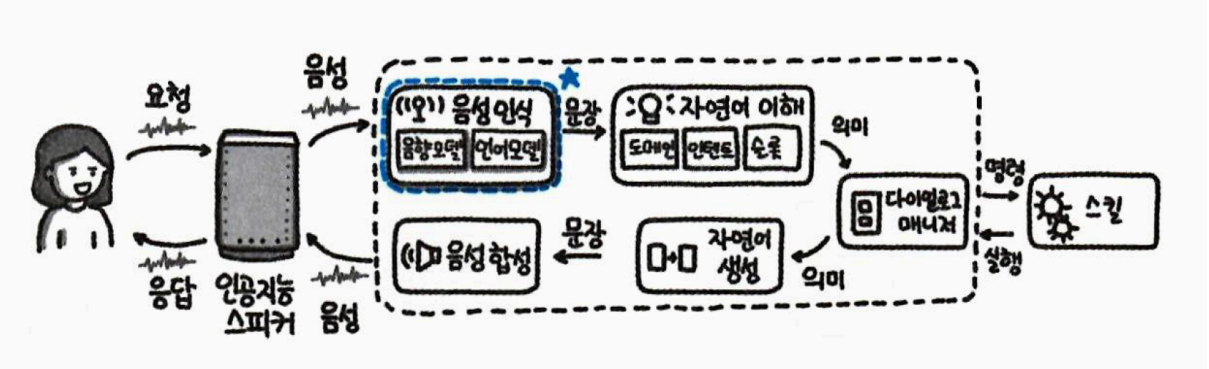
호출어를 "Alexa"로 바꾸어 고대 이집트 도서관에 대한 오마주를 실현.
"Alexa"가 플랫폼 브랜드로 성장 → 미국에서 아이 이름으로 사용이 급감.
Echo는 전 세계적으로 1억 대를 판매하며 스마트 스피커 시장을 선도.

스마트 스피커는 어떻게 말을 알아들을까?

스피커 자체는 "블루투스 스피커 + 마이크"라기 때문에 음성 인식·대화 기능을 클라우드에 맡김

웨이크-업 전까지는 "키워드 인식 모델"만을 실행하고, 나머지는 클라우드에서 처리

사람의 목소리를 알아듣는 음성인식 과정



단계	역할
음성인식(ASR)	음파 → 음성 특징 추출 → 단어/문장 인식
자연어 이해(NLU)	인식된 텍스트 → 의도·개체 추출
다이얼로그 매니저	대화 흐름 관리, 상태 추적
스킬(서비스) 호출	외부 API/서비스 활용(날씨, 음악, IoT 등)
자연어 생성(NLG)	사용자에게 전달할 답변 문장 생성
음성 합성(TTS)	생성된 문장을 음성 신호로 변환
스피커로 반환	재생용 음성 데이터 전달 → 스피커에서 재생

음성인식의 어려움(책에서 설명한 내용)

- 사람의 발음은 문자를 읽는 것보다 변동성이 많고, 음성은 시계열 파형으로 불규칙적이다.
- 음소(phoneme) 규칙 기반은 사람마다 발음이 달라 생략·변형이 빈번해 실패.
- 단어 사이에 명확한 공백이 없고, 억양·속도·음량이 의미를 바꿀 수 있다.
- 인간도 소음이 있는 상황에서는 이해가 힘들고, 컴퓨터에게는 더욱 어려움.
- 과거 규칙 기반(언어학적 if-then) 접근은 한계가 뚜렷해졌고, 현재는 대규모 데이터와 딥러닝 모델을 활용해 음성 패턴을 학습한다.

음향 모델, 음성의 파형에서 단어를 인식하다

음성인식률을 높이기위한 방법들을 고안

첫째로

음성의 파형을 규칙으로 구분하려고 시도 → 쉽지않음

두번째로 (40~15% 오류인식률)

은닉된 상태와 관찰 가능한 결과로 구성된 통계적 모델

ex) 은닉 마르코프 모델

```
지난 5일간 한 행동을 관찰해봤습니다.
동거인이 '청소-청소-쇼핑 -산책-청소'를 했다고 한다면,
지난 5일간의 날씨 는 과연 어 땠을까요?
```

```
확률 모델을 이용해 은닉된 상태인 날씨의 확률 을 알아내는 것을
디코딩이라고 합니다. 디코딩 과정을 거쳐 가장 가능성이 높은 날씨를
확률적으로 추정하여 확률로 나타냄
```

세번째로

딥러닝 도입 + 순환신경망 고안

순환신경망(RNN)은 '이전 시점의 은닉 상태'를 현재 입력과 함께 사용해 다음 은닉 상태를 만들고, 이를 통해 시간-연속 데이터(음성, 텍스트)의 문맥을 '기억'하고 예측하는 딥러닝 모델이다.

언어 모델, 오인식 단어를 보정하다

언어 모델은 과거에 본 문장 통계(확률)를 바탕으로 음성 인식 결과를 보정한다.

ex) 기계는 텍스트(뉴스, 책 등)로부터 단어/문장 빈도를 학습해, 흔히 쓰이는 단어·구조에 높은 확률을 부여.

자연어 이해, 언어를 이해하다

문장이 어떤 의미를 가지는가 이해가 필요

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

자연어 이해 과정

그렇다면 자연어이해로 받은 문장을 실행하는 역활은?

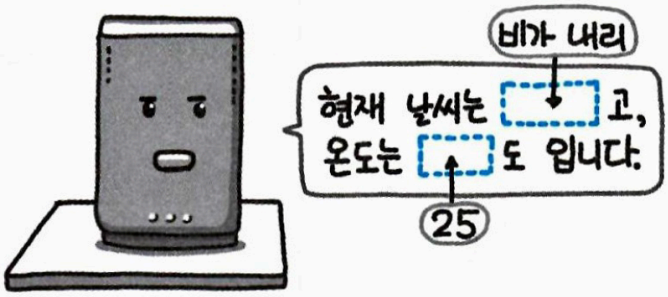
다이얼로그 매니저, 명령을 실행하다

음악을 틀 수 있는 서비스, 예를 들어
멜론 같은 서비스에 접속해 최신가요 목록을 찾아서 음악을 재생하
도록 명령을 내리는 역할

이렇게 음악을 재생한 후에는 결과를 통보

ex) 참고로 멜론 같은 여 러 서비스를 이용하는 기능을 스킬Skill

자연어 생성, 대화를 디자인하다



"현재 날씨는 비가 내리고, 온도는 25도입니다."

좋습니다. 이제 생성된 문장을 스피커가 읽기만 하면 됩니다. 물론 단순한 템플릿을 쓰면 매번 똑같은 방식으로 답변하기 때문에 식상할 수 있으니 더욱 풍부한 대화를 위해서 템플릿을 다양하게 구성합니다. 여러 개의 템플릿으로 번갈아가며 대답한다면 훨씬 더 사람처럼 다양하게 대답할 수 있겠죠.

연결 합성, 문장을 자연스럽게 읽을 수 있을까?

1. 문장 생성
   - 제한된 템플릿에 slot 값 삽입 → 빠르고 품질 보장.
2. TTS
   - Unit-Selection: 녹음된 음편(Unit Library)에서 가장 적합한 단위 선택 → 신호 수준 합성.
   - Neural TTS(딥러닝): 직접 음성 파형을 예측 → 더욱 자연스러운 억양, 감정 표현 가능.

3. 최종 출력
    - 생성된 음성 파형은 스마트 스피커 하드웨어(스피커 모듈)를 통해 재생 → 사용자가 들을 수 있는 “음성”이 된다.

음성 합성, 인간보다 더 자연스러움을 향해

현재 딥러닝 기반 음성 합성은 규칙/통계·보코더→엔드-투-엔드 네트워크→딥-보코더 순으로 단순화되고 있으며, 음질은 거의 사람 수준에 근접했지만 여전히 억양·감정·속도-미세조정에 한계가 있다.향후는 전 단계*(*텍스트→멜-스펙트로그램*)* 없이 직접 음성을 생성하는 *True End-to-End* 모델과 멀티-스피커·감정-연동이 결합된 다중모달 *AI*가 상용화될 것으로 예상

스마트 스피커가 말하는 과정

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.