멀티미디어시스템 REPORT

음성인식 AI 스피커

2020 년 06 월 26 일

제출: 32191105 김지민

1. 과제 개요

1) 기술 소개

인공지능 스피커는 음성인식 기능을 통해 음악 감상, 배달 서비스, 정보 검색, 기기 관리(ex TV 끄기) 등 다양한 기능을 수행하는 음성비서이다. 음성인식 스피커 기기의 주요 입출력 장치는 마이크와 스피커이며, 이 스피커 기술이 성공하면서 세계는 원래의 터치 방식에서 음성 기반 플랫폼으로 이동하고 있다. 인공지능 음성인식 스피커 서비스는 명령 수행 능력이 좋다. 손을 이용한 타이핑은 1분에 40단어를 기록하지만, 음성인식은 같은 시간에 150단어 기록이 가능하다. 또한 직접 손을 이용하여 기기를 조작하는 것이 아니기 때문에 작동법이 쉽다. 이러한 사람과 대화하듯 말하며 서비스를 제공 받을 수 있는 우수한 능력 덕분에 사람들이 점점 더 필요로 하고 있으며 기술이 발전하고 있다.

2) 기존 기술의 현황, 문제점 및 개선 방안, 기대효과

음성인식 인공지능 스피커 기술을 이용해 다양한 기업들이 제품을 출시하고 있는데, 그 중 대표적인 것이 SK의 '누구(Nugu)', KT의 '기가지니' 애플의 '시리(Siri)' 등 이다. 여러 기업이 출시한 음성인식 AI 기술은 인식률, 맥락 이해 향상 기술과, 더 나아가 대화 구현 단계에 들어섰다. 이 기술은 개별 각각의 숫자, 음절 인식에서 고립된 단어 인식, 연결된 단어 인식, 대어휘 연속 음성인식 등 점점 높은 단계로 진화하면서 빠르게 발전하고 있다. 대어휘 연속 음성인식 기술에서 인체가 신경세포 신호를 전달하는 방식과 비슷한 딥러닝 기술이 음석인식 전반에 걸쳐 적용되며 인식률이 급속히 개선되었다. 일반 상황이나 소리내어 말을 하는 상황에 대한 인식의 정확도는 90% 이상의 수준으로 발전했다. 이처럼 많은 글로벌 IT기업들이 인공지능 기술을 발전시키고 있고 스피커, 스마트 폰 등에 음성인식 서비스를 탑재하여 거대한 데이터를 수집하고 있다.

음석인식 AI가 많은 발전을 하고 있지만 문제점 또한 적지 않다. 우선 수많은 양의 음성, 문자 데이터 확보가 필요하다. 이 기술은 딥러닝을 기반으로 빅데이터를 처리하기 때문에 다양한 사용자와 환경 등에서 수집된 데이터들이 필요하다. 또한 여러 환경에서 잡음, 교차대화 등이 있을 때 인식하는 것이 어려운 문제점이 있다. 말하는 이의 성별, 나이, 사투리, 지역 등이 반영된 다양하고 거대한 음성 데이터베이스의 수집, 구축이 필요하다.

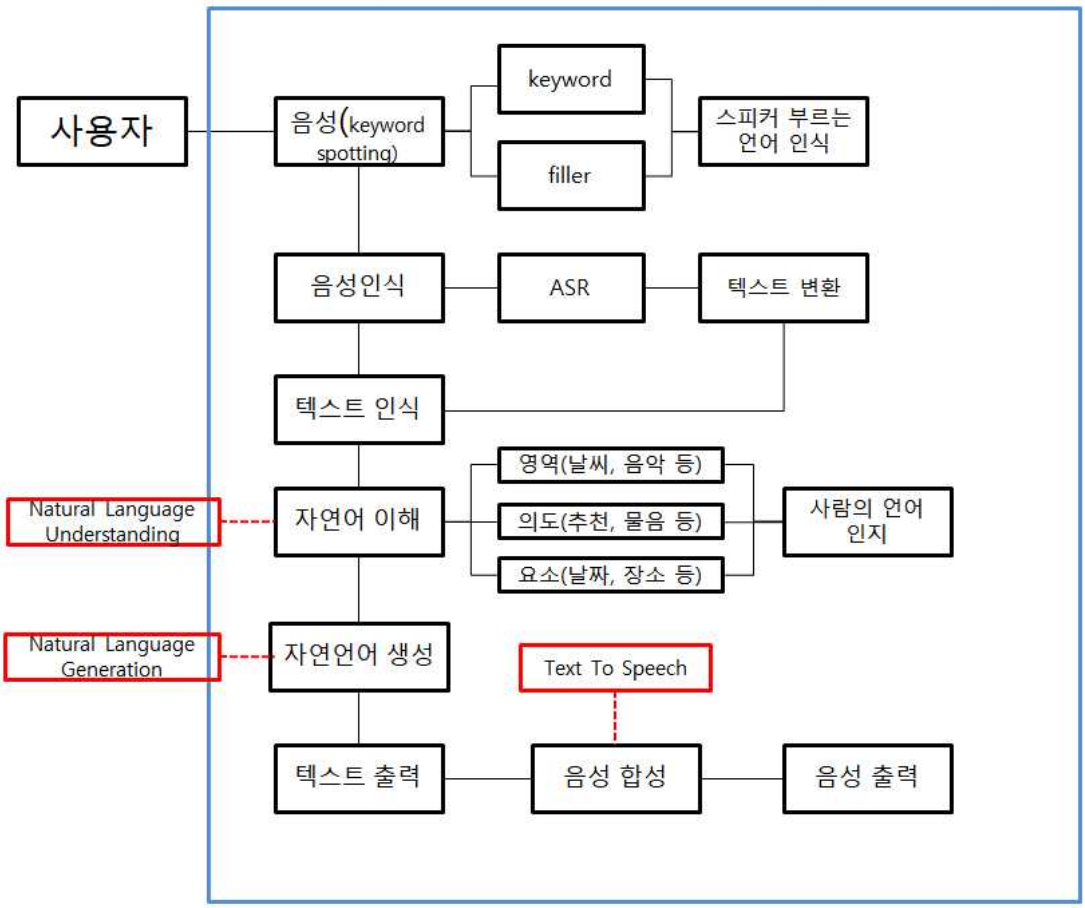
궁극적인 음석인식 AI의 목표는 사람과의 자연스러운 소통이다. 위의 문제점들을 모두 해결해 나가면 이 기술은 사람과 같은 방법으로 대화를 이해하고, 피드백하며, 거기에 추가 지식을 포함하는 등 기기와 사람이 소통할 때 가장 편리한 인터페이스로 발전할 것이다. 또한 세밀한 개인화, 다국어 음성인식, 보안 등의 해결도 기대되고 있다.

2. 과제 연구내용

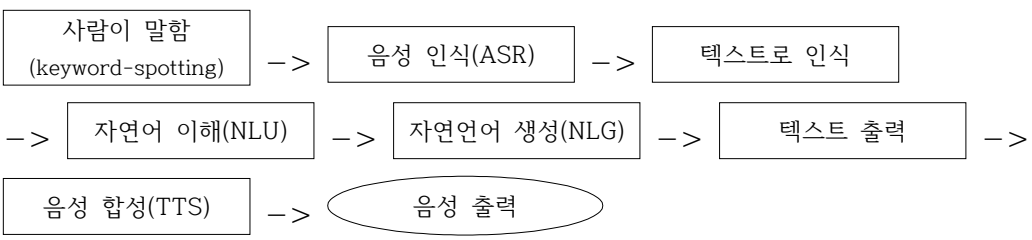
1) 기술 세부 소개

(1) 음성인식~음성출력 기술

NUI(Natural User Interface) 지능을 이용하여 인간처럼 음성을 인식하고 출력할 수 있도록 한다. 인식부터 출력까지의 블록다이어그램은 다음과 같다.



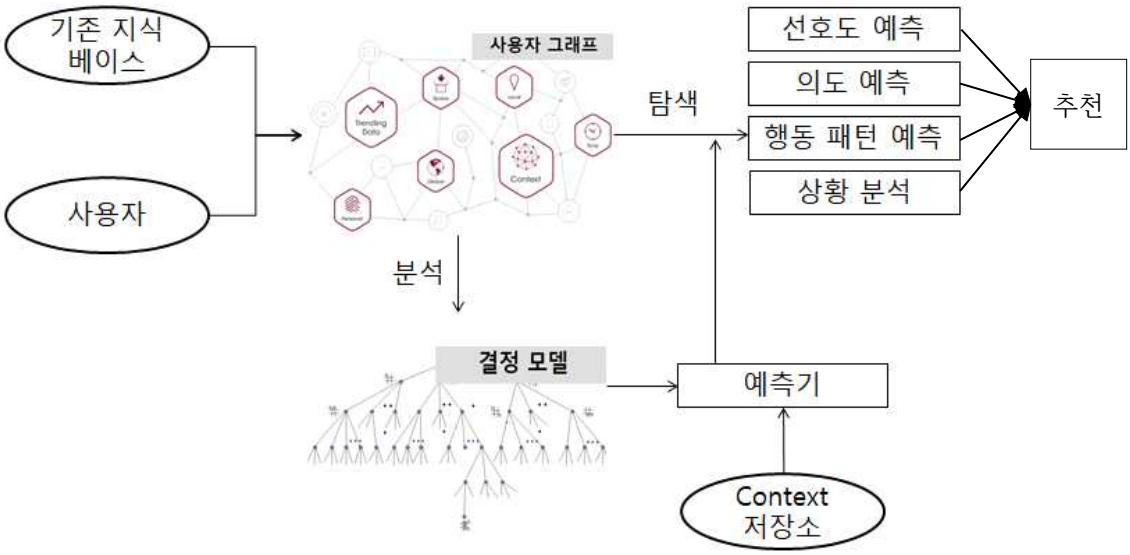
-알고리즘



①keyword spotting을 통해 스피커를 부르는 소리인지 아닌지 인식한다.
ex)아리아~, 기가지니!
②부르는 소리를 인식하면 음성인식을 시작하고 사람이 말하는 음성신호를 텍스트로 변환하여 인식한다. (ASR)
③인식한 텍스트를 domain(영역), intent(의도), entity(요소) 로 나누어 이해한다. (NLU)
ex)오늘 서울 날씨 알려줘
domain : 날씨
intent : 물음
entity : 날짜-오늘, 위치-서울
④이해한 자연어에 대한 응답을 자연어로 생성하여(NLG) 텍스트로 출력한다.
⑤텍스트를 음성으로 합성한다. (TTS)
⑥합성한 음성을 출력한다.
ex)오늘 서울은 더운 날씨가 계속되며, 최고기온은...

(2) 개인화/추천 기술

음성인식 스피커는 사용자의 선호도, 상황을 예측,분석하고 그에 따른 정보 제공한다. 시스템 구조는 다음과 같다.

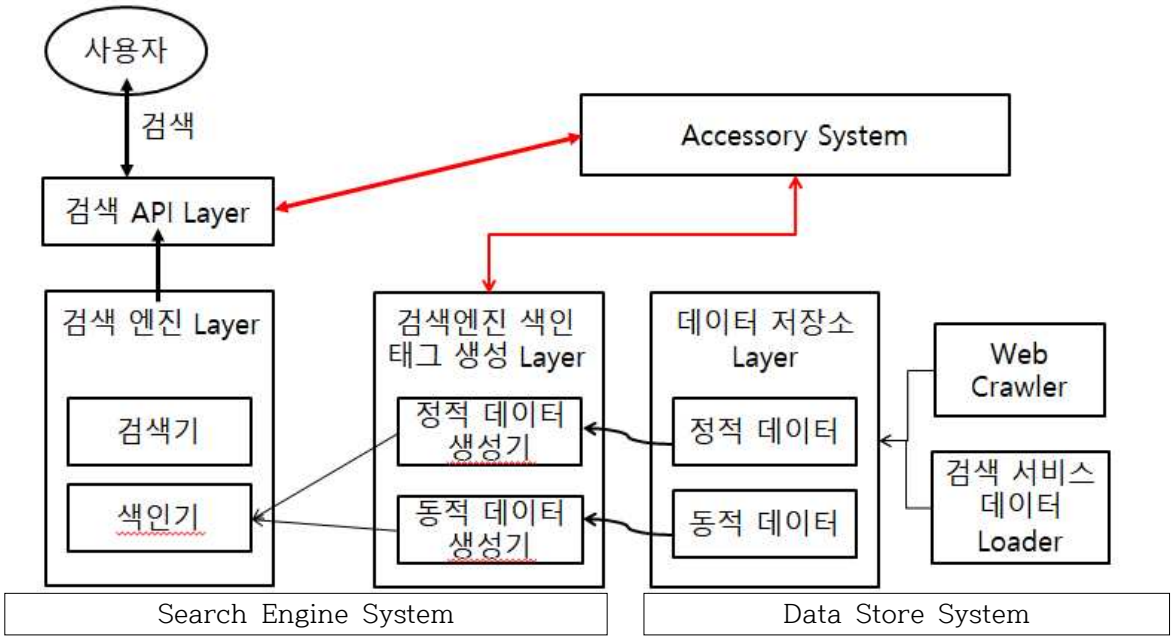


①기존 지식베이스와 사용자의 데이터를 바탕으로 사용자 그래프 데이터를 구축한다.
②사용자 그래프를 분석하고 탐색한다.
ex)사용자가 평소에 듣는 음악 장르 : 힙합

③분석하여 생성한 결정모델과 context 저장소를 통해 사용자의 선호도, 의도, 행동 패턴, 상황 등을 예측하고 분석한다.
ex)사용자는 힙합 같은 신나는 음악을 좋아함
④예측, 분석한 데이터를 바탕으로 사용자에게 맞춤 추천을 한다.
ex)신나는 음악 추천

(3) 정보 검색 기술

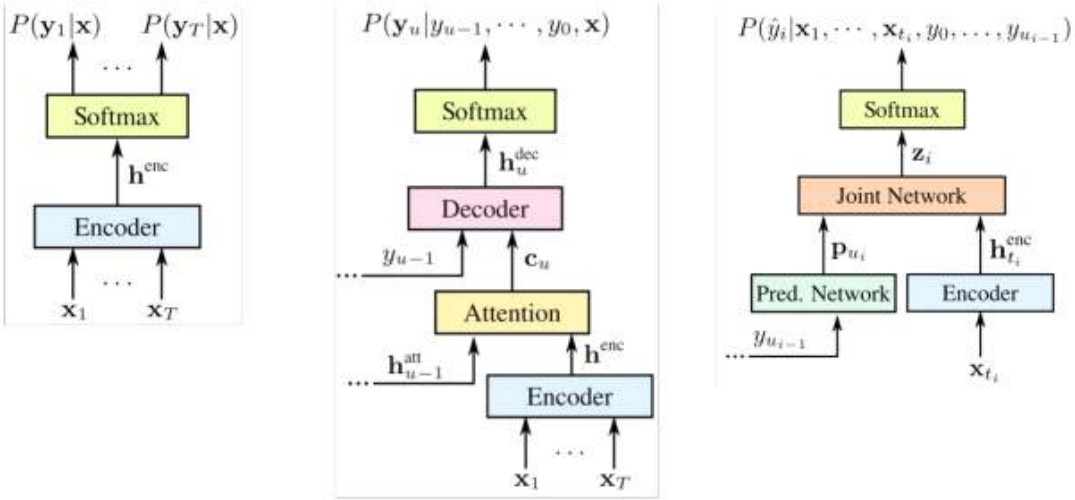
정보 검색 기술은 사용자가 요청한 정보를 알맞게 검색해주는 기술로 시스템 구조는 다음과 같다.



①사용자가 검색을 요청하면 검색 API를 이용해 검색을 시작한다.
②Web Crawler나 검색 서비스 데이터 Loader 등을 통해 데이터 저장소에 데이터를 저장한다.
③데이터 저장소에서 저장한 정적데이터와 동적데이터를 색인 태그 생성 layer에 전달한다.
④전달한 데이터들을 검색 엔진 Layer에서 색인(중요키워드 추출, 저장)하여 검색 API layer에 전달한다.
⑤사용자가 요청한 검색결과를 반환한다.

(4) 음성인식 기술

음성인식 기술은 음향모델, 단어발음사전, 언어모델을 사용한다. 음향적 학습 부분을 담당하는 음향모델은 소리를 기호로 전환하는 역할을 맡는다. 소리 정보에는 많은 잡음 환경과 음향 데이터가 수집, 포함되어 있으므로 방대한 데이터가 이 모델에 수집될수록 올바른 기호정보로 바꿀 수 있다. 단어발음사전은 단어의 발음들을 기록해 놓은 것이다. 언어모델은 단어 간 관계를 확률 관계를 통해 모델링한 것이다. 이 모델에 의해 표현되지 않은 단어들의 관계에 관해서도 성능을 향상시키기 위해 RNN을 이용한 Rescoring방법이 많이 이용된다. 효율적인 모델을 만들기 위해 이 세 가지의 전체 정보들을 모아놓고 중복을 제거하는 방법인 wFST방법을 대부분 사용한다. 최근에는 음향모델, 발음사전, 언어모델 등을 통합하여 하나의 딥러닝 네트워크로 표현하는 E2E(End to End) 기법들이 많이 쓰인다. E2E 중 RNN-T, LAS 방법이 가장 주목 받고 있다. 아래는 E2E음성인식기의 딥러닝 구조를 비교 설명한 그림이다.



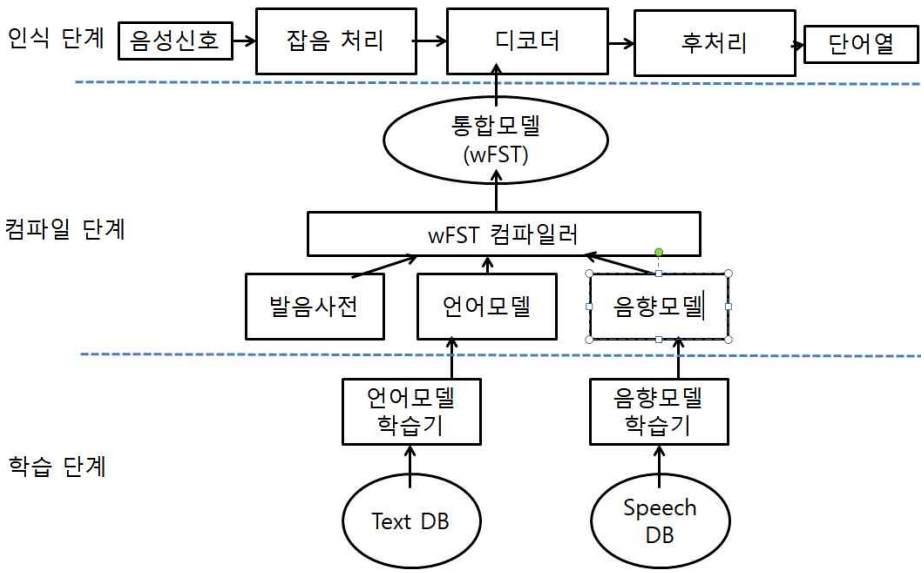
Connectionist Temporal Classification (CTC) [Graves et al., 2006]

Listen Attend Spell (LAS) [Chan et al., 2015]

RNN-Transducer (RNN-T) [Graves et al., 2012, 2013]

(참고 : AI 스피커 주요 기술 및 발전 방향-김영준)

음성인식기술 구조는 크게 학습단계, 컴파일단계, 인식단계로 구분된다. 구조도는 다음과 같다.



①언어 데이터베이스와 음향 데이터베이스를 통해 언어, 음향을 학습한다.
②학습된 언어모델, 음향모델과 발음사전의 중복을 제거하는 wFST 컴파일러 단계를 거친다.
③컴파일러를 거친 통합모델을 이용하여 인식하는 단계로 넘어간다.
④인식 단계에서 음성신호의 잡음을 처리, 디코더, 후처리 단계를 거쳐 단어열을 생성한다.

2) 응용 사례

(1)AI스피커 기반의 스마트홈 서비스

AI스피커 기반의 스마트홈 서비스는 사용자의 음성 명령을 기반으로 인공지능 플랫폼을 통해 분석되고 사물인터넷 플랫폼, 클라우드에 명령을 전달하며 그에 따라 집안의 IOT 기기를 작동시키는 시스템이다. 기존의 스마트홈 플랫폼인 개별적인 IOT 기기를 조작하는 방식과는 달리 AI스피커 서비스를 중심으로 많은 사물인터넷 서비스가 인공지능 서비스와 상호작용하여 하나의 플랫폼을 구성한다. 다수의 전자기기에 사물인터넷 기술이 결합됨에 따라 AI 스피커가 각종 기기들을 연결하고, 사용자와 서로 소통 하는듯한 창구 역할을 할 수 있어 많은 글로벌 IT 기업들이 인공지능 스피커 기반의 스마트홈 환경을 만들고, 이 기술의 시장을 선도하려고 노력하고 있다.

인공지능 스피커를 이용한 스마트홈 서비스의 대표적인 예로 스피커와 TV를 연동하여 음성으로 TV를 조작하며 특정 프로그램 다시보기 등을 명령하는 것이 있다. 스피커와 스마트폰을 연동하여 스마트 플러그, 디지털 도어록, 홈캠 등을 제어할 수도 있다. 또한 냉장고, 에어컨 등 가전제품들도 인공지능 스피커를 이용해 조작 가능하다. 향후에는 전자제품뿐 아니라 모든 사물들에 사물인터넷 기술이 탑재되어 인공지능 스피커로 조작할 수 있는 미래가 올 것이다.

(2)시니어 사용자를 위한 음성인식 스피커

사회가 고령화됨에 따라 혼자 사는 노인들의 우울감, 소외감, 외로움, 위험함 등 부정적인 문제가 발생하고 있다. IT기술은 이에 발맞추어 적절한 기능들을 제공하고 있다. 그 중 하나가 챗봇활용 음성인식 스피커이다. 챗봇활용 음성인식 스피커는 고령 사용자가 간단하게 음성으로 대화하는 식의 명령을 주고, 사용자의 상태를 지켜보며 모니터링 할 수 있다. 또한 사용자의 음성을 분석하고 챗봇 서비스를 통해 대화 내용을 파악한다. 이를 통해 사용자가 원하는 서비스를 연결해주고 그 결과를 음성으로 다시 제공해준다. 이 기술은 우울감, 소외감, 외로움 등을 느끼는 사용자에게 누군가와 같이 있는 듯한 느낌을 받을 수 있도록 해준다. 가족들과 연락하고 싶은데 방법이 어려운 사용자에게는 간단한 음성 명령 조작을 통해 스피커가 알아서 연락해주기 때문에 간편하다. 위험한 일이 발생할 때는, 예를 들어 불이 났을 때 혼자 사는 시니어 사용자가 빨리 신고해야할 때는 직접조작보다 빠른 음성명령을 통해 재빠른 조치를 취할 수 있다. 이처럼 혼자 사는 노인을 위한 인공지능 스피커는 고령화된 사회적 문제를 일정 부분 해결해줄 수 있을 것으로 기대된다.

3. 제안 사항

1) 기술에 대한 독창적 생각, 제안, 활용 방안

장애인을 위한 휴대용 음성인식 스피커가 있다면 좋을 것 같다. 집안에서 또는 밖을 돌아다닐 때 장애인들은 많은 불편함을 안고 살아간다. 예를 들어 버스정류장에서 버스를 기다리는 시각장애인은 잠시 후 도착하는 버스 외에 차가 몇 분 남았는지 알 수 없다. 이때 음성인식 스피커를 이용해 원하는 버스의 남은 시간을 안내를 받을 수 있다. 또, 안내용 개를 데리고 다니지 않아도 음성인식 스피커가 카메라로 바깥상태를 인식하고 원하는 길을 안전하게 안내해줄 수 있다. 청각장애인들을 위해서는 음성인식 스피커에 진동기능과 시각적 기능을 추가하면 좋을 것 같다. 이들이 스피커에 어떠한 명령을 했을 때, 예를 들어 스피커를 통해 배달주문을 했을 때 명령을 알아들었다는 의미로 진동을 한번 울려주거나 스마일 표정이 나타나는 등으로 출력하면 청각장애인들도 알아볼 수 있기 때문이다.

2) 레포트, 멀티미디어 교과목 수강에 대한 후기, 제안 의견

이번 멀티미디어 레포트는 요즘 가장 이슈화 되고 있는 AI기술을 주제로 잡았고, 그중에서도 주변에서 쉽게 접할 수 있는 AI스피커에 대해 다뤘다. 특정한 기술에 대해 이렇게 자세히 알아보고 조사한 적이 없었는데 레포트를 쓰며 한 기술을 자세히 파헤칠 수 있는 계기가 되어 좋았다. AI스피커 기술을 조사하면서 기기 안에 이렇게 무수히 많은 기술들이 있다는 것에 놀랐고, 흥미로웠다. 가장 신기했던 기술 중 하나는 추천/개인화 기술이다. 매우 다양하고 방대한 양의 데이터들을 이용해 사용자에게 맞는 것을 추천하고 개인화하는 과정이 인상 깊었기 때문이다. 기술에 대한 알고리즘, 블록다이어그램 등을 직접 짜보면서 한층 더 깊게 이해할 수 있었고, 그냥 글을 읽는 것보다 훨씬 머릿속에 오래 남을 것 같다는 생각이 들었다. 레포트를 어떤 식으로 써야하는지 감을 잡을 수 있게 되었고, 논문을 처음 찾아보면서 어디에 어떤 논문이 있는지도 알게 되었다.
멀티미디어 과목을 수강하면서 들어보기만 했던 멀티미디어의 정의에 대해 정확히 알게 되었고, 여러 가지 흥미로운 기술들을 배워서 좋았다. 특히 아날로그를 디지털로 변환시킬 때 샘플링, 양자화 하는 과정에 대해 배웠던 것이 기억에 남는다. 아날로그를 디지털화 하는 것도 신기했는데 그 과정을 배우고 이해할 수 있어서 더욱 신기했기 때문이다. 그래픽 기술 등 다른 멀티미디어 기술도 쉽게 강의해주셔서 재밌고 이해하기 쉬웠다.

참고 : SKT