제5장 스마트 스피커

5.1 인공지능 비서의 탄생

- 스마트 스피커는 어느 날 갑자기 하늘에서 뚝 떨어진 기술이 아니고. 스마트 스피커 뒤에 숨어있는 핵심 콘셉트와 기술(에이전트 기반의 아키텍처, 자연어 이해, 온톨로지 등)은 수십 년 동안 연구소의 연구 주제 중 하나였고, 마침내 기술이 무르익어 카키오처럼 하나의 제품으로 빠르게 탄생되었다

5.2 애플 시리, 음성인식 비서의 시대를 열다

- 2011년 애플 아이폰에 탑재된 시리가 등장한 이후부터 본격적으로 음성 인식 비서의 시대가 열었다

시리는 처음에는 음성인식 기능이 없었다. 텍스트로 메시지를 입력하면 텍스트로 응답하는 제품이였는데

음성인식 기능을 도입하고 1년 뒤에 출시하게 된다 훗날 빅스비랑 같은 뿌리는 공유하는 셈이다

5.3 아마존 알렉사 스마트 스피커의 시대를 열다

- 첫 음성인식 비서는 애플의 시리였지만 스마트 스피커라는 카테고리를 처음 만든 회사는 아마존이다.

오늘날 에코는 전 세계에 1 억대가 넘게 팔린 베스트셀러이며, 미국 시장조사 기업 이마케터가 조사한 바에 따르면, 미국 가정에 있는 스마트 스피커의 약 70%가 에코라고 한다.

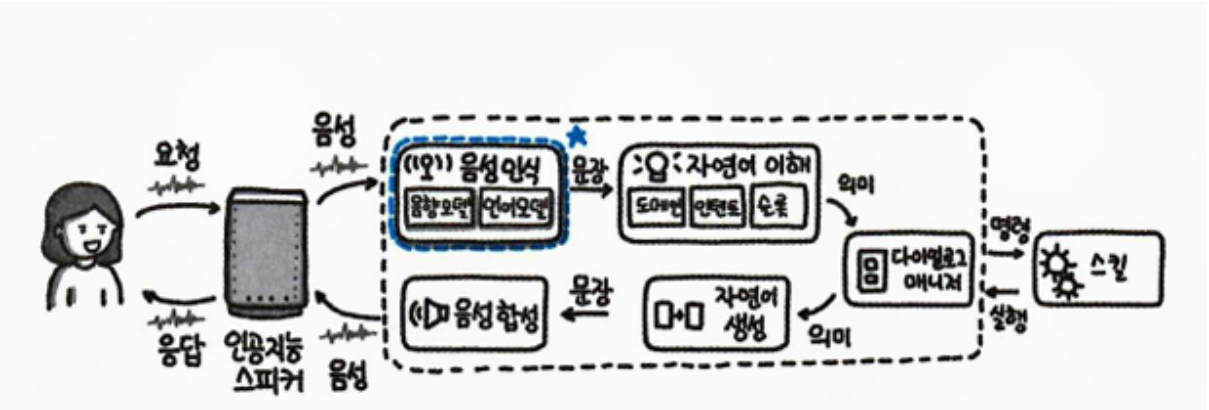
5.4 스마트 스피커는 어떻게 말을 알아들을까?

-실제로 사람의 말을 알아듣는 과정은 음성을 녹음하여 서버로 보내 분석하는 과정이고, 사람에게 말을 하는 기능은 녹음된 음성을 서버에서 받아와 재생하는 것입니다.

웨이크업(Wake-Up) ⇒ 헤이 카카오”라고 부르면 스피커가 “네?”하고 반응

웨이크업 이후에는 본격적으로 음성 파일을서버로 전송하여 분석을 진행, 이제부터 마이크에 녹음된 모든 음성을서버로 전송하여 분석 이제 여기서부터는 거대한 서버 시스템이 관여한다.

5.5 사람의 목소리를 알아듣는 음성인식 과정



-음성인식은 한마디로 시간의 흐름에 따라 역동적으로 변동하는 음성의 파형을 다루는 일이다

과거에는 음성의 파형에서 음소를 인식한 다음, 음소의 고유한 배열을 기반으로 단어를 인식했다

이 방식은 철저하게 규칙에 따라 음성을 인식했지만 곧 한계를 드러낸다. 사람들은 제각각 음소를 다르게

발음하고, 음소의 패턴은 가까이 있는 음소에 영향을 받으며, 생략되는 음소도 많다. 심지어 같은 사람이라도 발음하는 방식이 늘 일정하지 않다

초기에는 단어 사이에 명확한 공백이 존재할 거라 예상했지만 실제로는 그렇지 않다는 점도 문제를 복잡하게 만들었다.

5.6 음향 모델, 음성의 파형에서 단어를 인식하다

- 1970년대 중반부터 은닉 마르코프 모델 이라는 방법을 응용하기 시작

은닉 마르코프 모델은 말 그대로 은닉된 상태와 관찰 가능한 결과로 구성된 통계적 모델

은닉 마르코프 모델을 적용한 통계 기반은 규칙 기반보다 훨씬 더 좋은 성능을 보였다.

하지만 음성의 15~40%이상을 잘못 인식하였다.

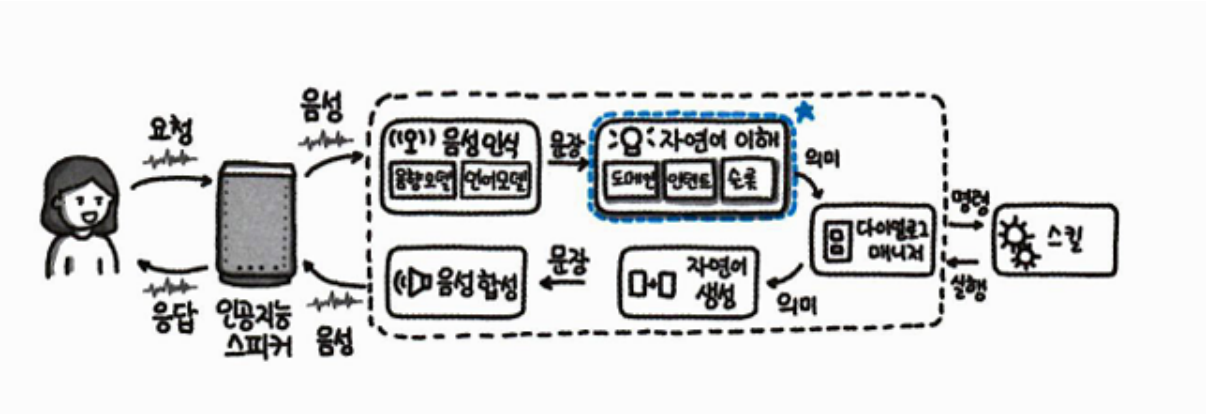
하지만 딥러닝이 좋은 성과를 내기시작하면서 음향 모델의 성능도 급격히 좋아졌다

딥러닝과 빅데이터를 이용해 성능이 뛰어난음성인식 모델을 만들 수 있게 됐는데 이를 음성의 파형으로 단어를 인식하는 음향모델 이라고 한다.

5.7 언어 모델, 오인식 단어를 보정하다

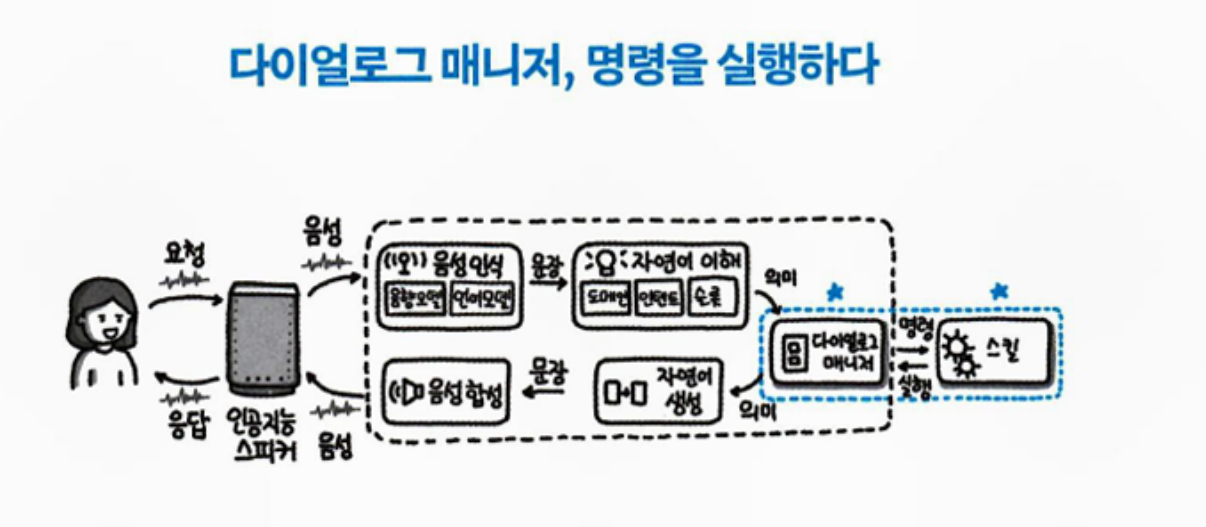
-규칙 기반의 음성인식이 좋은 성과를 내지 못했던 이유는 오인식을 보정할 수 있는 두뇌의 역할이 결여되어
있었기 때문이다. 만약, 음성을 잘못 인식하더 라도 그동안의 학습 결과를 토대로 사용할 확률이 높은 단어로 보정해 준다면 훨씬 더 좋은 성과를 낼 수 있고 이것이 바로 언어 모델 Language Model의 역할이다

5.8 자연어 이해, 언어를 이해하다



-기계는 문장의 의미를 파악해야 하는데. 기계가 이러한 난관을 뚫고 과정을 처리하는 것을 자연어 이
해(Natural Language Understanding, NLU)라고 한다. 말이나 글의 의미가 무엇인지 알 수 있도록 언어를 구조화하는 것

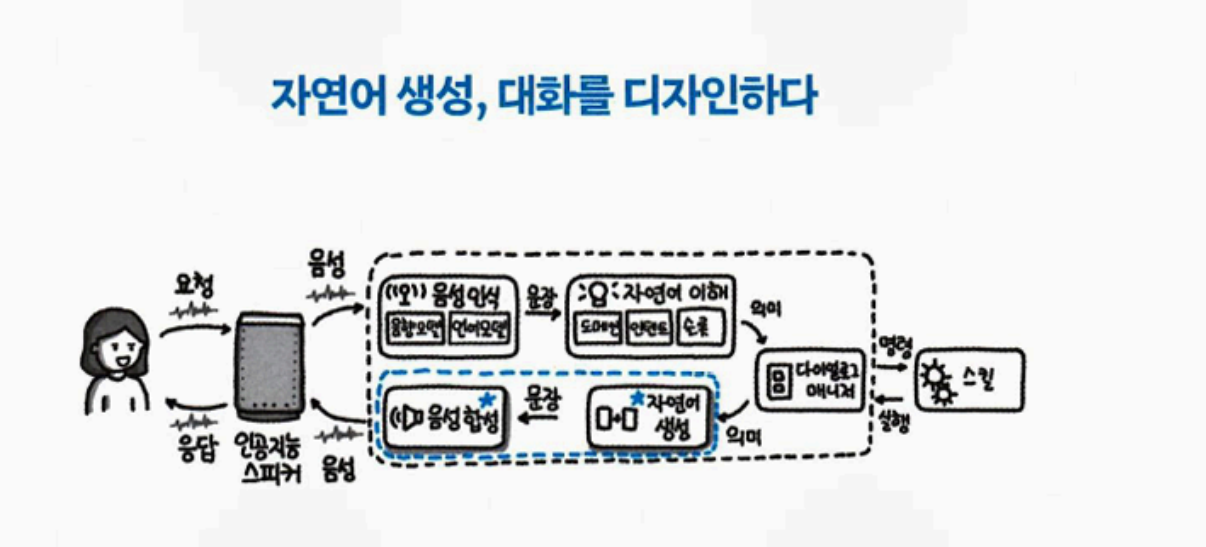
5.9 다이얼로그 매니저, 명령을 실행하다



-다이얼로그 매니저의 역할은 하나에 국한되지 않는다. 앞서 음성 인식이나 자연어 이해가 자신의 역할에 맞는 특정 기능만 수행했다면, 다이얼로그 매니저는 사실상 그 이외의 모든 작업을 수행한다.

스마트 스피커에서 스킬은 스스로 작동하지 않는다. 반드시 다이얼로그 매니저가 개입하고, 판단하여 스킬에 명령을내린다. 다이얼로그매니저가'음악 재생'이나 '날씨 조회' 같은 명령을 스킬에 내리면 스킬에 등록된 서비스가 실행되어 결과를 받아오는 구조다

5.10 자연어 생성, 대화를 디자인하다



-이제 최종 결과를 사용자에게 알려주는 일만 남았다

'이해' 영역(음성인식, 자연어 이해)에서는 딥러닝이 좋은 성능을 보였고, 그렇다면 자연어 생성도 딥러닝을 적극 도입하여 통계적인 방법을 거쳐 기계가 자유롭게 문장을 생성하도록 두면 될까?

만약 잘못 생성한 문장이 "인간은 모두 죽어야 해!"라는 문장이라면 이런 문제 때문에 '생성' 영역에서는 아직까지 딥러닝의 활용이 조심스럽다.

5.11 연결 합성, 문장을 자연스럽게 읽을 수 있을까?

-그렇다면 스마트 스피커는 어떤 과정을 거쳐 문장을 읽을까

녹음한 소리를 조합하면 하나의 자연스러운 문장을 만들어낼 수 있는데 이러한 방식을 연결 합성 또는 USS(UnitSelection Synthesis(음편 선 택 합성) 라고 한다, 미리 녹음된 음성을 기준에 따라 잘게 쪼개어 음편Unit을 만들고 가장 적합한 음편을 선택Selection 하여 음성을 합성Synthesis하는 방식을 말한다.

원음을 그대로 사용하므로 음질이 매우 자연스럽다는 장점이 있고. 연결 합성 기술은 내비게이션에도 사용되며, 이처럼 항상 일정한 답변을 하는 경우에 무척 유용하다

5.12 음성 합성, 인간보다 더 자연스러움을 향해

-음성 합성 분야는 딥러닝이 가장 빠르게 발전하는 분야다.

최근에는 문장 전체를 딥러닝으로 합성하려는 시도도 많이 하고 있다

구글이 제안하고 엔비디아에서 구현한 음성 합성 모델 타코트론2(Tacotron2)는 사람과 거의 구분할 수
없을 정도로 자연스러운 음성을 합성해낸다

음성을 합성하는 과정은 크게 두 단계로 요약할 수 있는데

1 . 텍스트 ―> 멜 스펙트로그램
2 . 멜 스펙트로그램 ―음성

타코트론2의 역할은 이 중 첫 번째 단계인 텍스트를 멜 스펙트로그램(Mel Spectrogram)으로 만드는 과정을 담당한다.

멜 스펙트로그램이란 소리나 파동을 시각화하여 파악할 수 있도록 표현한 것으로, 음파와 비슷하게 생겼지 만 색상의 차이 , 농도를 포함해 더욱 풍부한 정보를 표현할 수 있으며, 이를 인간이 인지할 수 있는 주파수 대역으
로 변환해 낮은 해상도로 압축한 것을 말한다

다음으로 멜 스펙트로그램을실제 음성으로 바꾸는 작업이 필요한데 이 단계를 처리하는 기술을 보코더(Vocoder)라고 하며, 얼마나 노이즈 없이 깨끗하고 선명한 음질을 생성할 수 있는지가 이 기술의 핵심이다

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