인공지능 서비스 누구(NUGU) 기술 소개

2018. 6. 22

SK Telecom 정규준

Contents

AI Assistant : Speakers

NUGU

https://www.youtube.com/embed/YrJSgg_2mEg

음성인식 인공지능 스피커 (국외)

amazon

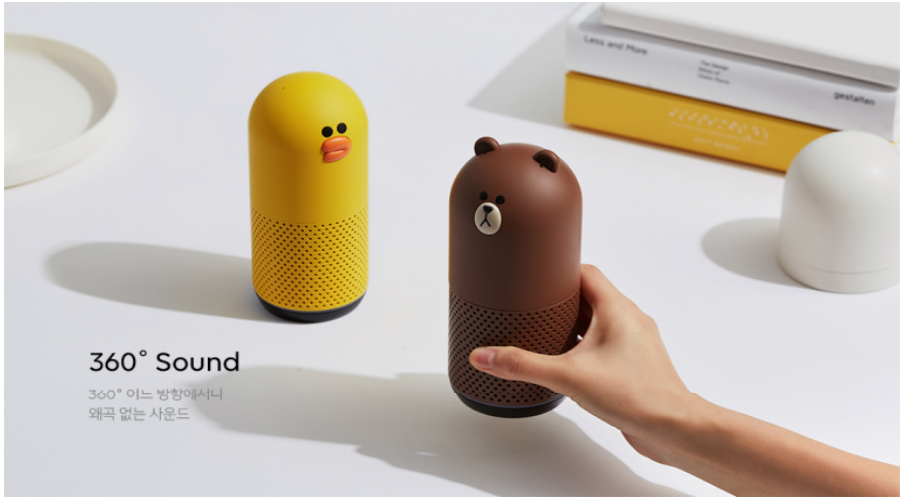
Echo

Google

Home

음성인식 인공지능 스피커 (국내)

NAVER



kakao

음성인식 인공지능 스피커 NUGU (1/2)

세계 최초 한국어 음성인식 지원 음성인식 스피커

스마트폰 음성비서

음악감상, 팟캐스트

날씨, 뉴스, 경기결과, 운세

제어, IoT

알람, 스케줄

음식 주문, 쇼핑

Artificial Intelligence (AI) speaker

음성인식 인공지능 스피커 NUGU (2/2)

’15. 3月
Smart Box I Prototype

’15. 8月
Smart Box II Prototype

’16. 9月
NUGU 스피커 상용화

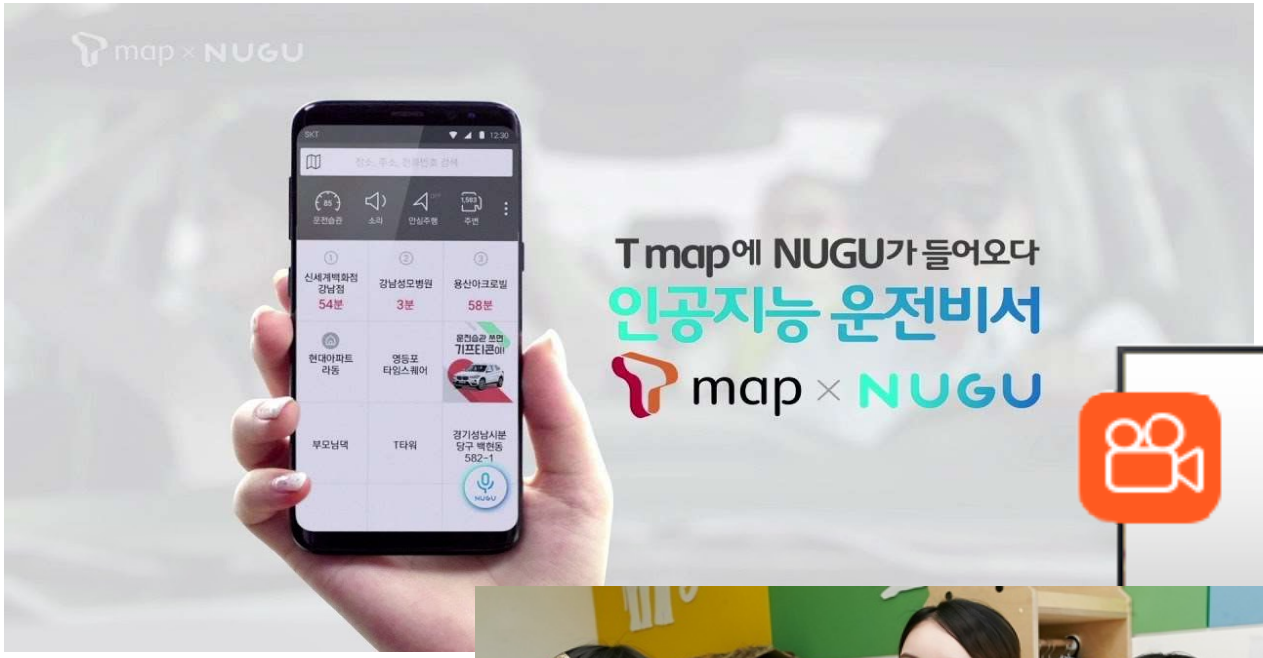
AI Assistant : 내비게이션

AI Assistant : 내비게이션

https://youtu.be/TXw7468D8U0?list=PLhp6UuAYrPtqVdkta98K0zY0zuyeveKZN

NUGU Platform

Tmap x NUGU



Btv x NUGU

JOON x NUGU

NUGU Platform

2018
API Open

Social bot

Pet bot

Commerce bot

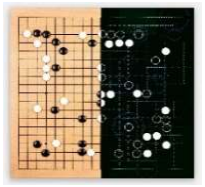
Toy bot

1. 현재 Business에 접목하면서 미래를 위한 포석을 만들고 있음

Google

번역

게임



운전

"AI as Platform"

생활 전반의 모든 서비스를
구글을 통해서만 접근

amazon

물류

비서

배달

"Commerce"

고객의 선호도를 누구보다
빠르게 파악하고 유통

IBM

헬스케어

Finance

"AI Expert"

전문가의 식견을 AI화하여
Solution 제공

SKT의 AI (1/2)

- **Life Companion : “Connect everything if you want”**

[Human Life]

Connect

Intelligent Agent

Integrated Service Platform

Smart Advisor

- Commerce
- Contents
- Communication

[contents/service/info.]

MelOn

btv mobile

WIKIPEDIA The Free Encyclopedia

map

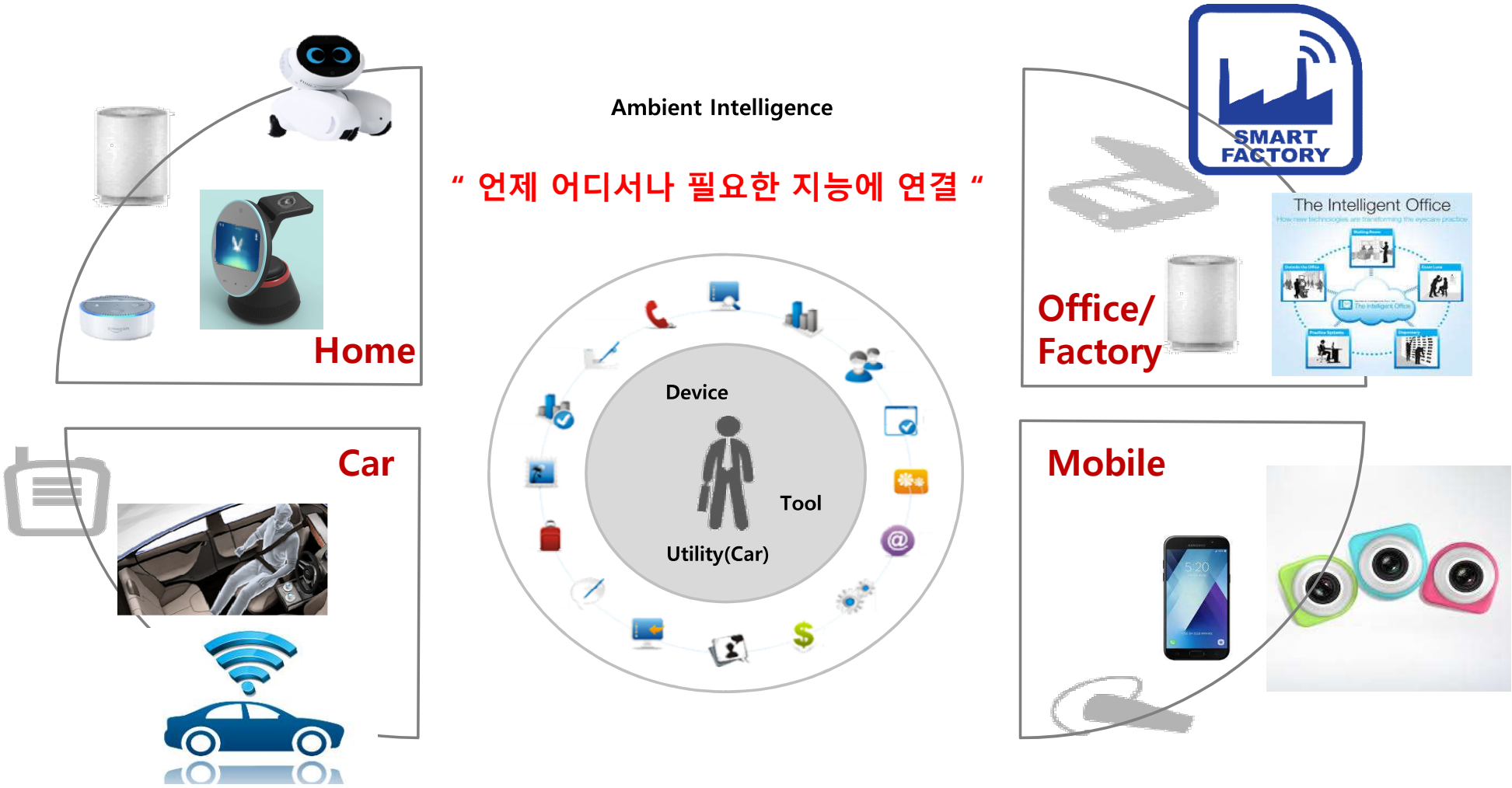
1ST

table

[SKT “AI R&D target” report, Sep. 2015.]

SKT의 AI (2/2)

- 다양한 공간에서 다양한 Device를 통해 Ambient Intelligence를 제공하여 새로운 가치를 창출함



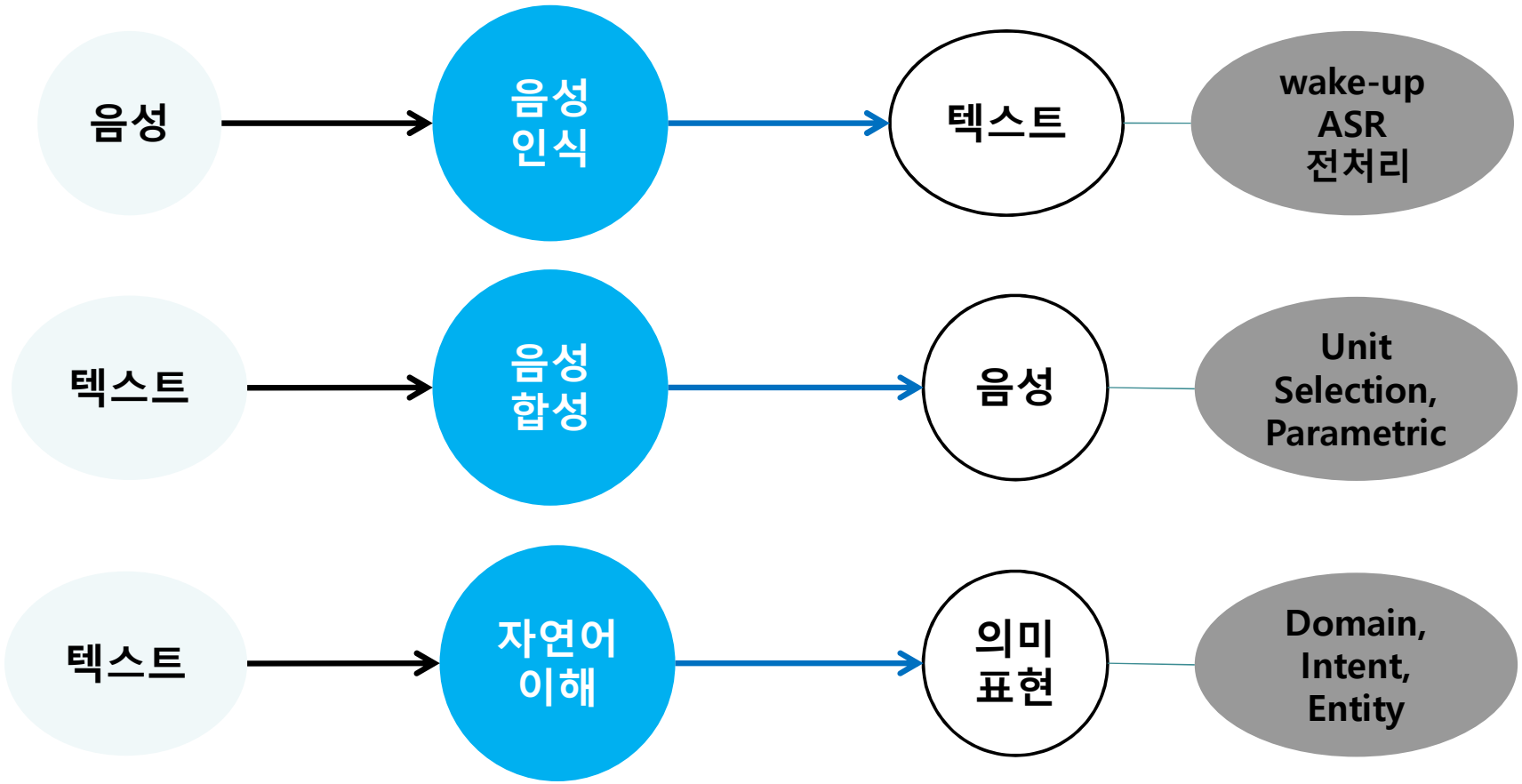
Contents

Natural User Interface

Intelligence

15

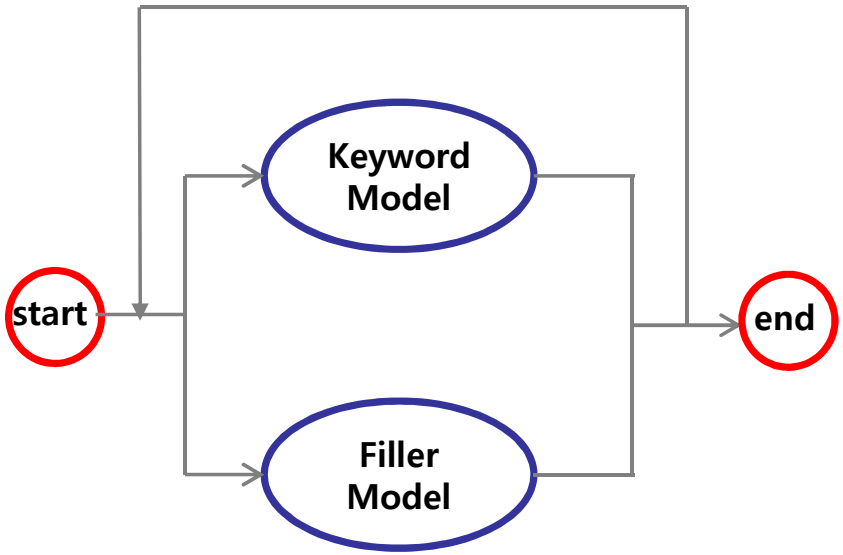
Core Technology - NUI

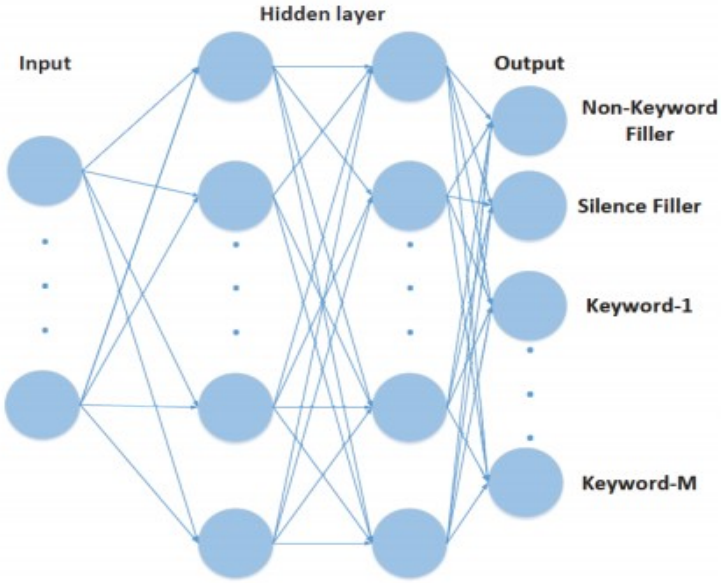


16

Wake-up (Keyword Spotting)

【 Basic structure 】





Advantage

- Use low power resource
- Available for wake-up word
  (Echo: 'Alexa', NUGU: 'Aria', 'Tinkerbell)

Dis-advantage

- Performance degradation compared to server-side ASR recognition
- pre-registered keywords only

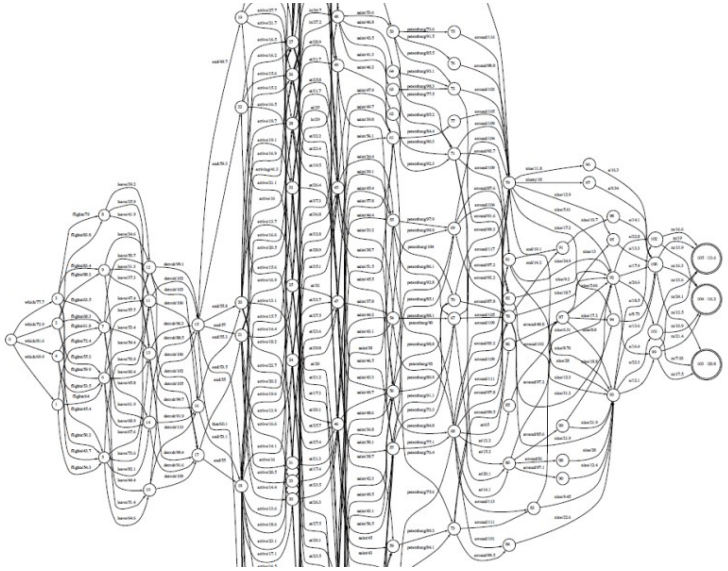
Automatic Speech Recognition (ASR)

Speech Signal → ASR → Recognized text

"서울 날씨 알려줘"

Challenges

- Background noise
- Low power speech
- Far-field
- Cross-talk
- Incorrect pronunciation
- Similar pronunciation
- New word, non-standard word
- Large vocabulary
- Natural conversation



Natural Language Understanding (NLU)

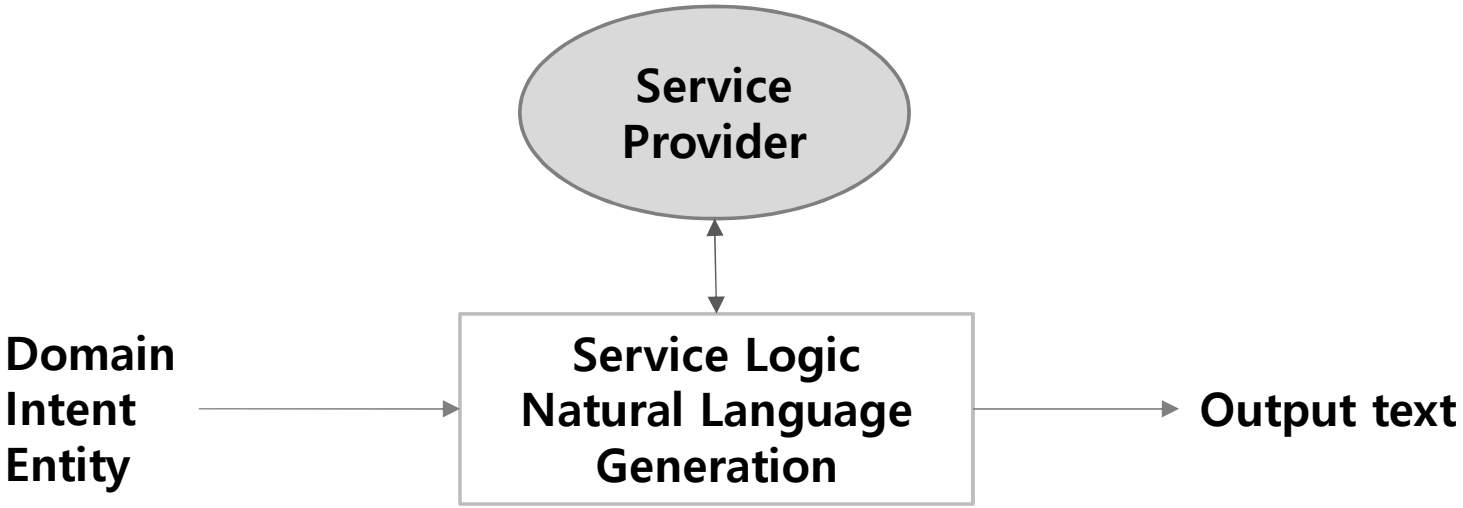
"서울 날씨 알려줘"

Domain: Weather
Intent: ask.weather
Entity: date.tomorrow, location.Seoul

Challenges

- Ambiguity of expression
- Proper noun
- pronoun
- Domain increasing
- New word, non-standard word
- Natural conversation
- ASR error

Service / NLG



Domain: Weather
Intent: ask.weather
Entity: date:tomorrow, location:Seoul

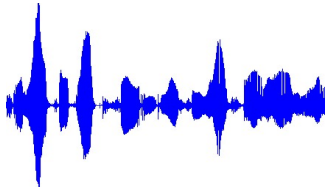
"내일 서울은 쌀쌀하고 구름이 다소 낄 예정입니다. 최저 기온은 영하 1도, 최고 기온은 … "

Challenges

- **Looks like human**
- **Domain increasing**
- **Handling Input outside coverage**

Text To Speech (TTS)

"내일 서울은 쌀쌀하고 구름이
다소 낄 예정입니다. 최저 기온은
영하 1도, 최고 기온은 … "



Challenges

- **Naturalness**
- **Tone, pause, speed**
- **New word, non-standard word**
- **Exception pronunciation processing**

Core Technology - Intelligence

검색 기술

목표: 모든 CP 데이터 검색 내재화

(완료) 지식검색, 감성대화검색, T map 주소록 검색

(진행 중/예정) 뉴스 검색, 음악 검색, B tv 검색,
T114 검색, 팟캐스트 검색

요약 기술

뉴스 요약

통합 지식베이스

이종의 DB 통합

Entity Linking

("소녀시대 서현 " – "서주현" vs "분당구 서현)

추천 기술

개인화된 음악 추천 제공 (2017년 12월 출시)

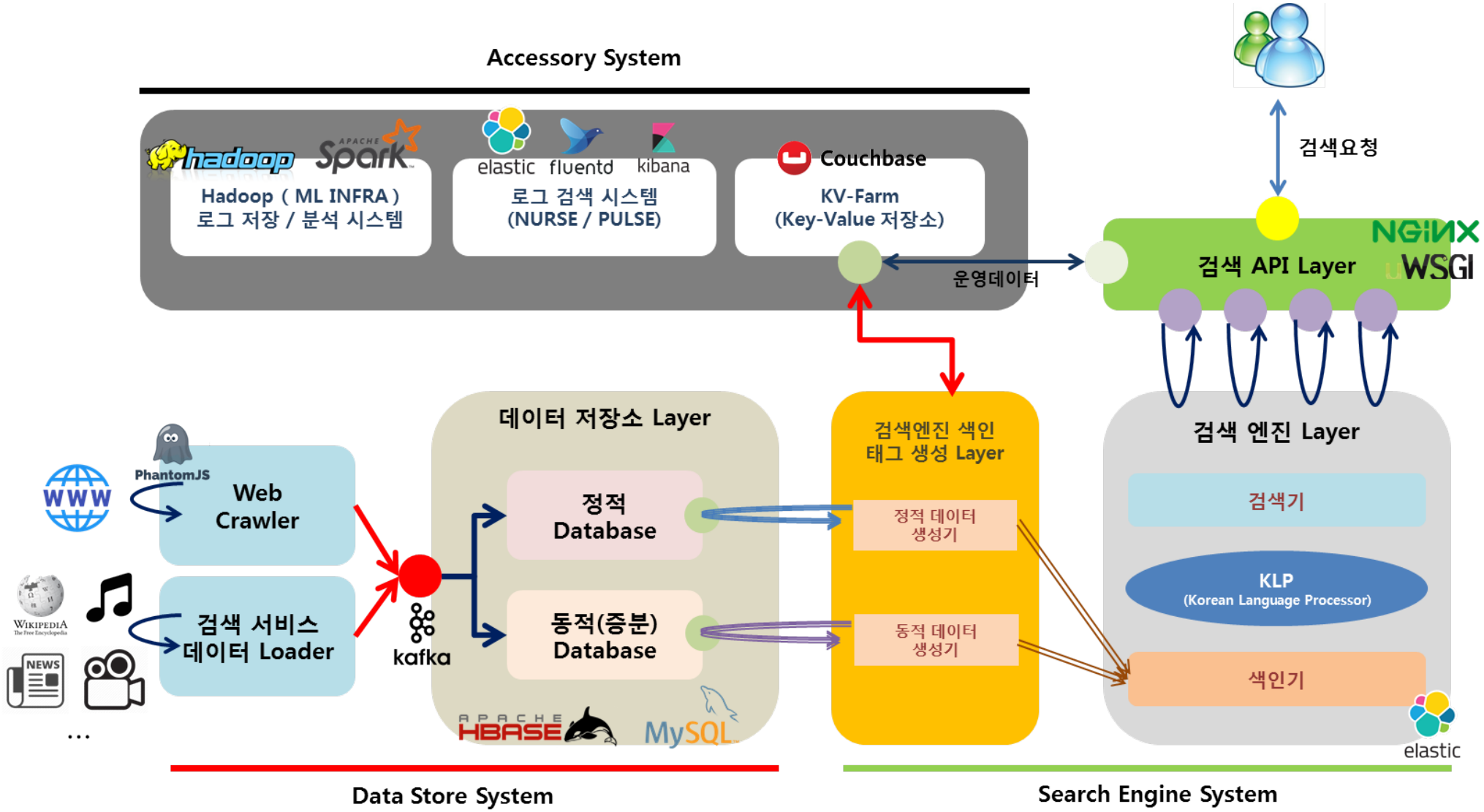
추천곡 중, 3곡 이상 (33%), 5곡 이상 (24%),
7곡 이상 (17%)

채택 기준: 전체 재생 시간의 60% 이상 청취

22

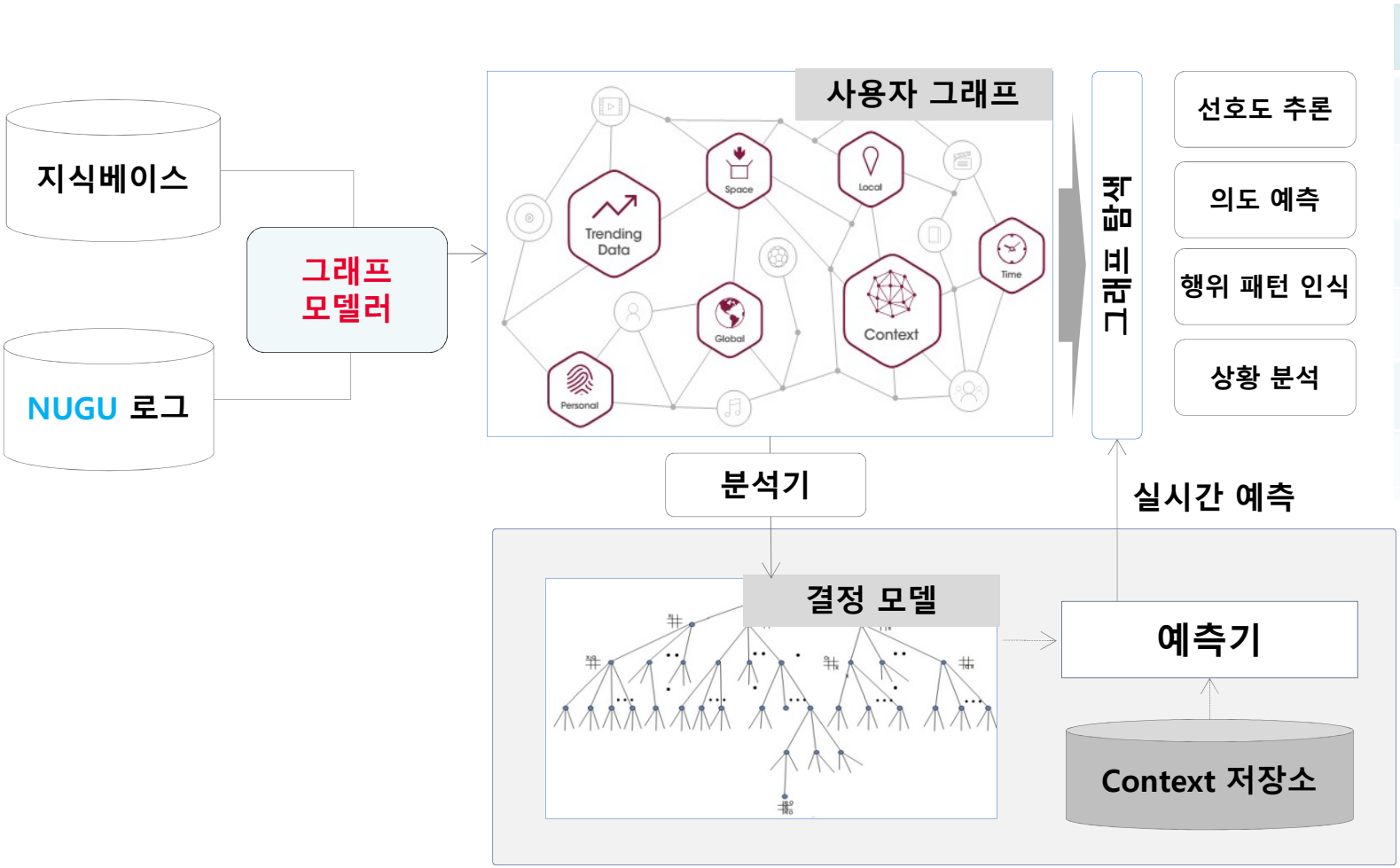
NUGU 검색 시스템 구조

수집, 제휴를 통한 데이터 생성과 유실 없고 빠른 서비스 투입을 위한 검색 시스템 구축
검색 로그 분석을 위한 빅데이터 처리 시스템 구축



23

NUGU 개인화/추천



제공 정보
노래 선호도
가수 선호도
의도 요청 비율
개체-의도 연관도
개체 간 결합 선호도
자주 하는 요청 패턴

시나리오

- 사용자 선호도를 반영하여 음악 추천
  - 예) "노래 추천해줘" → 좋아할 만한 음악 추천
- 명시적 의도 없이 개체명만 발화하는 경우, 의도를 예측
  - 예) "뽀로로" → 뽀로로 동영상이 아닌 뽀로로 주제곡 재생

→ **개인화 추천**
의도 예측

NUGU Knowledge Base 현황

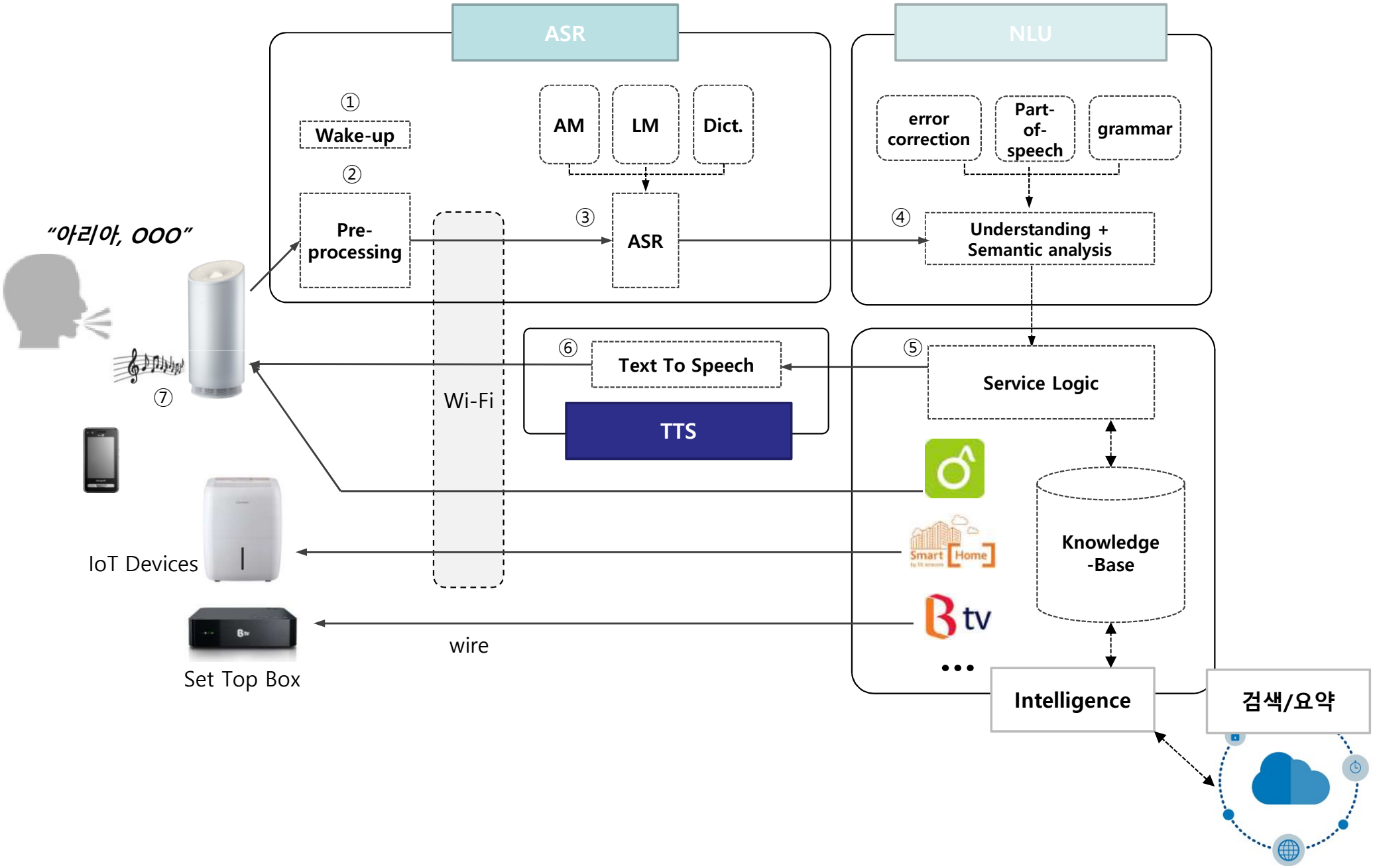
총 2천 7백만 Entity, 2억 5천만 Triple의 KB를 구축 중 (2018. 03. 31 기준)

DOMAIN	CATEGORY	ENTITY 수	Triple 수
음악	앨범	1,274,126	14,182,888
	노래	11,232,220	96,288,283
	가수	1,019,178	6,367,323
방송/영화	**VOD**	241,413	12,020,348
	인물	23,001	666,472
	방송	32,306	1,177,776
	코너	53,344	1,587,367
	채널	232	928
	편성표	713,917	9,844,981
	영화/상영관	5,174	15,648
교통	**POI**	5,023,558	52,623,268
	도로명	6,245,744	37,448,542
	행정명	46,015	230,075
오디오	**오디오북**	9,745	139,909
	팟캐스트	10,969	96,124

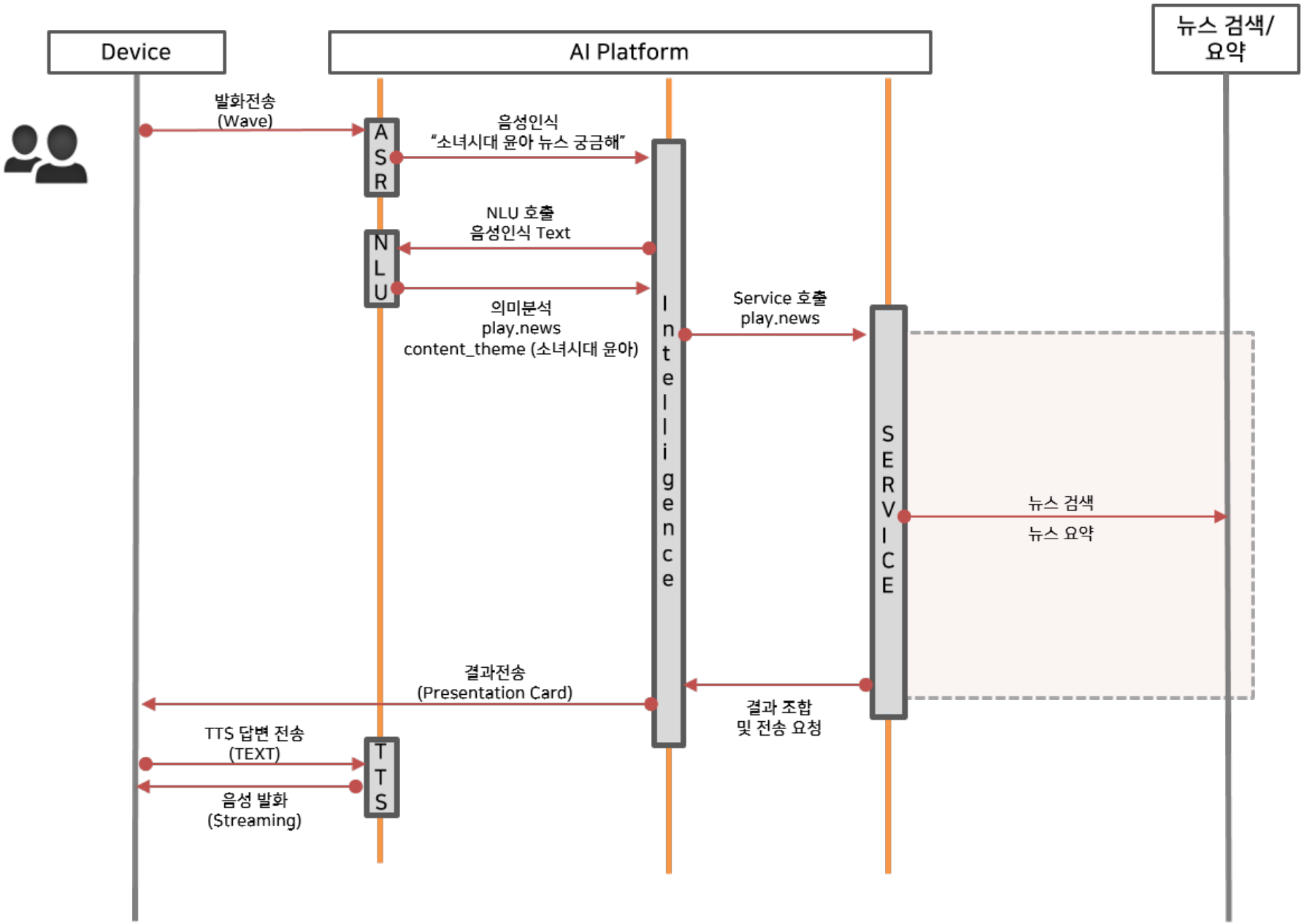
DOMAIN	CATEGORY	ENTITY 수	Triple 수
지식 백과	위키피디아	375,252	4,893,250
	기네스북	187	747
	고사성어	364	2184
	공공사전	42,299	302,1998
	민족문화사전	69,206	526,127
	큐레이션	7,343	45,137
	우리말 사전	921,156	11,492,687
	시사 용어	63,420	172,450
	속담/학습	3,084	19,788
스포츠	평창올림픽	250	1,740
증권	KOSDAQ	1,236	1,388
	KOSPI	1,333	2,060
음식	음식명	26,646	133,230
	상점명	1,000	10,177

25

NUGU – Service Flow



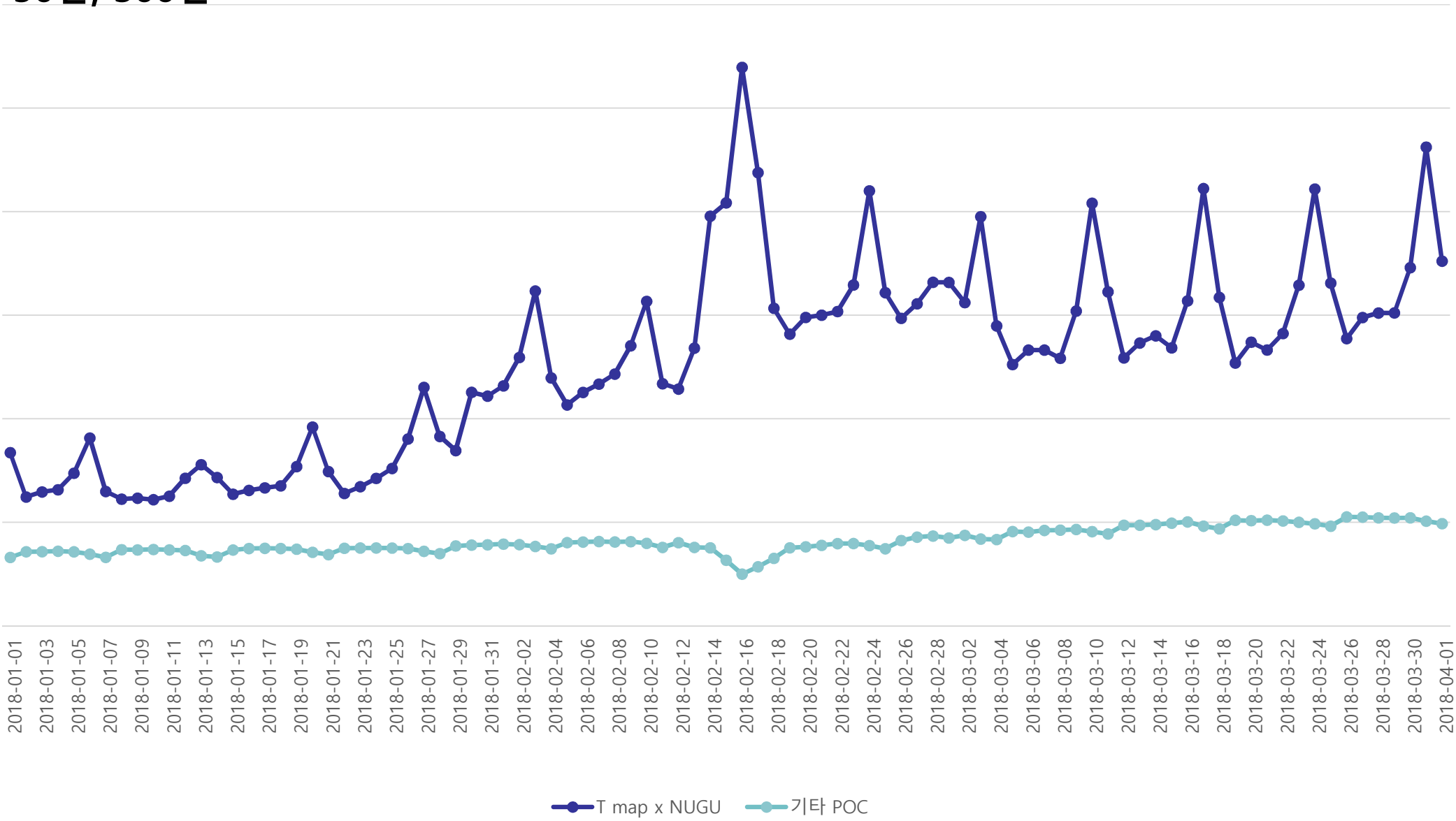
NUGU Platform (상용 적용 예제)



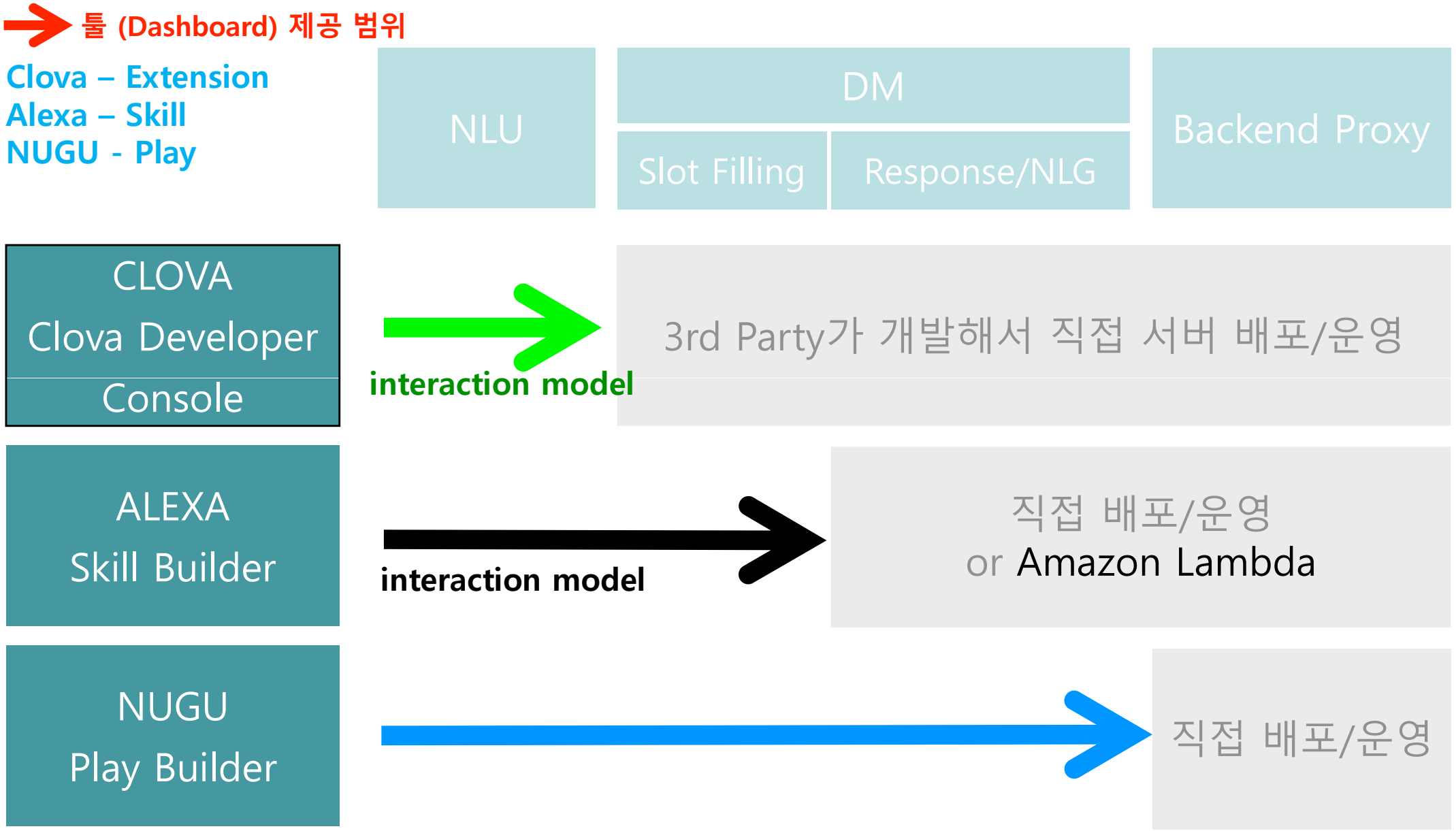
27

NUGU 사용자 지표 – Daily Active User (DAU)

50만, 300만



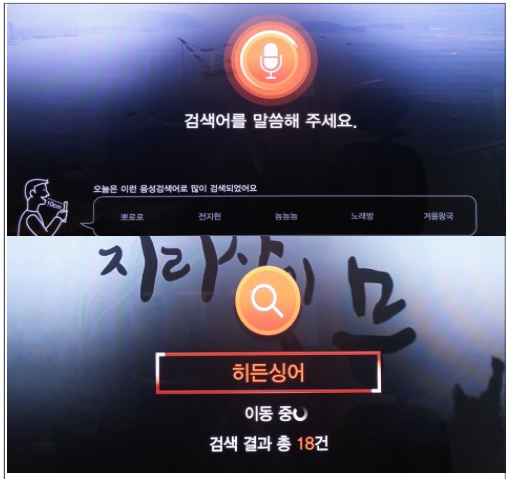
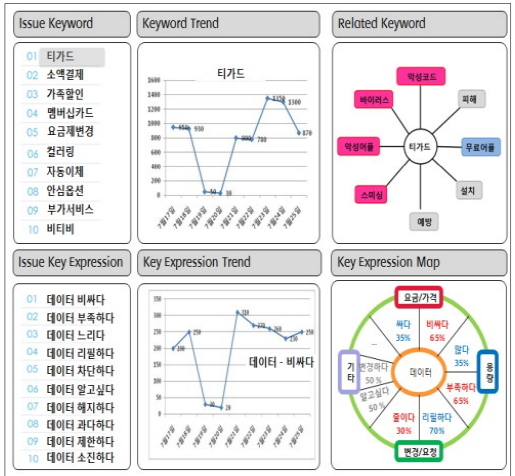
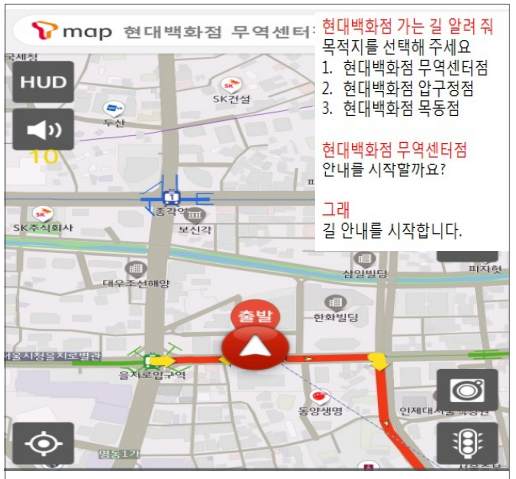
Open Platform NUGU Play 개발



Contents

SKT 음성 인식 기술

1. **2014년부터 상용화 시작**
2. **NUGU, Btv, T-map, 고객 VoC 분석 등에 적용**

Voice Search (Set-Top box)	VOC analysis	T map (Navigation)	NUGU (speaker)
SK Broadband ('14.09)	SKT ('14.11), Shinsegae ('16.03)	SKT ('16.12)	SKT ('16.09)
			
• ASR based contents search • Contents title, channel, people name search	• ASR for call center • Keyword trend, issue keyword, related word extraction	• POI Voice Search • Command and Control • Natural Language Processing & Dialog Management	• Far-Field ASR, Echo cancelation • Wake-up, TTS • Music, Weather, Smart Home (add services continuously)

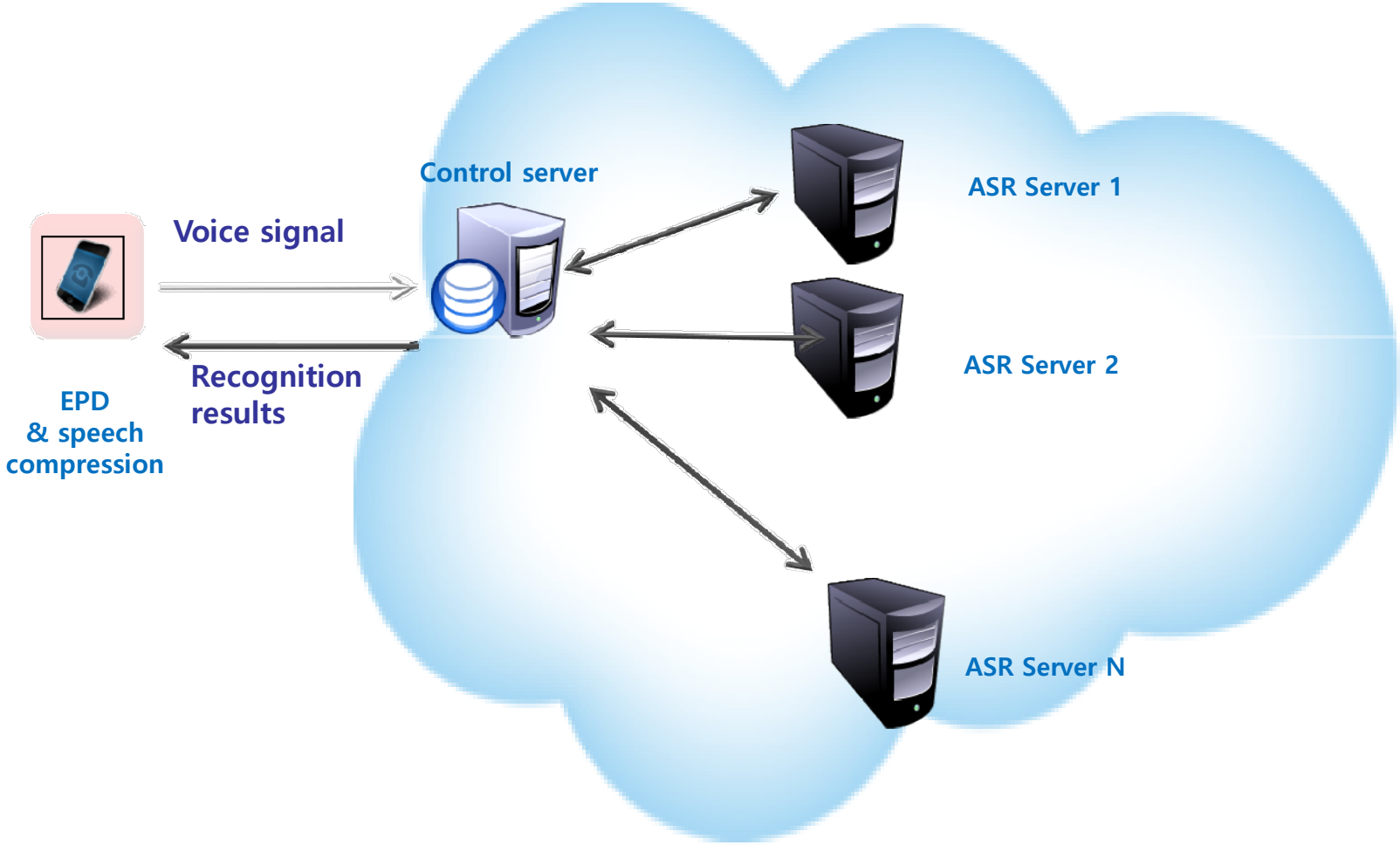
GMM → **DNN**

Cloud based speech recognition system

Client : end point detection, speech capture and compression

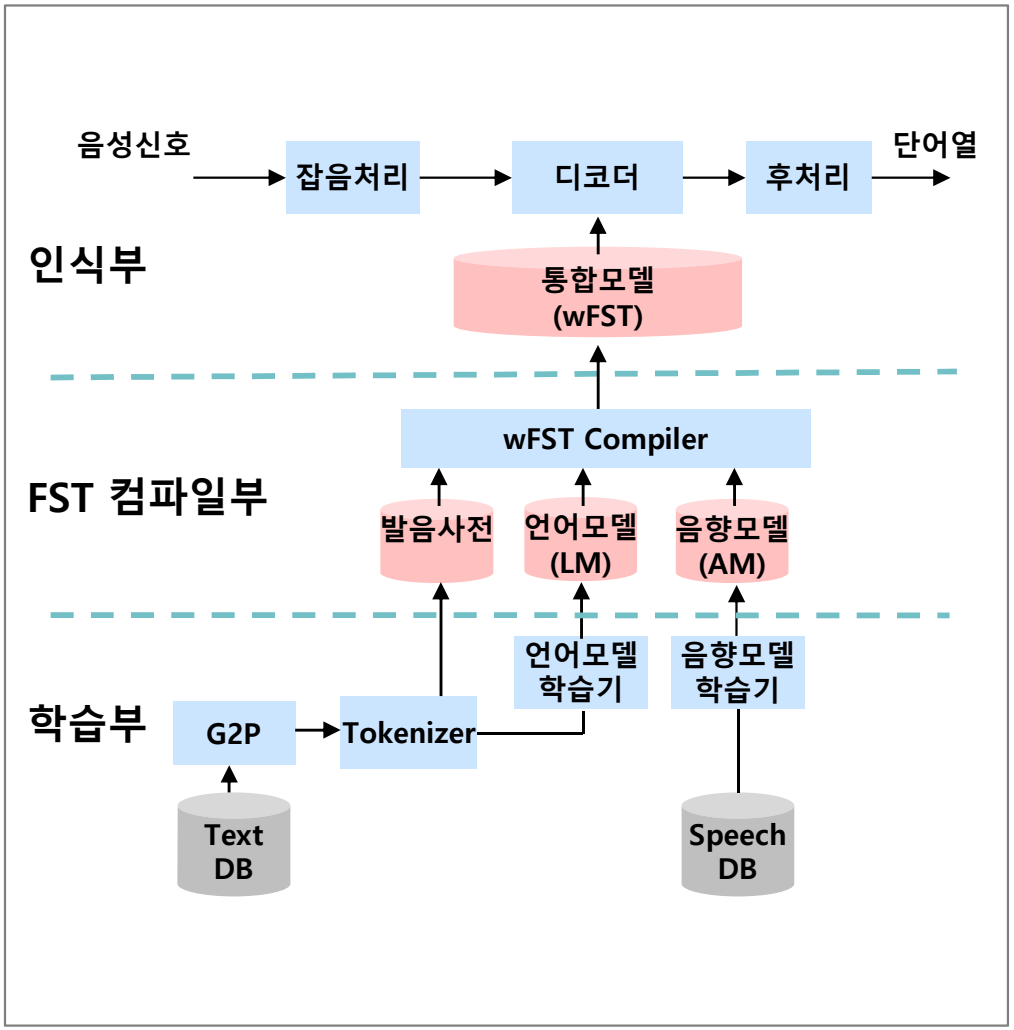
: send streaming data into the cloud server

Server : run large vocabulary speech recognition system and return results



음성인식 기술 구성

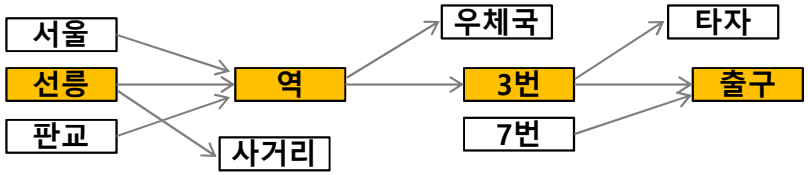
음성인식 기술은 크게 모델을 학습하는 단계, 학습된 모델을 이용하여 인식하는 단계로 구분되고, 이 중 음향, 언어 모델을 학습할 수 있는 기술이 핵심



- **통합모델(wFST) 기술**
  - 문장 단위 학습에 최적화되어 속도와 인식률 향상
  - 향후 대용량 연속 어휘, 즉 자연어 음성인식을 위한 핵심 기술

- **모델링 기술 및 데이터**
  - 음향모델 : 입력 신호와 음소의 유사도

  - 언어모델 : 단어간 확률 관계 그래프



  - 발음사전 : 단어의 발성 정보 저장

    예) 선릉 : 설릉(ㅅ ㅓ ㄹ ㄹ ㅡ ㅇ), 선능(ㅅ ㅓ ㄴ ㄴ ㅡ ㅇ)

음성인식 기술 구성 - 자체 기술 개발

속도, 성능을 향상시킨 wFST, 컴퓨팅 파워의 향상에 기반한 DNN 기술 적용

전처리	Feature	HLDA, STC, Equalization / Wiener, Kalman Filter / Model Space
	Neural Net	Bottleneck Feature
학습	Discriminative Training	MPE, fMPE
		MCE, MMI
	Big LM	Distributed Modeling, Long Span LM
	Deep Neural Network	DNN based Acoustic Modeling Training
인식	Dynamic Network	FSN
		Lexical Tree
	Static Network	wFST (weighted Finite State Transducer)

음성인식 기술 구성 - Neural Network 기반 언어모델링 (NN-LM)

Neural Net기반 언어모델에서는 아직까지 음향모델만큼 큰 성능향상을 이루지는 못함

Probabilistic Language Model

- 단어열 $w_1^n = w_1 \ldots w_n$
- N-gram $P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-N+1}^{k-1})$
- 활용 분야
  - Machine Translation
  - Spell Correction
  - Speech Recognition
  - Summarization, question, answering, etc.
- 예제

나는 학교에 간다
학교에 학생이 있다
학생이 학교에 있다

\2-grams:	prob.
나는 학교에	1/1
학교에 간다	1/3
학교에 있다	1/3
학교에 학생이	1/3
학생이 있다	1/2
학생이 학교에	1/2

P(학생이 학교에 간다)
= 1/2 * 1/3

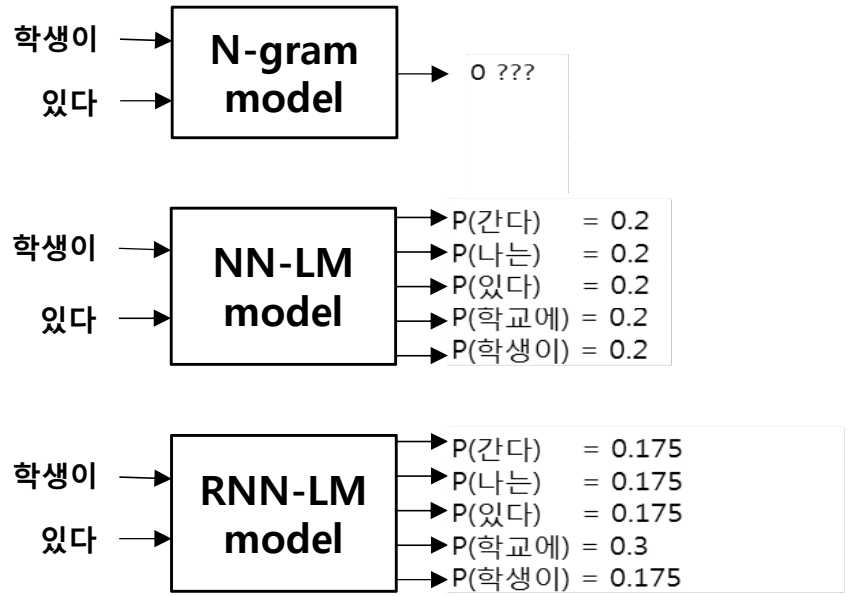
P(학생이 있다 학교에)
= 1/2 * 0 ???

✓ Unseen !!!

✓ longer history !!!

NN기반 언어모델과 N-gram 비교

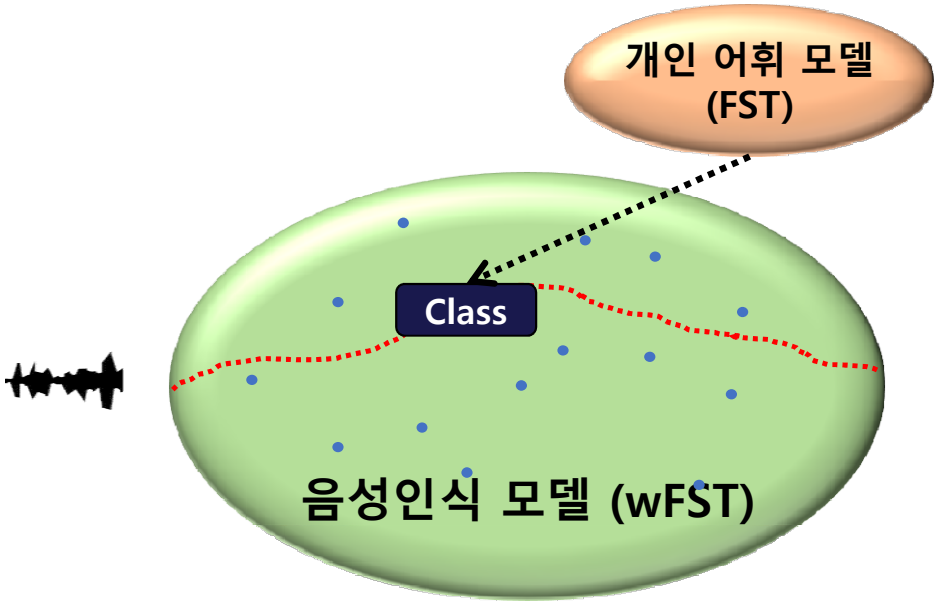
❖ P(학교에 | 학생이 있다)



1. N-gram vs NNLM
   - Better to unknown n-grams
   - Heavy computation
2. NNLM vs RNNLM
   - Utilize short term memory
   - Clustering of similar histories

✓ Still N-gram approach is best !!
✓ NNLM supports n-gram model

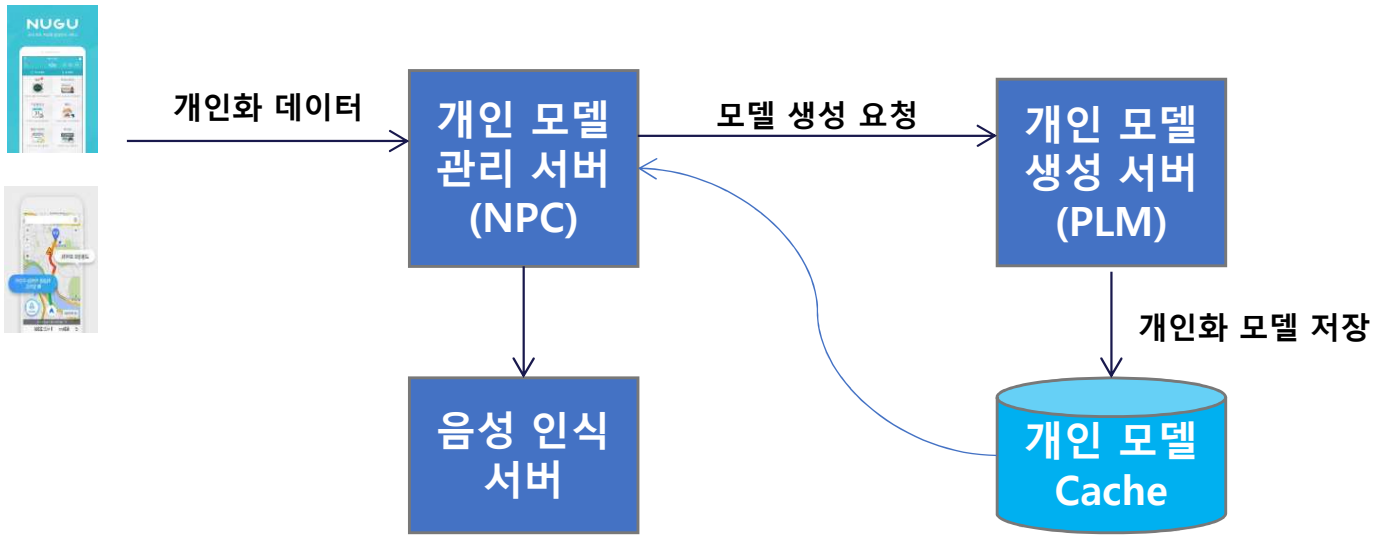
음성인식 기술 구성 – Personalized Language Model (PLM)



- 개인화된 어휘가 사용될 위치를 사전에 class 형태로 모델링

  NAME__ 에게 전화, NAME__ 한테 문자,
  NAME__ 이한테 전화

  NAME__ Class를 개인별 주소록을 이용해 확장 (NAME__ = 엄마, 김강율, 한동근 등)



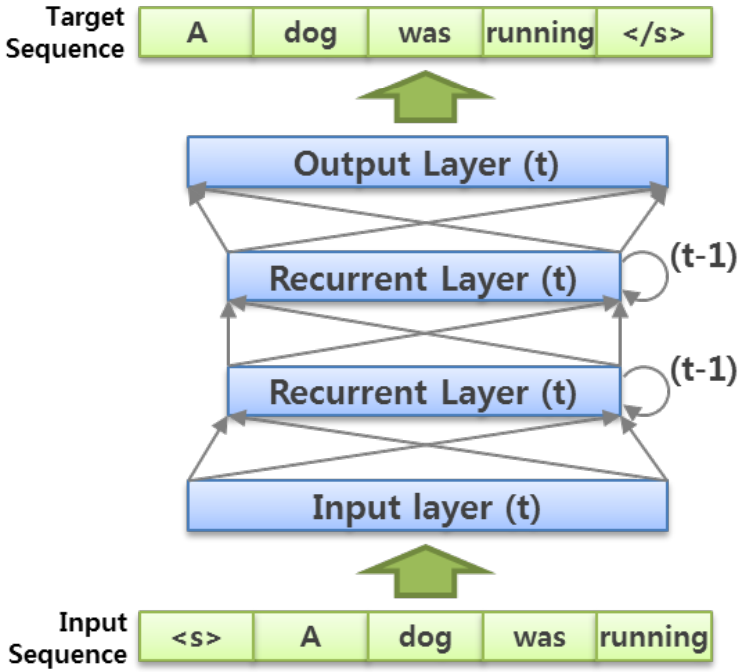
36

음성인식 기술 구성 - Sequence to Sequence Learning, CTC

RNN-LSTM을 음절 기반의 띄어쓰기 모델에 확장 적용함으로써 성능 향상.
발음열 생성기술에 CTC(Connectionist Temporal Classification) 적용 등 다양한 영역에서 DNN 적용을 시도 중

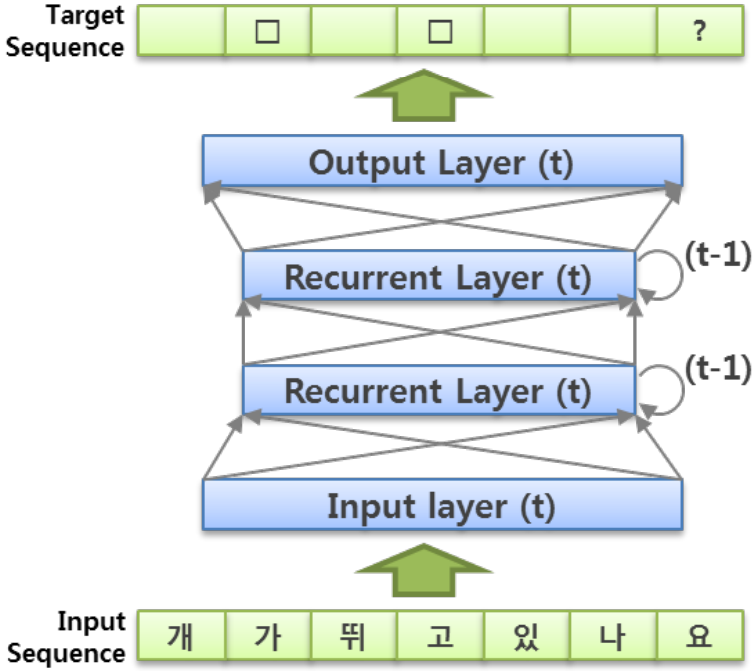
LSTM 언어모델링

문장을 구성하는 단어 sequence에 대해
다음 단어의 sequence를 target으로 학습



LSTM 띄어쓰기 모델

한글 corpus의 음절 sequence에 대해
각 음절 별 띄어쓰기 및 문장부호를 target으로 학습



음성인식 기술 - 원거리 음성인식

전용 Device를 이용한 음성인식은 기존 방식과는 음향 환경 및 요구조건의 차이가 매우 커 많은 것을 고려하여 개발 필요

【 기존 방식과 차이점 】

1 원거리 음성인식
- 2~3m의 먼 거리에서도 음성인식이 가능해야 함

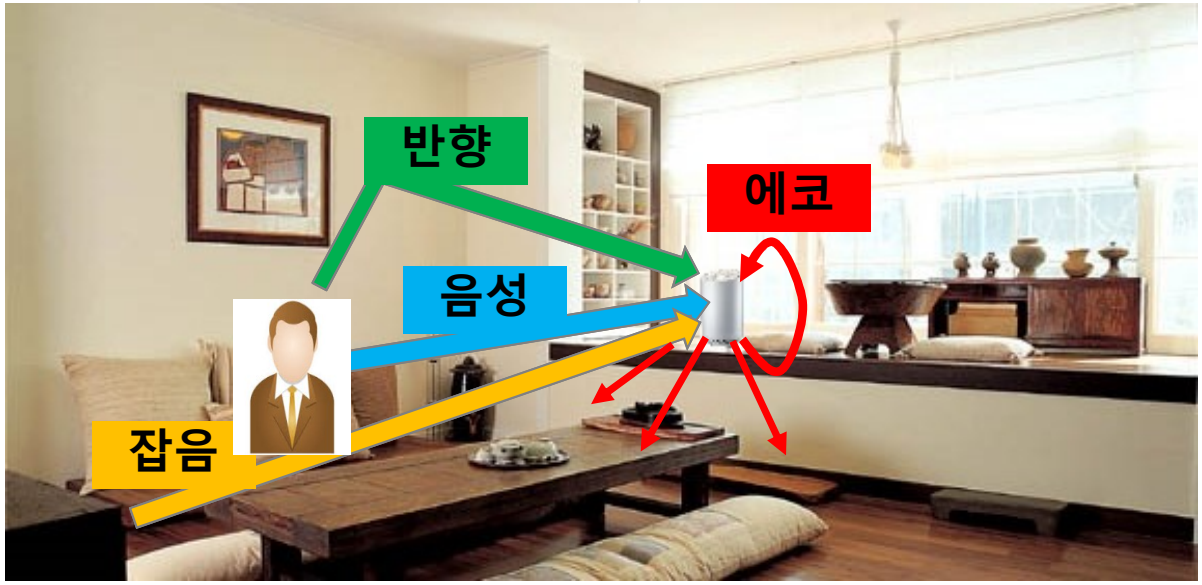
2 에코 제거
- 음악이 나오는 상황에서도 음성인식 기능이 동작

3 음성 Trigger
- 음성을 이용한 서비스 시작이 가능

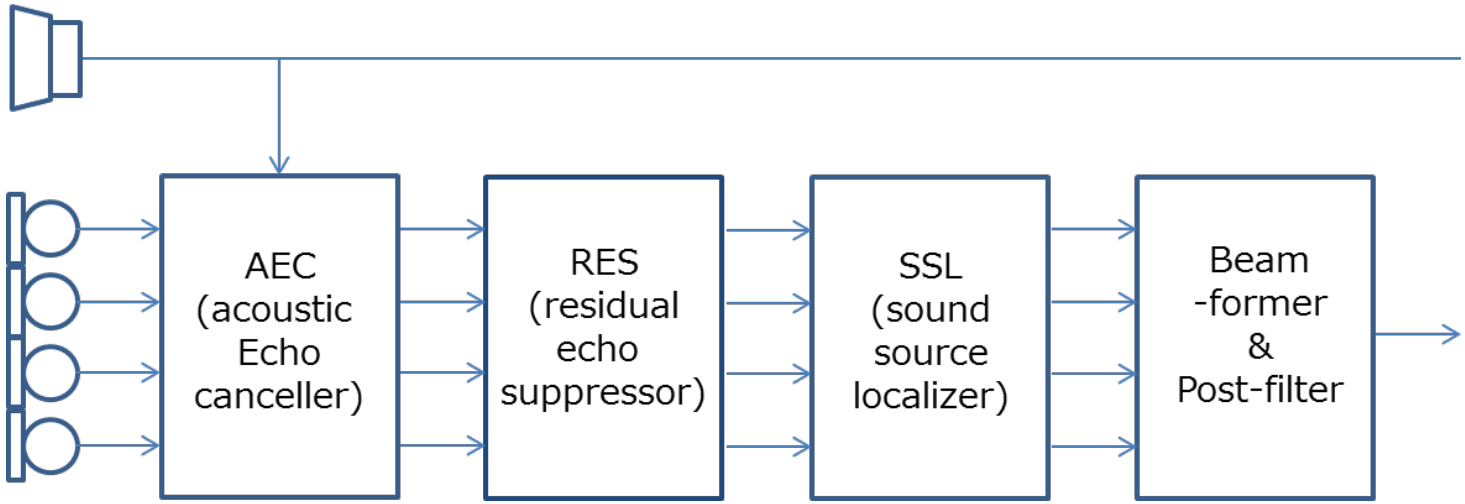
【 해결 방식 】

2개 Mic를 이용한 전처리
- 인간의 귀와 같이 2개의 마이크를 이용하여 Gain 보상, 에코 제거 기능을 구현하여 적용

DNN을 이용한 음향모델 적응
- 전처리를 통해 왜곡된 신호를 보상하기 위한 Simulation기반 음향모델 학습 수행



음성인식 기술 – 다채널 음성전처리 개발

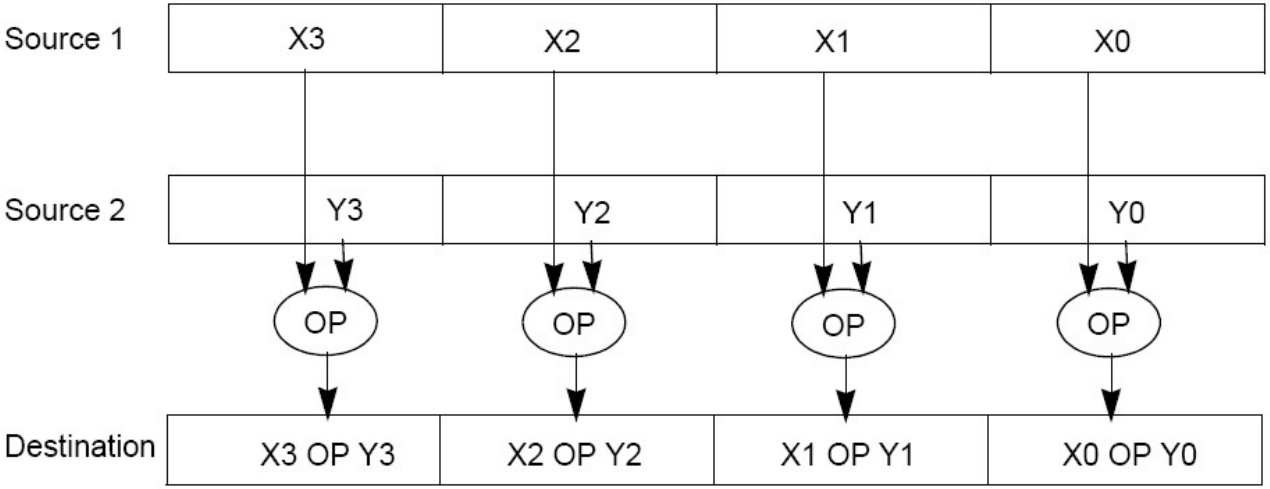


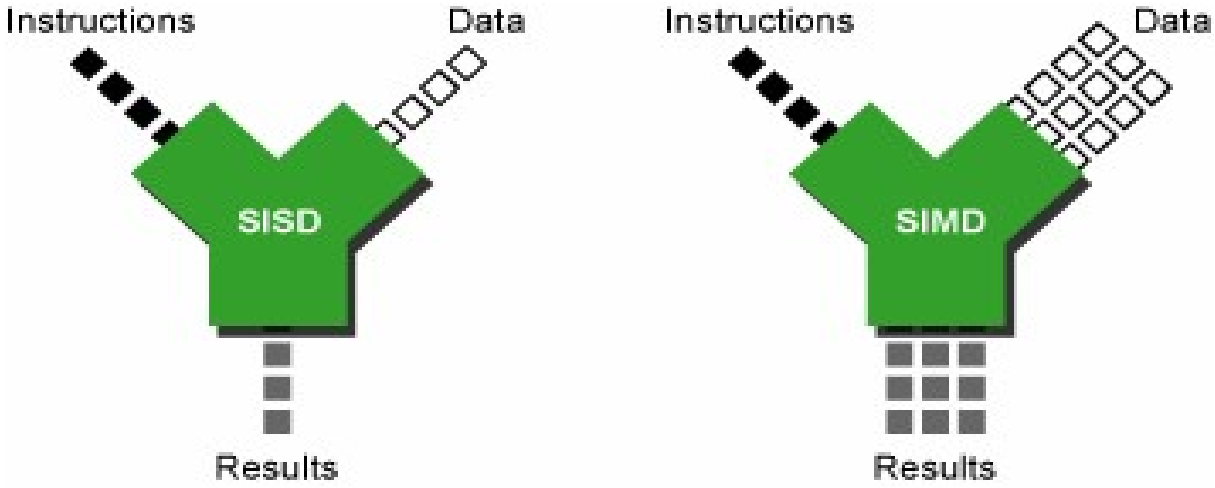
- AEC
  - 낮은 SER 환경에서 음성 인식 성능 확보
  - Sub-block 필터 구조 적용하여 음향 전달 함수의 long-tail 제어력 강화
- SSL
  - 발화자 방향 정보 예측
  - 잡음과 잔여 에코 예측 기반 가중치를 적용하여 wake-up 발화 대응력 강화
- Beam-former
  - MVDR (Minimum Variance Distortionless Response) 빔포밍 적용으로 왜곡 없는 음성 취득
  - 방향별 필터 상수 차별 적용으로 분해능 강화

39

음성인식 속도 개선 – SIMD (single instruction multiple data)

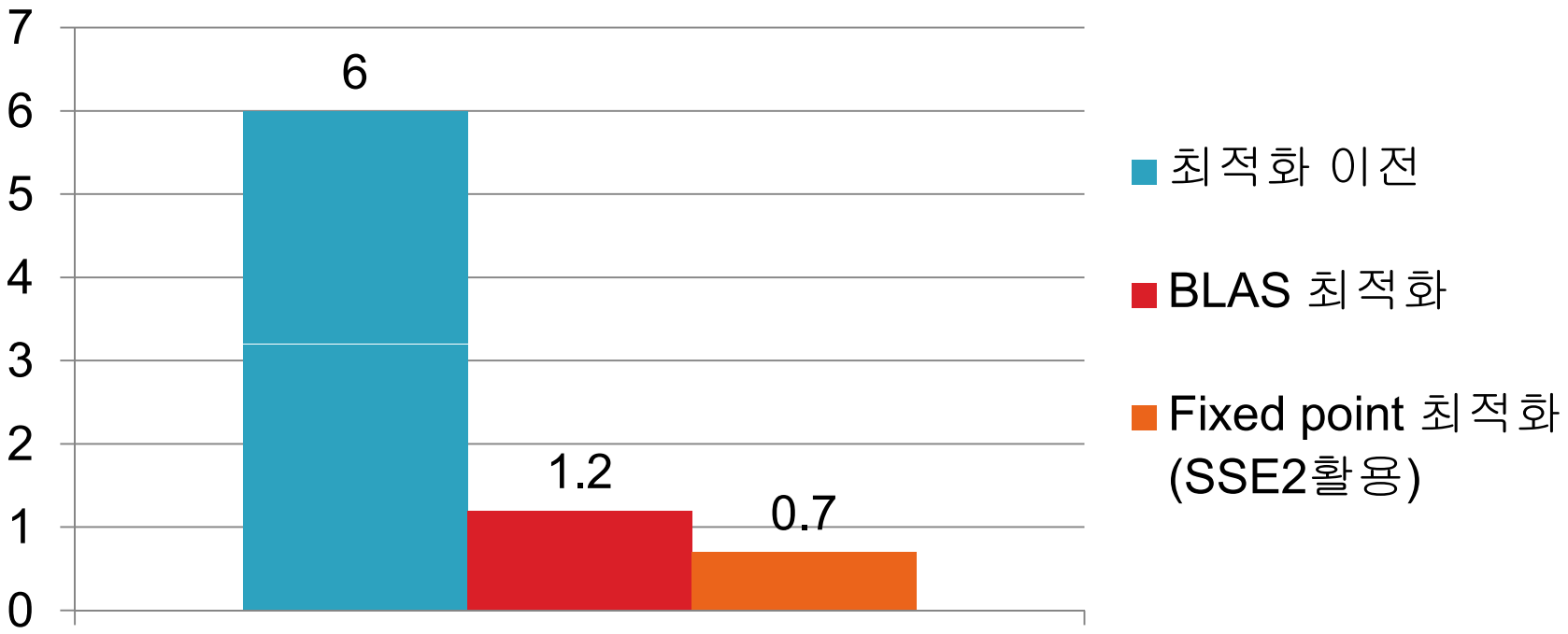
하나의 Instruction에서 vector의 연산수행이 가능한 알고리즘에 대해서 병렬 core를 이용한 효율적인 계산 알고리즘을 제공





음성인식 속도 개선 – SIMD를 이용한 online Decoder 개발

연산 최적화에 따른 속도 개선



음성인식 기술 발전 방향 – H/W와 S/W의 모두를 잘 활용

학습 순서

1) Data 분배

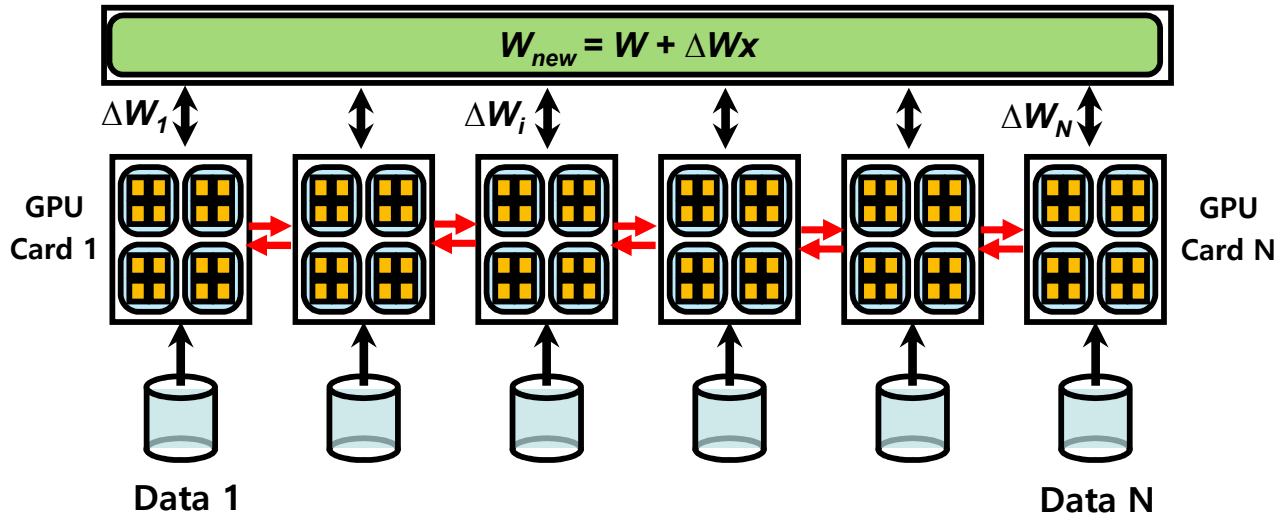
2) 중간결과 산출(GPU-> CPU)

3) 새로운 parameter(W) 계산

4) W 재분배 (CPU -> GPU)

5) 수렴할 때 까지 Step1부터 반복

개념도



단순히 나누기만 하면 되나? ➔ 아니다! 잘 나눠야 한다

고려 사항

학습 속도: 연산 시간, 네트워크 통신(mini-batch 크기) ➔ Hardware 구성까지 고려해야 함!

수렴 여부: learning rate 설정, Asynchronous SGD(HOGWILD)

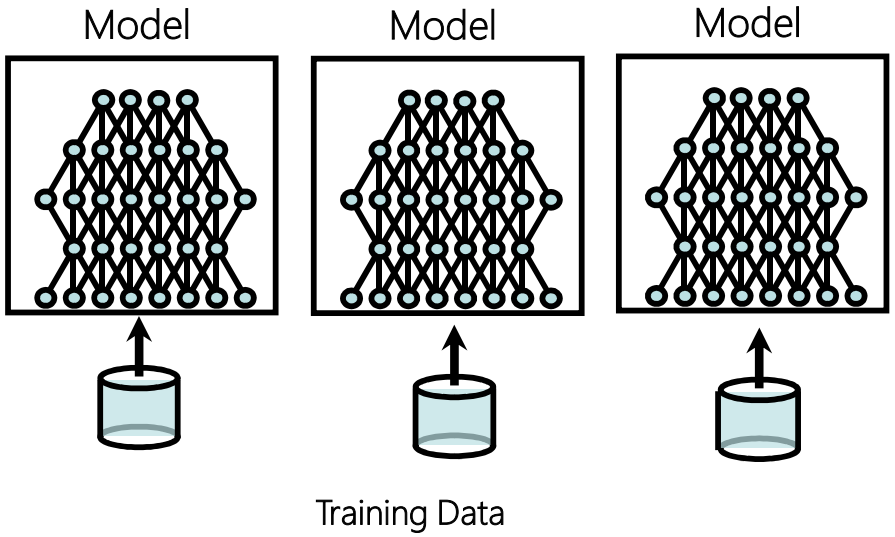
Model Size: Data 분할 vs. Model 분할 vs. Matrix 분할

분산처리 기반 DNN 학습 (Data vs. Model Parallelism)

【 Data Parallelism 】

같은 모델을 이용하며 다른 데이터를 이용해 학습

➔ 주기적 동기화



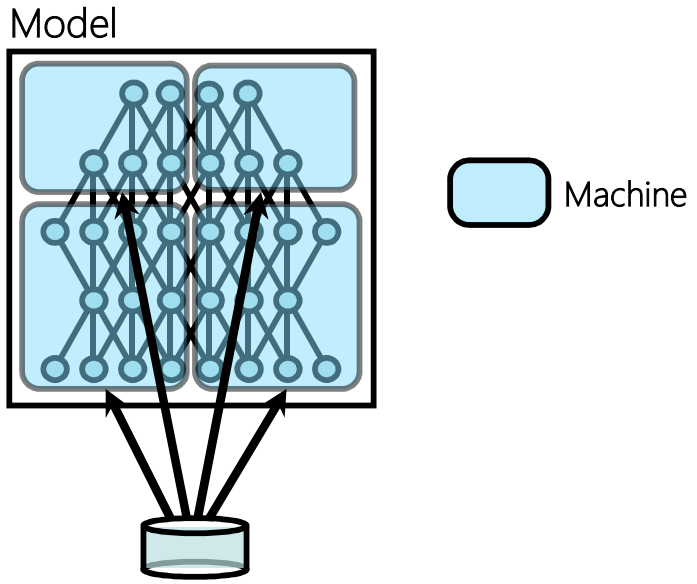
- 장점:
  1. 작은 size의 DNN 모델에 대해서 빠른 성능
  2. 프로그램 제작이 간단함

- 단점: 큰 size의 DNN 모델에 대해 학습 불가능

【 Model Parallelism 】

모델을 나눠 다른 서버로 전송

➔ 동일한 데이터를 이용해 나눠진 모델을 학습



- 장점: 큰 size의 DNN 모델의 학습이 가능

- 단점:
  1. 모델 특성에 따라 parallel 부분이 바뀌기 때문에 코드 제작이 어려움
  2. 모델 업데이트 시 machine간(GPU간) 통신 load를 고려해서 설계해야 함
  3. 모델의 업데이트의 주기를 고려한 설계가 필요

SKT 음성인식 기술 발전 방향

음성 UI가 필요한 신규 Device 의 확산, 인식 대상 범위가 확대되면서 다양한 상황의 새로운 도전을 하게 될 것으로 전망

- **신규 Device 확산** : 원거리 음성인식(다채널 음성전처리), 끝점검출기, 발전된 음향모델
- **인식 대상 범위 확대** : 3rd party Toolkit, 대형 언어모델, 인식기 효율화, Partial & Incremental Decoding

감사합니다