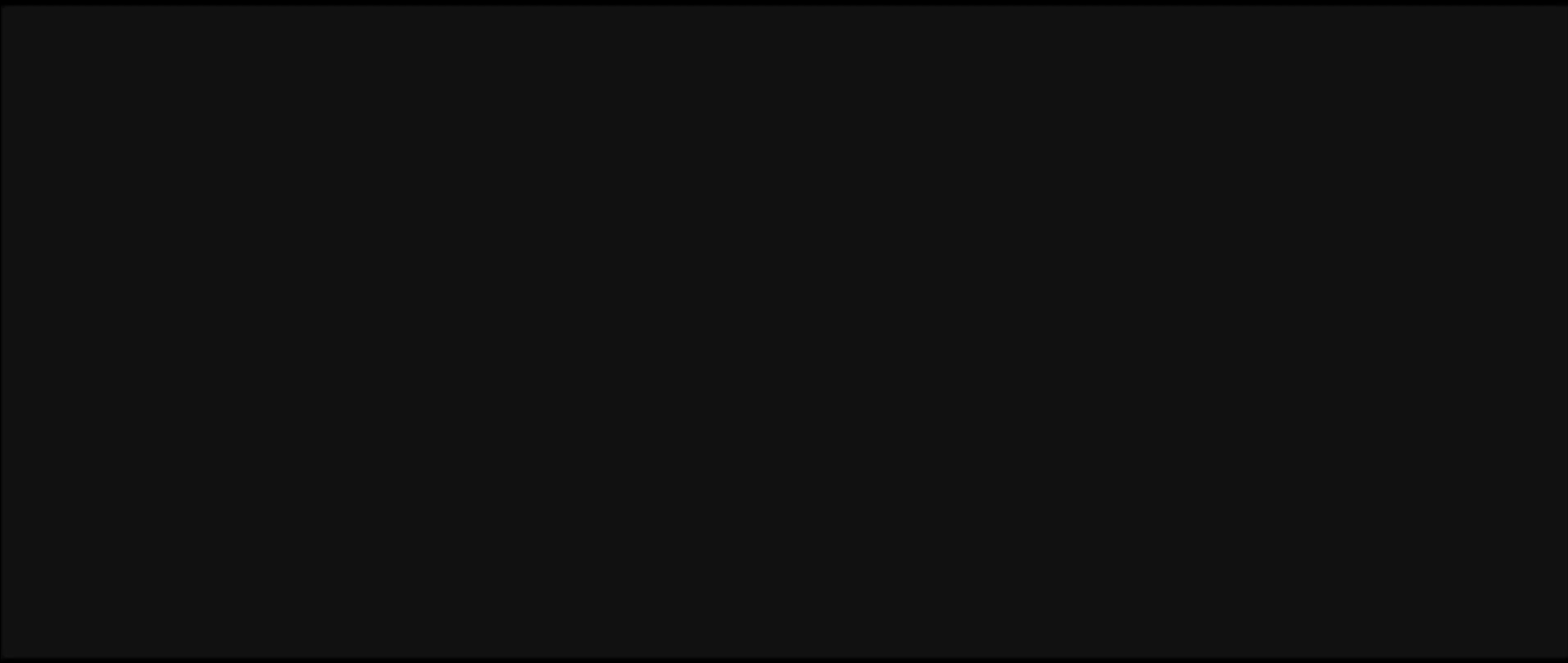
딥러닝을 활용한 실시간 보이스피싱 탐지 서비스

2018. 11. 21

과학기술정보통신부 NIA 한국정보화진흥원 IBK기업은행



Contents

Contents

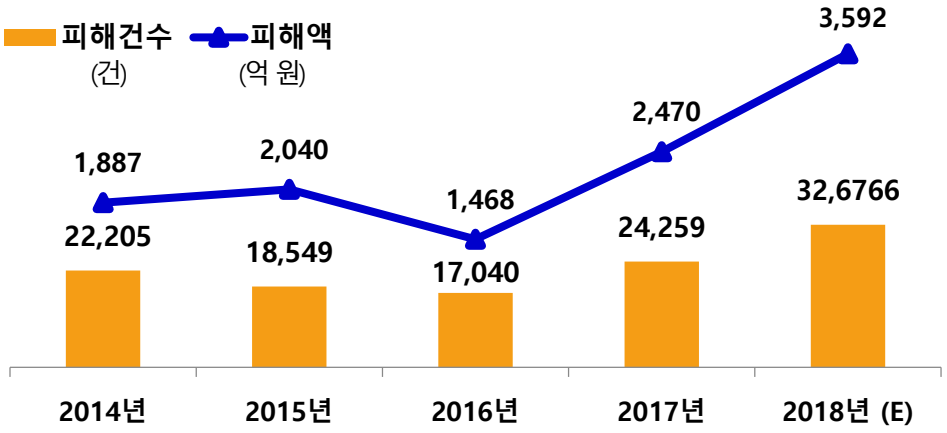
I 사업개요 (연합뉴스TV 2018. 09. 28)

Ⅰ 사업개요

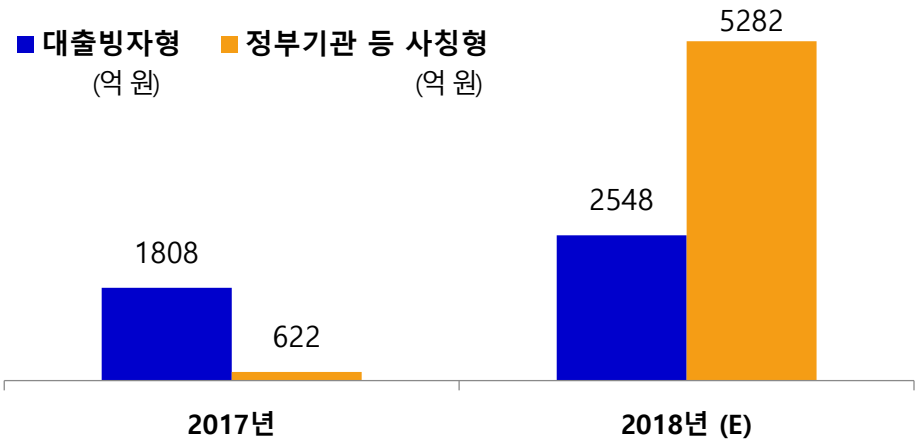
5년간 보이스피싱 피해액은 1 , 10 이며, 116 1 8,600 발생 중

성별로는 여성 피해액이 남성의 2.4배에 달하며, 전 연령대에서 고루 발생한 것으로 조사됨

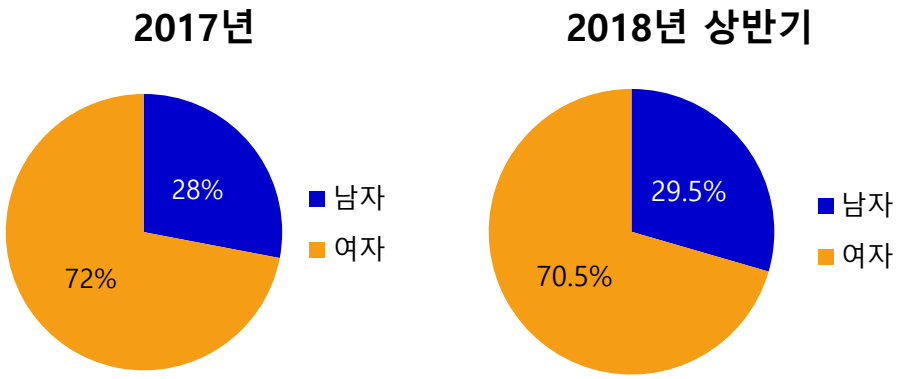
년간 보이스피싱 피해 상황 (출처 : 경찰청)



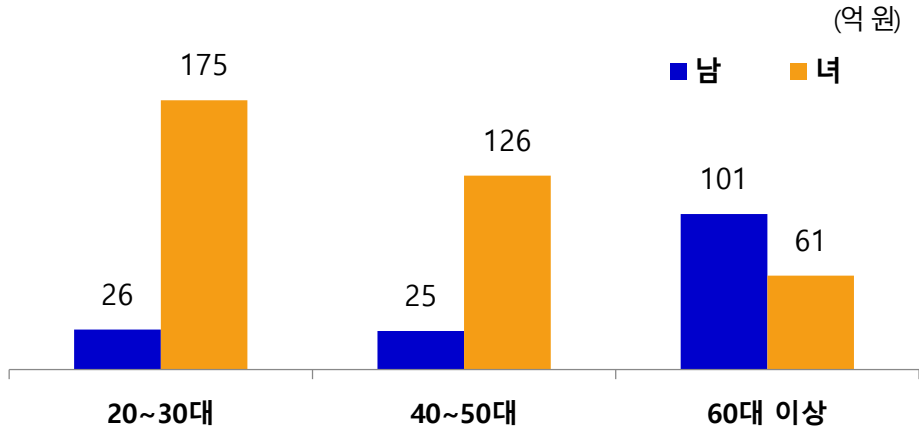
년간 보이스피싱 유형별 피해액 현황 (출처 : 경찰청)



보이스피싱 남/녀 피해 비율 (출처 : 경찰청)



'18년 성별/연령별 보이스피싱 피해액 (출처 : 경찰청)



I 사업개요

현재 통화 App산업은 T전화, 후후, 후스콜 3사가 가장 하고 있음

이는 통화 App 시장에 됨. 또한, 기업 이미지 제고 및 광고효과를 가지고 있음

후스콜

제작사 : gogolook (Naver)

출시일 : 2010. 08

다운로드 수 : 5,500만 건 (2015)

비고 : 전세계 31개국 진출

2016 구글 최고 App 선정

후후

제작사 : KT CS (KT)

출시일 : 2013. 08

다운로드 수 : 3,500만 건 (2017)

비고 : 국내 사용률 압도적 1위

711만명, 66.5% 사용

T전화

제작사 : SK Telecom (SKT)

출시일 : 2014. 02

다운로드 수 : 2,500만 건 (2017)

비고 : 어플 선탑제로 사용자 多

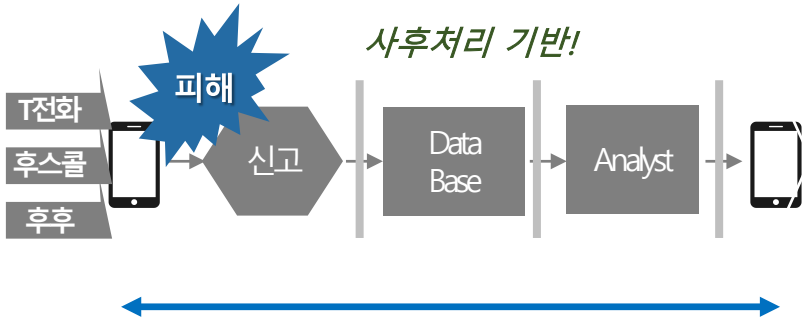
Ⅰ 사업개요

현행 통화 App은 이 맞춰져 있으며, 실시간으로

「IBK 실시간 보이스피싱 AI탐지 서비스」는 음성인식 및 딥러닝 기술을 활용하여 실시간

AS-IS

현재는 등록된 Data 기반으로만 동작하므로
실시간 보이스피싱 탐지 불가



- 사용자 : 전화 수신
- [사후처리] 통화 후 보이스피싱 전화 신고
- Database에 신고내용 기록
- 분석 및 검증 작업 및 배포
- 신고된 번호 탐지 및 알림 (보이스피싱)

TO-BE

음성인식 및 딥러닝 기반 기술을 통한
실시간 보이스피싱 탐지 가능

- 사용자 : 전화 수신
- [실시간] 딥러닝을 통해 보이스피싱 확률 산출
- 딥러닝을 통해 실시간 보이스피싱 동시 400건 산출
- 통화 중, 보이스피싱 의심 경고음 및 진동 알림
- 기계학습을 활용한 보이스피싱 탐지 정확도 향상

Ⅰ 사업개요

「딥러닝을 활용한 보이스피싱 탐지 서비스」를 통해 **매일 변화하는 보이스피싱**으로부터 고객을 보호하고,
기술 고도화를 통해 보이스피싱을 넘어 SMS 스미싱 방지 등 딥러닝 서비스로

고객 금융사기 보호

고객의 안전한 통화와
금융거래 니즈를 만족시킴

딥러닝 기반
「고객중심 서비스」
제공

OneShot 민생사기탐지

보이스피싱 뿐만 아니라
스미싱 또한 탐지

고객

- 사회초년생
- 고소득 전문직
- 고소득 중년
- 기업 CEO/임원
- 장년 직장인
- 중년 직장인
- ⋮

현재

「 실시간 보이스피싱 탐지 서비스」

- **보이스피싱 피해액 감소**
  ※ 보이스피싱 10% 예방 시 242억 원 절감 (출처:금융감독원)
- **공익 서비스 제공으로 안전한 금융거래 환경 조성**

향후과제

「 실시간 보이스피싱 탐지 서비스 고도화」

- **SMS 스미싱 차단 서비스 제공**
- **금감원, 경찰청 사례 제보 및 신고 기능 추가**

Ⅰ 추진계획

IBK, 금융감독원, 한국정보화진흥원 등 협업을 통해「딥러닝을 활용한 보이스피싱 탐지 서비스」 개발 사업의 성공 완수를 목표로 함. 이를 통해 **해외 판로개척** 등 추진 목표를 가짐

IBK

금융감독원

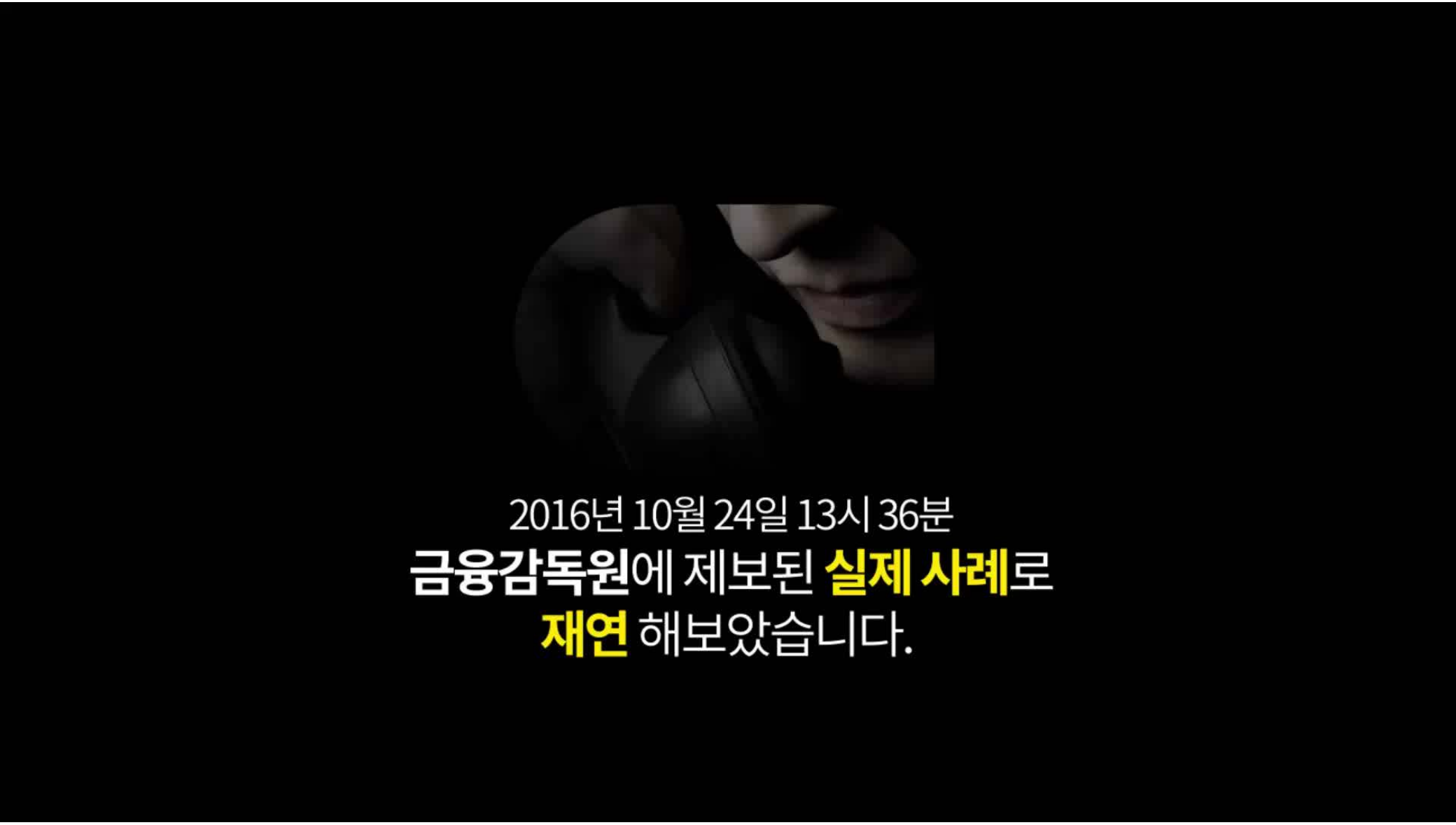
한국정보화 진흥원

「딥러닝을 활용한 보이스피싱 탐지 서비스」 성공적인 사업 완료

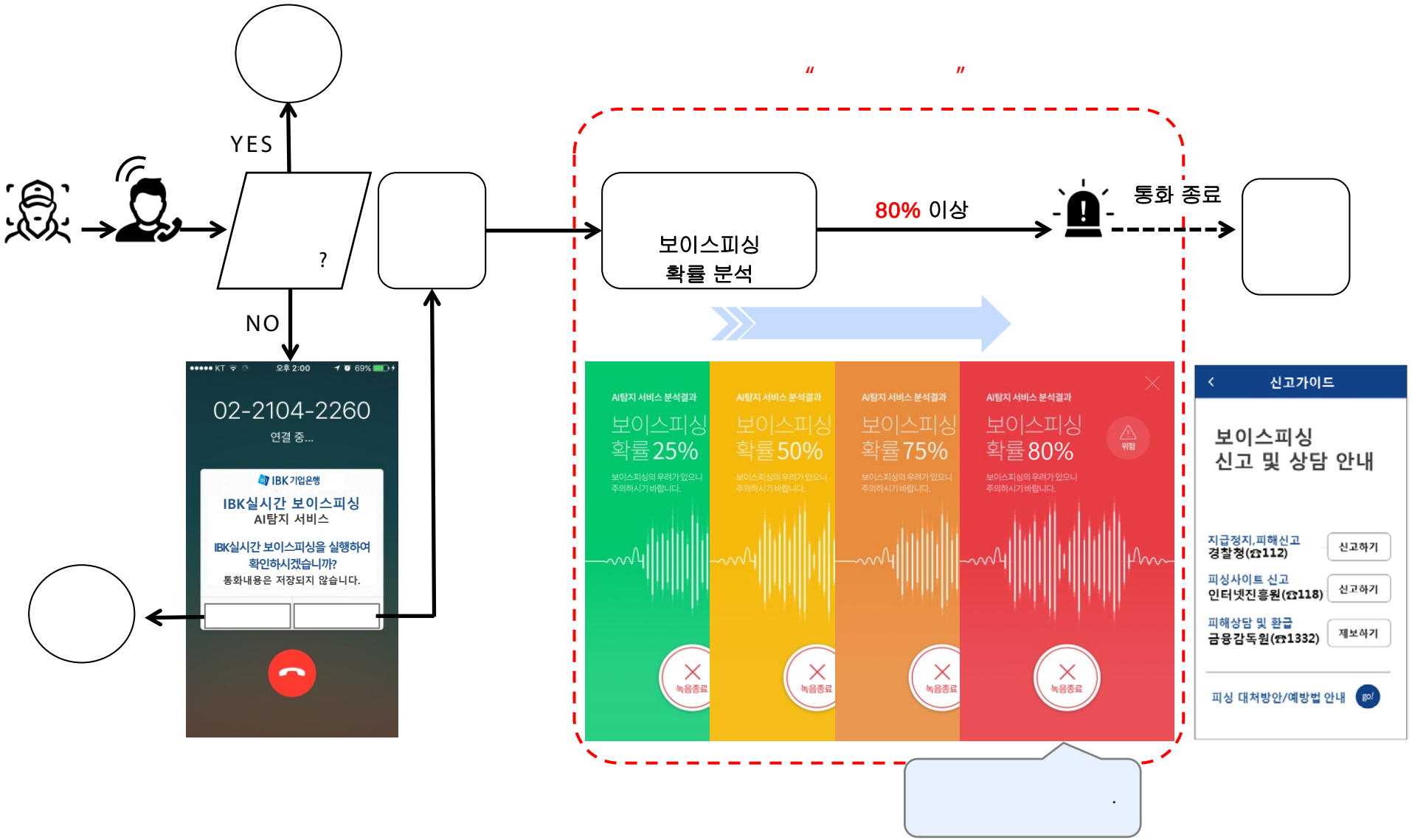
국제 특허출원, 해외 판로 개척 민생사기탐지 Platform

Contents

Ⅱ 시연 영상

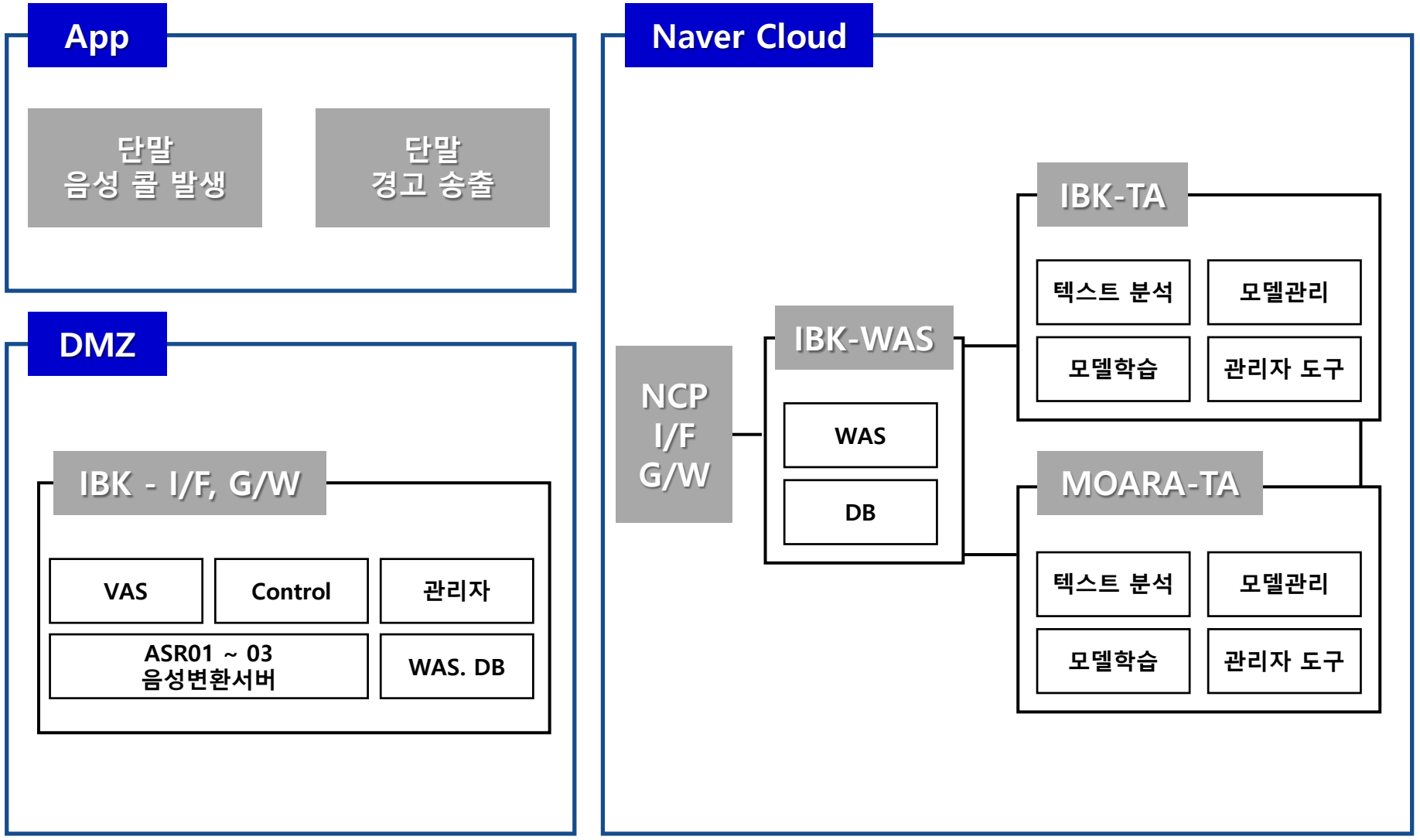


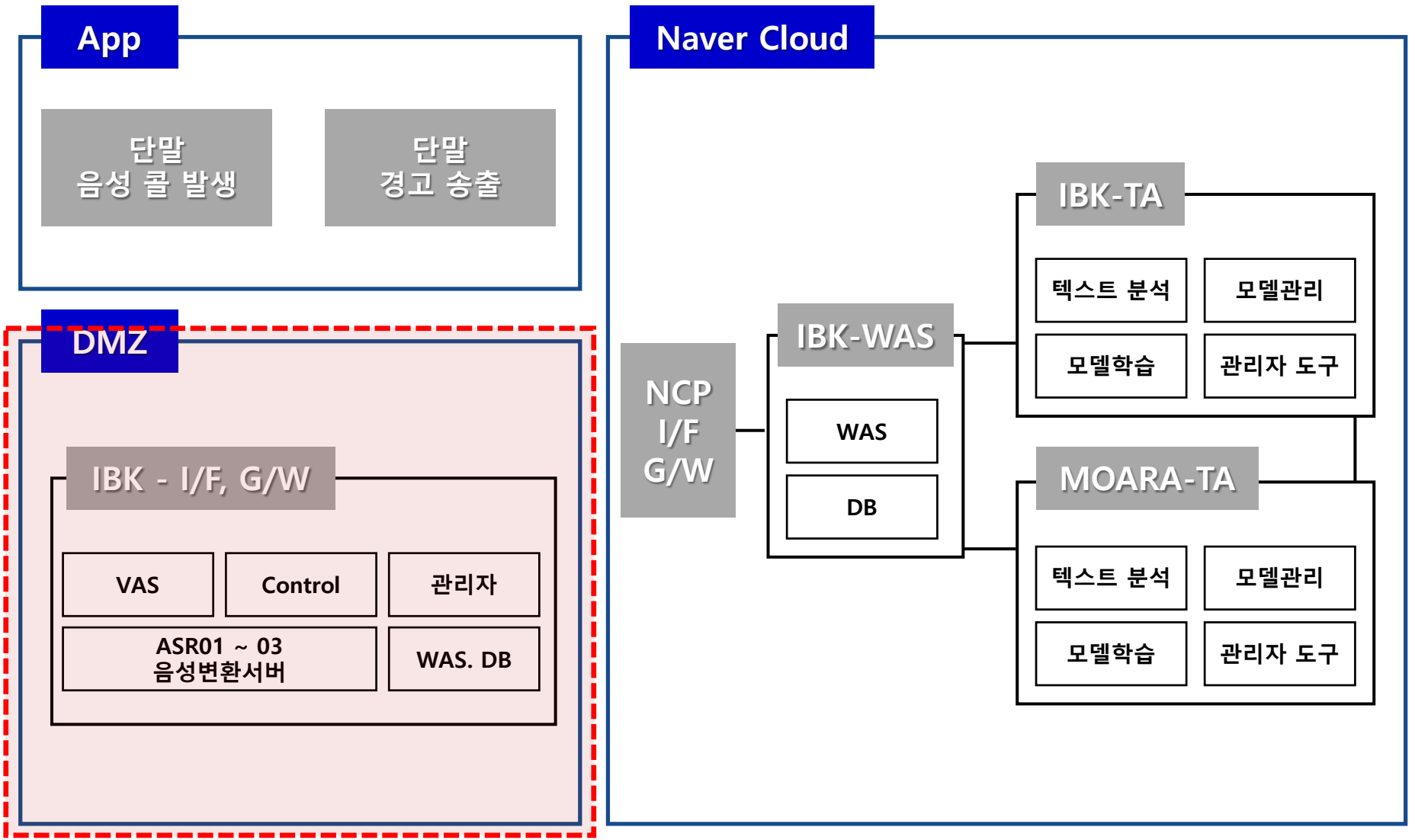
Ⅱ 서비스 프로세스


YES
?
NO
보이스피싱
확률 분석
80% 이상
통화 종료
02-2104-2260
연결 중...
IBK기업은행
IBK실시간 보이스피싱
AI탐지 서비스
IBK실시간 보이스피싱을 실행하여 확인하시겠습니까?
통화내용은 저장되지 않습니다.
AI탐지 서비스 분석결과
보이스피싱 확률25%
보이스피싱 확률50%
보이스피싱 확률75%
보이스피싱 확률80%
녹음종료
신고가이드
보이스피싱 신고 및 상담 안내
지급정지,피해신고 경찰청(☎112)
신고하기
피싱사이트 신고 인터넷진흥원(☎118)
신고하기
피해상담 및 환급 금융감독원(☎1332)
제보하기
피싱 대처방안/예방법 안내

Contents

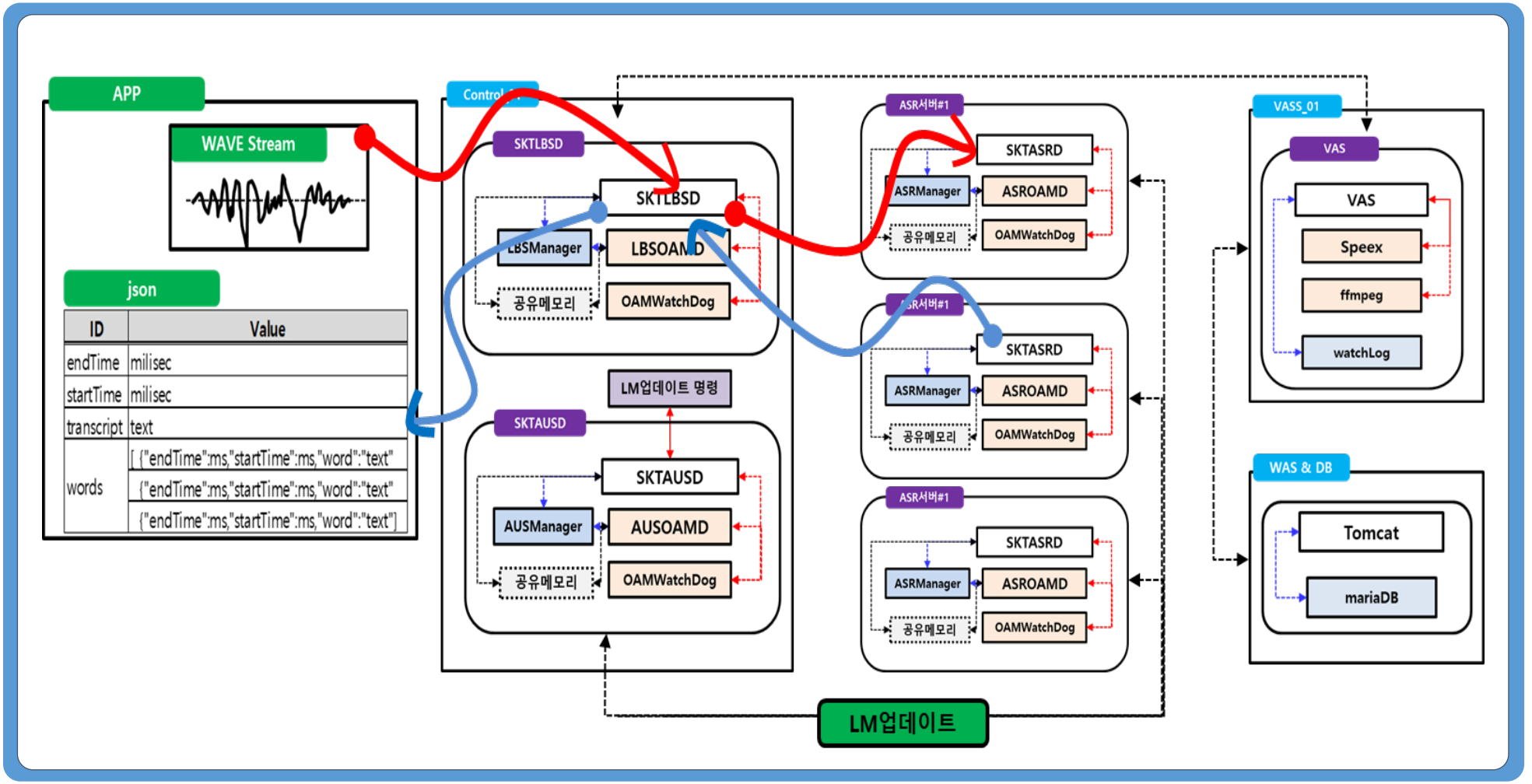
Ⅲ 전체 시스템 구성


App
단말
음성 콜 발생
단말
경고 송출
DMZ
IBK - I/F, G/W
VAS
Control
관리자
ASR01 ~ 03
음성변환서버
WAS. DB
Naver Cloud
NCP
I/F
G/W
IBK-WAS
WAS
DB
IBK-TA
텍스트 분석
모델관리
모델학습
관리자 도구
MOARA-TA
텍스트 분석
모델관리
모델학습
관리자 도구


App
단말
음성 콜 발생
단말
경고 송출
DMZ
IBK - I/F, G/W
VAS
Control
관리자
ASR01 ~ 03
음성변환서버
WAS. DB
Naver Cloud
NCP
I/F
G/W
IBK-WAS
WAS
DB
IBK-TA
텍스트 분석
모델관리
모델학습
관리자 도구
MOARA-TA
텍스트 분석
모델관리
모델학습
관리자 도구

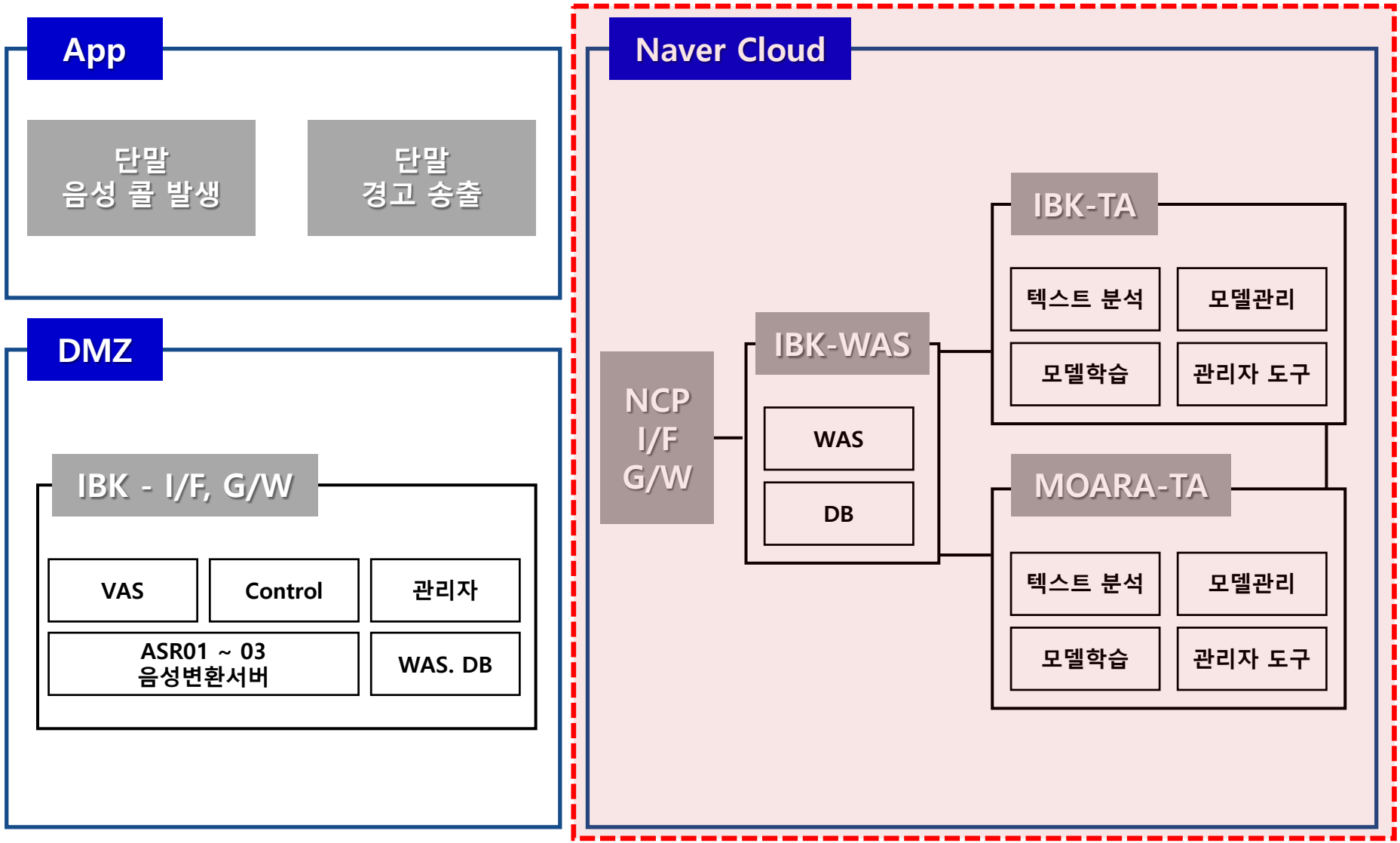
Ⅲ 음성 인식

실시간 STT 시스템 흐름도



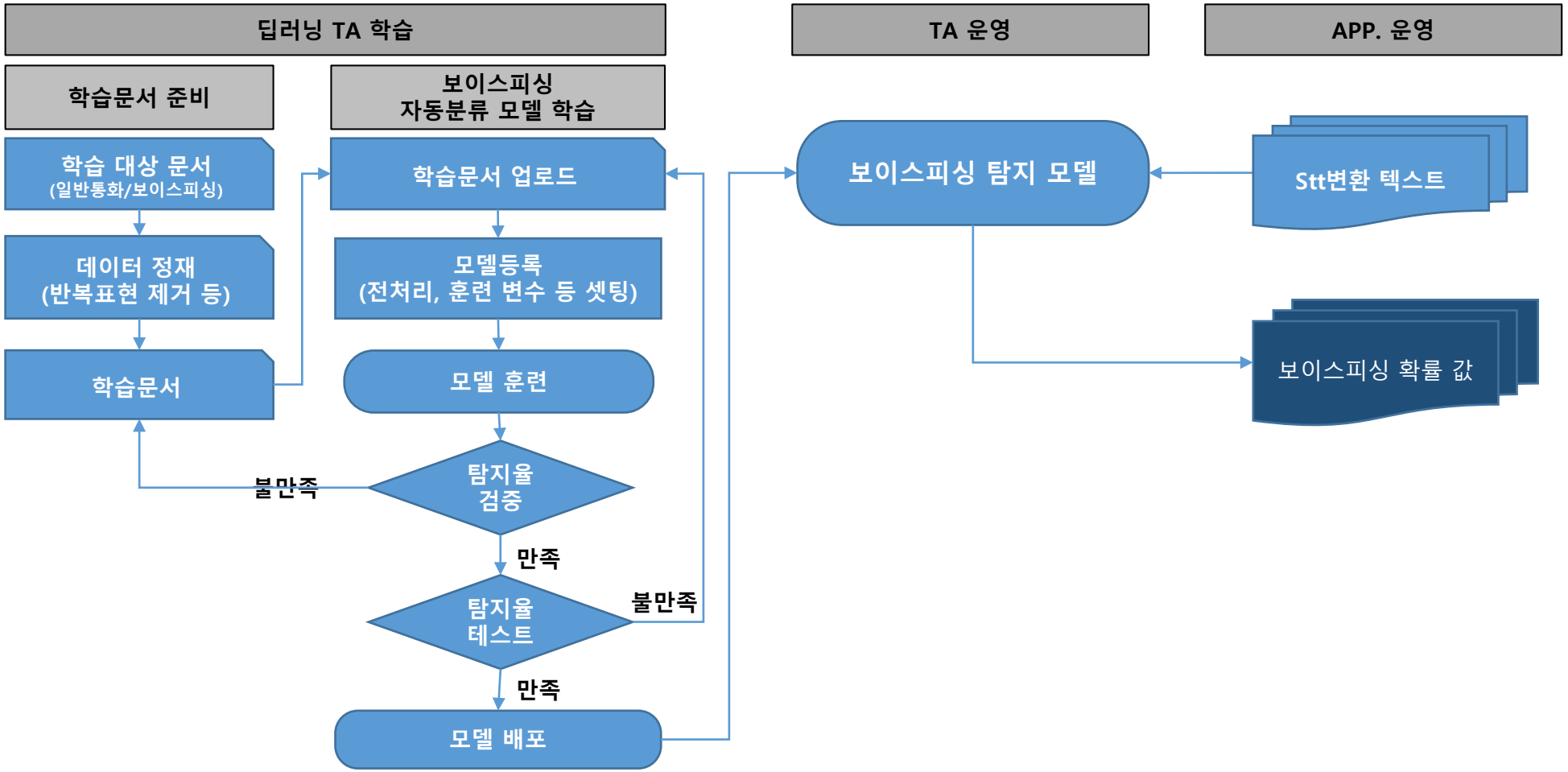
- **머신러닝을 기반으로 생성된 언어모델을 통해 음성을 문자로 실시간 변환**
- **SKT Nugu 음성인식 엔진 활용**

Ⅲ 텍스트 분석(MOARA-TA)



Ⅲ 텍스트 분석(MOARA-TA)

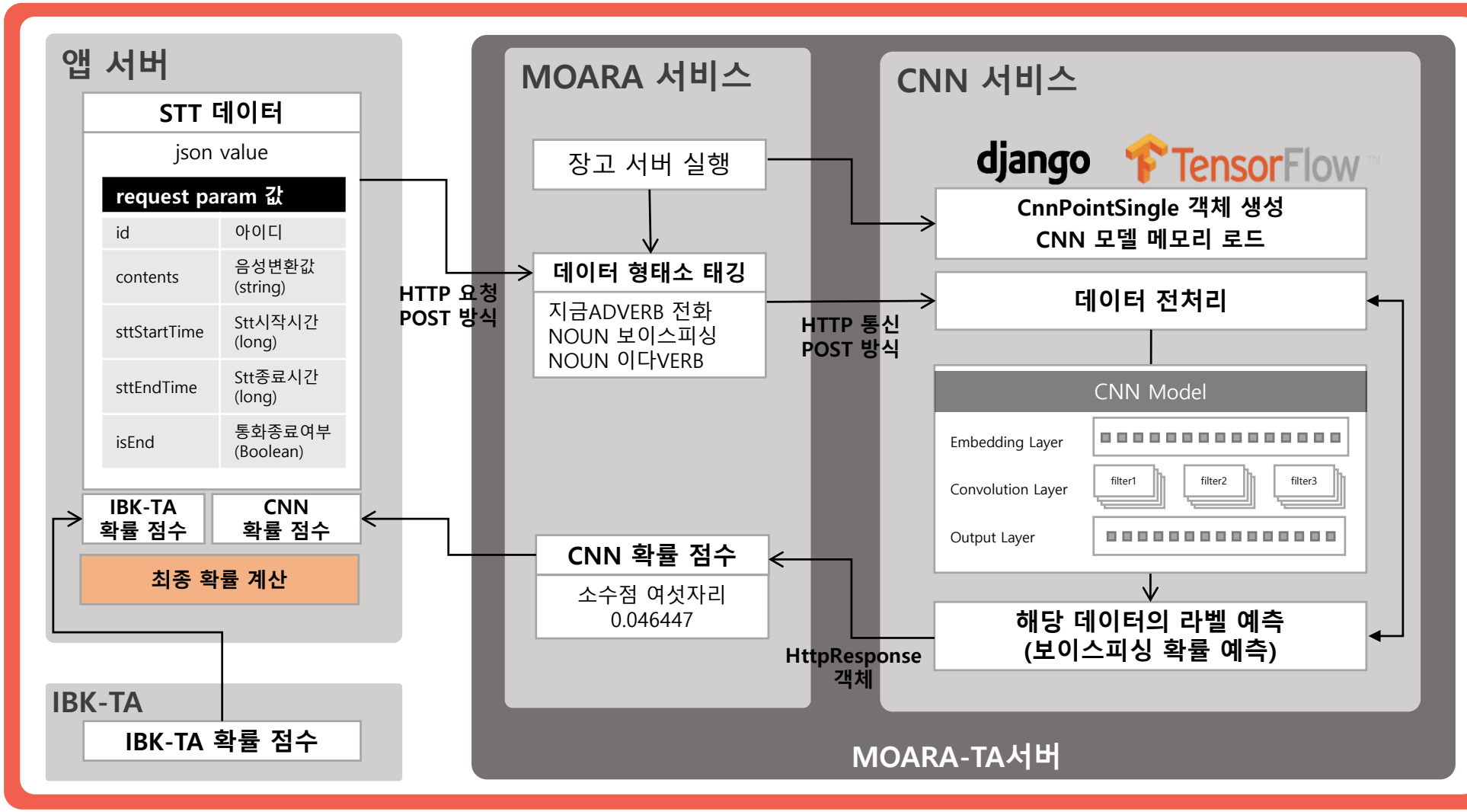
MOARA-TA 딥러닝을 활용한 보이스피싱 탐지 학습 및 운영 흐름도



- **딥러닝을 활용한 보이스 피싱 탐지 모델 개발 (CNN 알고리즘 적용)**
- **딥러닝을 활용하여 보이스피싱 모델을 지속적으로 학습하고 관리 할 수 있는 관리 시스템 개발**

*. CNN : Convolutional Neural Network

MOARA-TA 딥러닝을 활용한 보이스피싱 탐지 시스템 흐름도



Ⅲ 텍스트 분석(MOARA-TA)

MOARA-TA 딥러닝(CNN)을 활용한 보이스피싱 탐지 모델 학습 결과 (4차 ~ 6차 학습)

Test Data 글자수	2000자	3000자	4000자
200	78.79%	85.02%	85.86%
400	83.84%	89.56%	90.07%
600	86.53%	91.08%	90.24%
800	85.69%	91.58%	91.58%
1000	87.37%	92.42%	91.75%
1200	87.21%	93.10%	92.42%
1400	87.71%	93.10%	92.59%
1600	88.22%	93.43%	92.59%
1800	88.89%	93.77%	92.76%
2000	88.55%	93.60%	93.10%
2200	88.55%	93.94%	93.43%
2400	88.55%	94.11%	93.43%
2600	88.55%	94.11%	93.43%
2800	88.55%	94.11%	93.60%
3000	88.55%	94.28%	93.60%
3200	88.55%	94.11%	93.43%
3400	88.55%	94.11%	93.43%
3600	88.55%	93.94%	93.43%
3800	88.55%	93.94%	93.60%
4000	88.55%	93.94%	93.77%

<글자수 별 탐지 모델 정확도>

훈련 데이터 글자수	전체 정확도
2000	85.35%
3000	94.50%
4000	90.21%

<탐지 모델별 전체(모든글자) 정확도>

탐지단계	적용 보이스피싱 확률
안전	~ 34.99 %
보통	35 ~ 49.99 %
위험	50 ~ 69.99 %
경고	70 % ~

<탐지 단계별 확률 적용 범위>

- 2,000 ~ 4,000 자 모델로 학습하여 탐지율을 비교한 결과 3,000자 모델이 가장 정확 하였음
- 3,000자 모델을 기준으로 탐지 단계별 보이스피싱 확률 적용

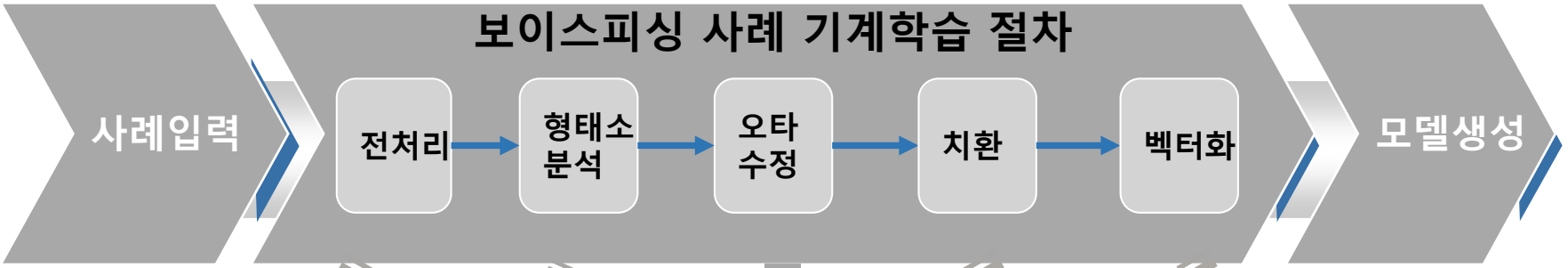
Doc2Vec, Jaro-Wrinkler 문서간 유사도 검색

- 전처리, 형태소분석, 오타변환, 치환, 벡터화

- 기계학습된 기존 보이스피싱 스크립트와 유사도 판단

1. Cos 유사도 산출
   → 기존 학습된 스크립트의 벡터와 보이스피싱 통화의 벡터와의 내적을 구하여 얼마나 가까운지(유사한지) 산출
2. JaroWinkler 알고리즘 이용 유사도 산출
   → 문장 內 단어들의 위치를 전이시켜 일치하는 정도를 파악하는 문자열 유사도 판단 알고리즘
3. 최종유사도 판단
   → Cos유사도 * 0.3 + JaroWinkler * 0.7 으로 유사도 판단
   (학습데이터의 양이 적어, Cos 유사도에 적은 가중치 부여)

Ⅲ 텍스트 분석 (IBK-TA)



상세내용

- 특수문자 제거

(ex. △, ★, & 등 ASCII 코드 기준 영문, 한글, 숫자를 제외한 값 제거)

- 형태소 분석기를 통하여 주어,동사 ,명사, 접속사등 형태소 분석 후 문구/태그 방식 문장변경

(ex. 나는 밥을 먹었다 → 나/S는/A 밥/N을/A 먹었/R다/A)

- 명사 사전에 저장되어 있지 않아 명사로 인식하지 못한 경우

1) 오타 사전에서 오타 여부 확인하여 치환
2) 없는 경우, 문자열 거리계산 알고리즘(Levenstein)이용하여 기존 명사와 가까운 단어선택 (ex. 빱 → 밥)

- 문장 內 핵심 단어사전, 기관 단어사전에 있는 단어가 등장 할 경우 형태소 치환

(ex. 대포통장**/n** → 대포통장**/hr** 금감원**/n** → 금감원**/pl**)

* 유사도 판단 및 벡터화 시 핵심단어 및 기관명에 가중치 부여를 위함

- Word2Vec의 Doc2Vec 기법을 이용하여 보이스피싱 스크립트 전체 벡터화 후 모델 생성

Jaro-Wrinkler 알고리즘 고도화

The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

where:

- m is the number of *matching characters* (see below);
- t is half the number of *transpositions* (see below).

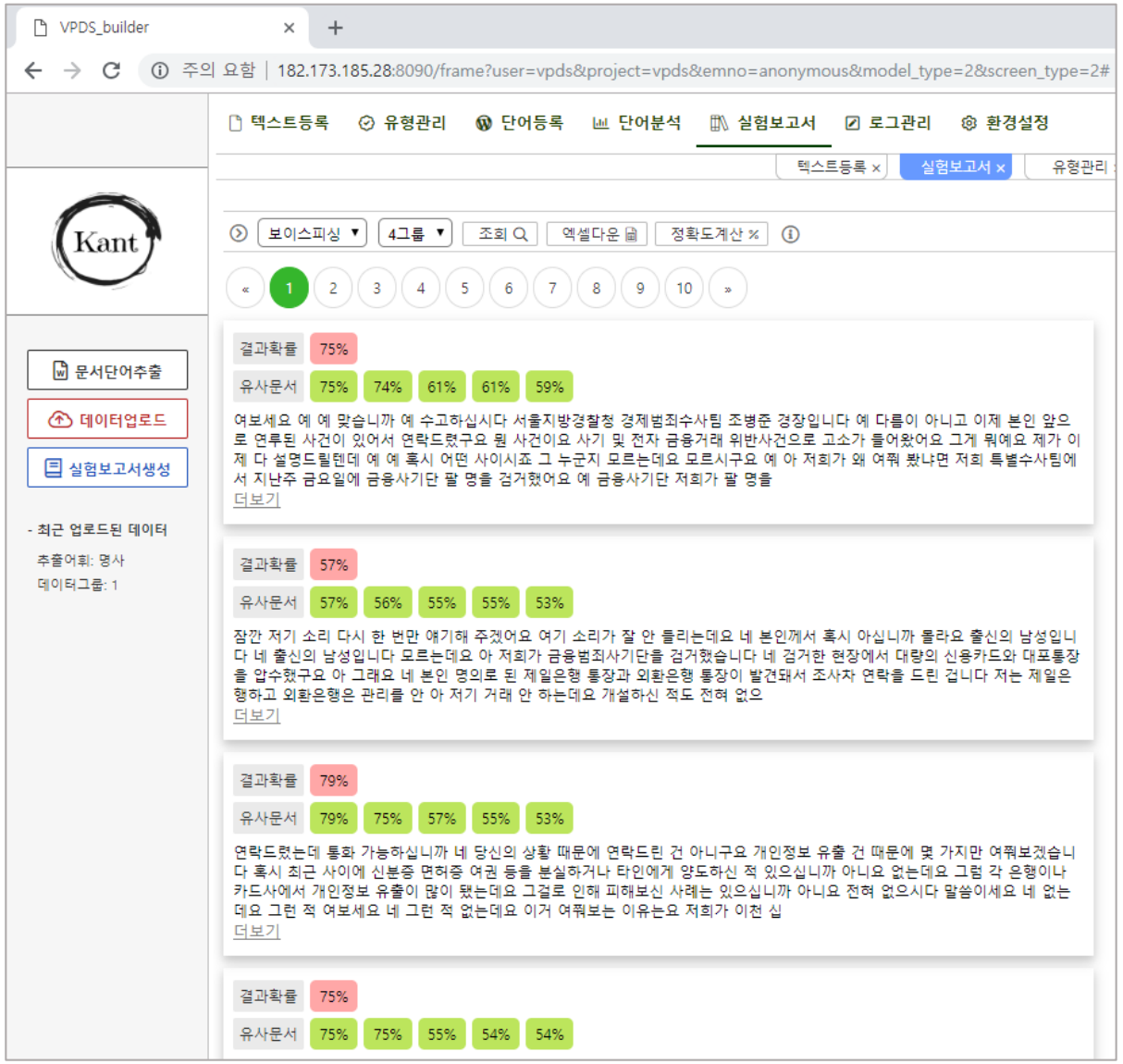
변경사항

1. s1, s2 (sentence) → d1, d2 (document) 로 변형
2. m, d1, d2 변수에 각 단어 별 가중치 부여 (가중치가 높을수록 높은 비율 차지)
3. 단어사전에 등록된 단어일 경우에만 t (transpositions) 에 포함

→ 변경전 대비 정확도 15% 향상

Ⅲ 텍스트 분석 (IBK-TA)

Jaro-Wrinkler 모델 실험보고서



보이스피싱 데이터 입력 시

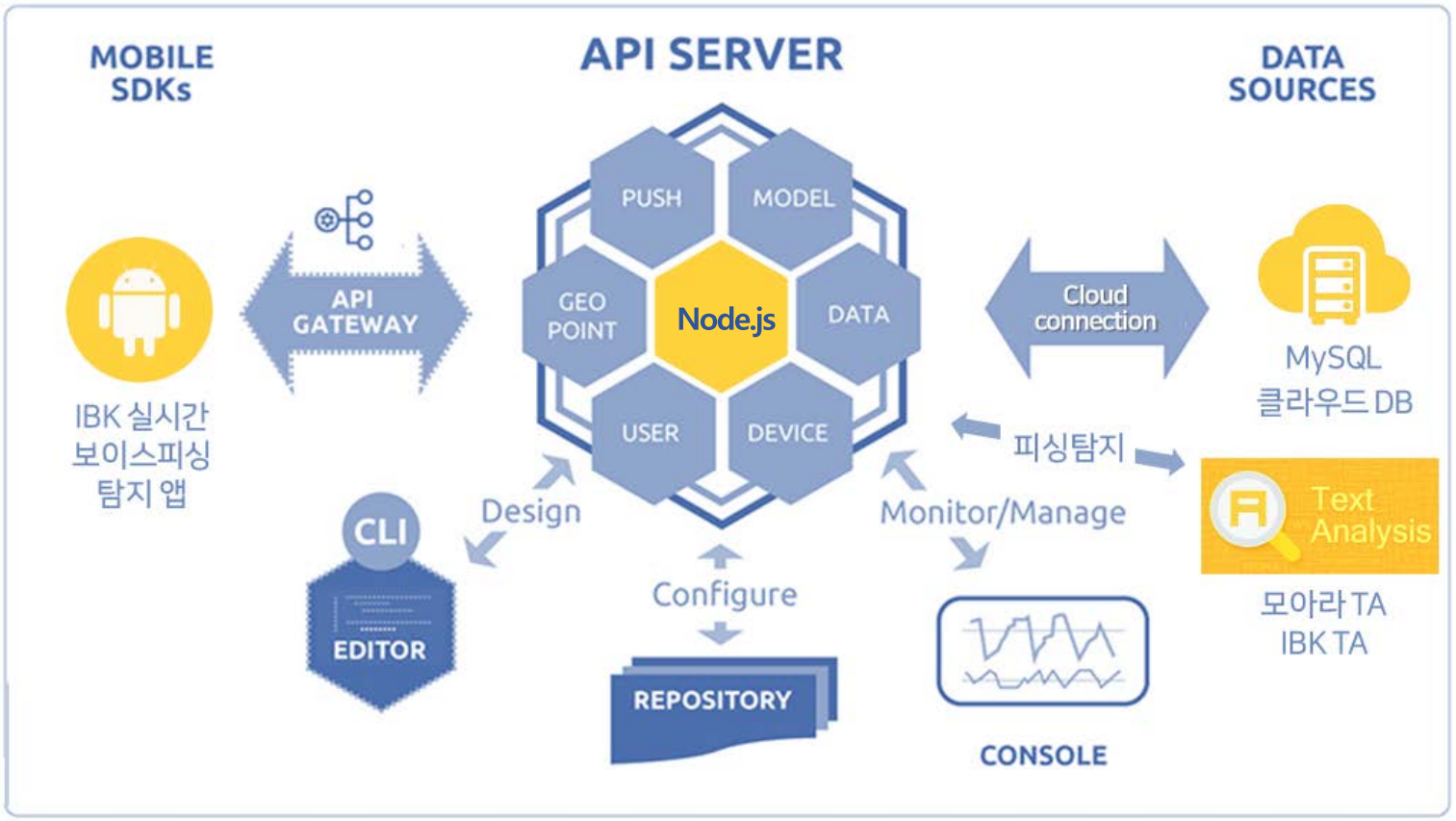
확률(%)	문서 수	비율
0~35	**3**	**4%**
35~50	**2**	**2%**
50~70	**49**	**66%**
70~	**20**	**27%**

일반통화 데이터 입력 시

확률(%)	문서 수	비율
0~35	**117**	**61%**
35~50	**50**	**26%**
50~70	**22**	**11%**
70~	**0**	**0%**

- 종합적으로 88~90% 정확도 달성

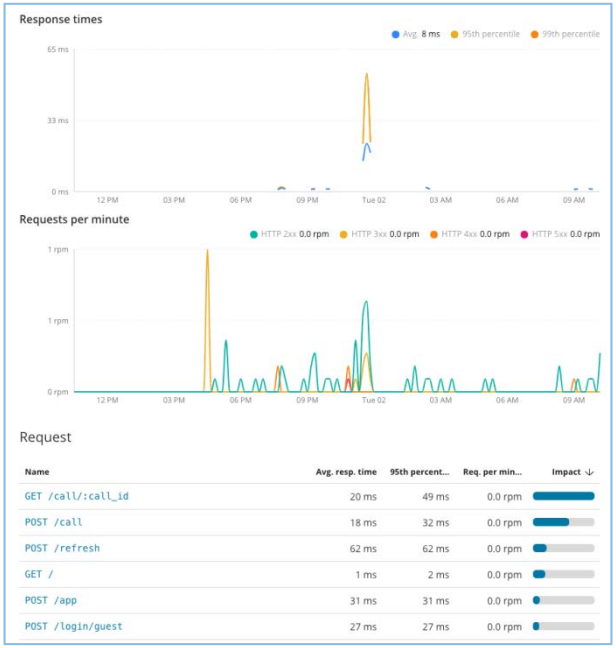
수행 내용 : 소프트웨어 아키텍처



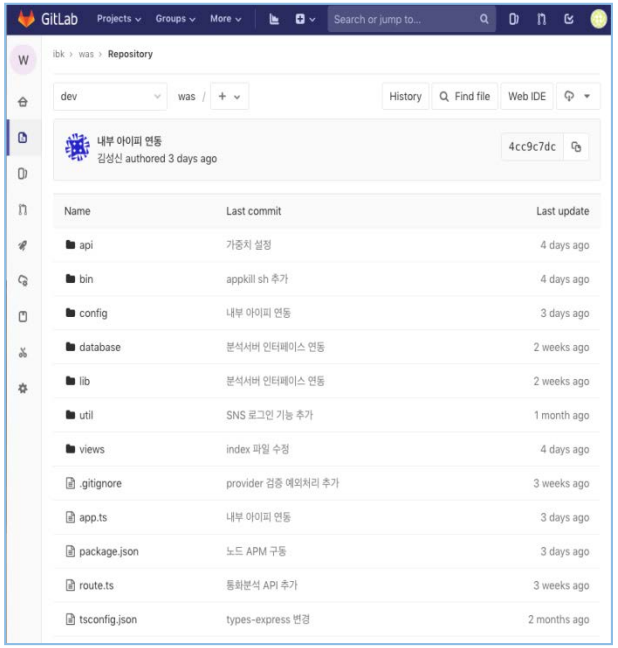
"성능 극대화를 위한 비동기(Async) 아키텍처"

Ⅲ 클라우드 서비스

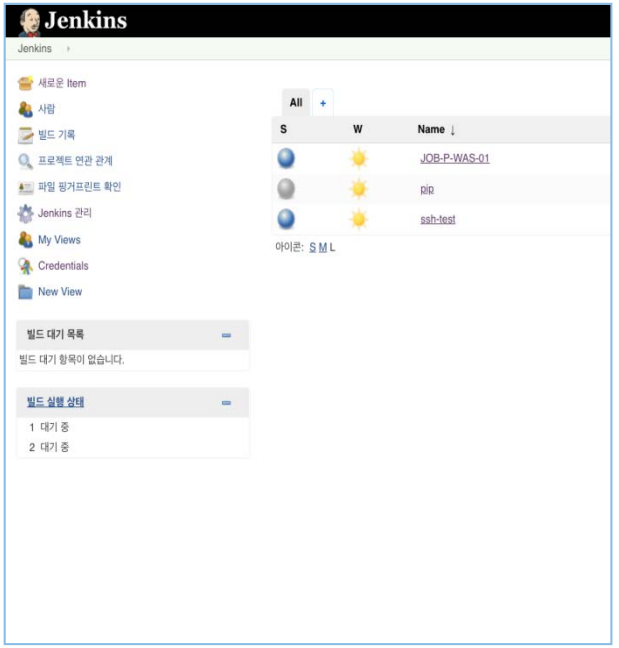
수행 내용 : 운영관리 소프트웨어



실시간 거래성능
모니터링



앱, AP, TA, ML
통합 형상관리



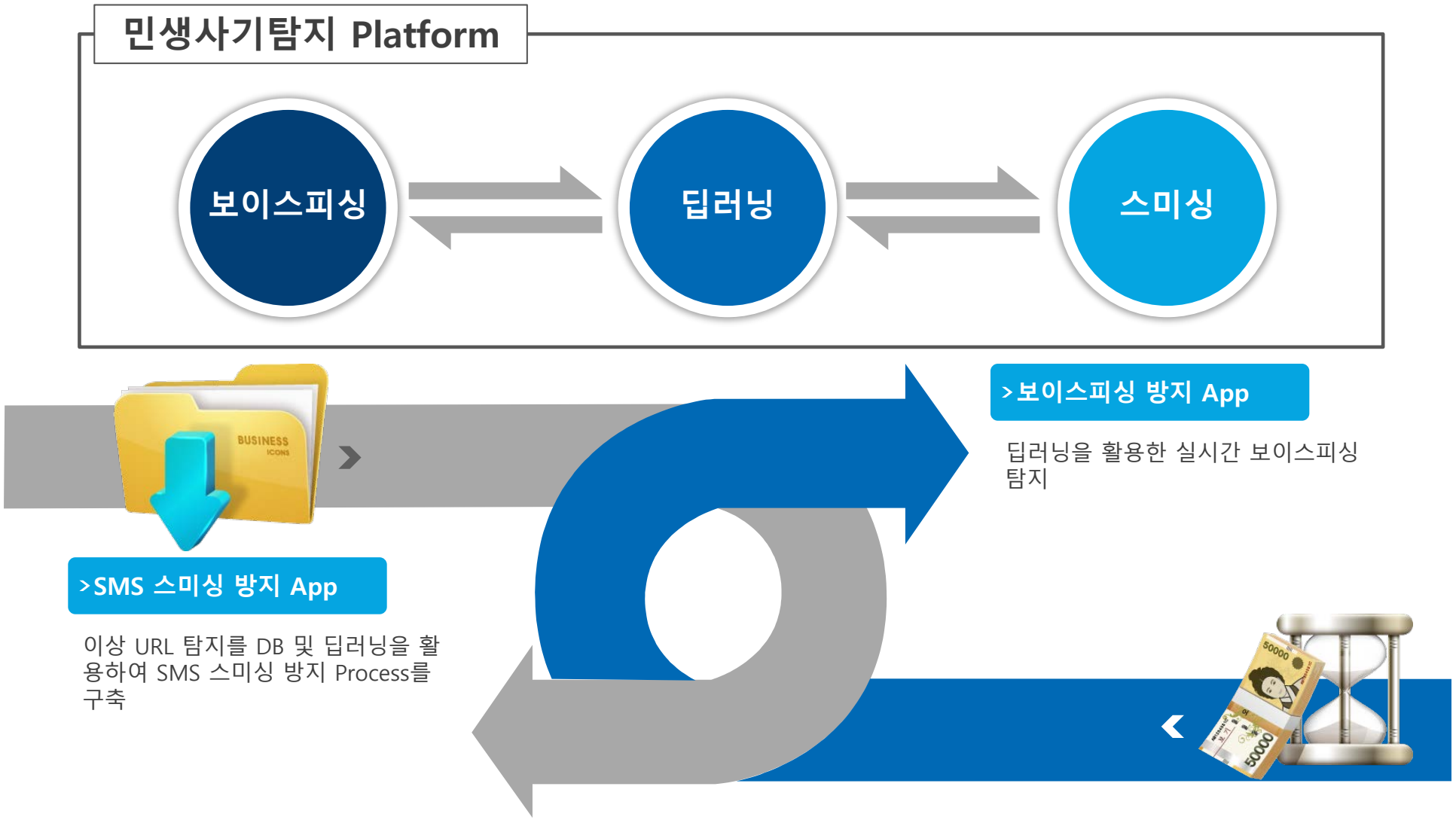
빌드 및 배포 관리
자동화

"클라우드 모니터링, 형상관리 및 빌드/배포 자동화 구현"

Contents

IV 향후 발전 방향

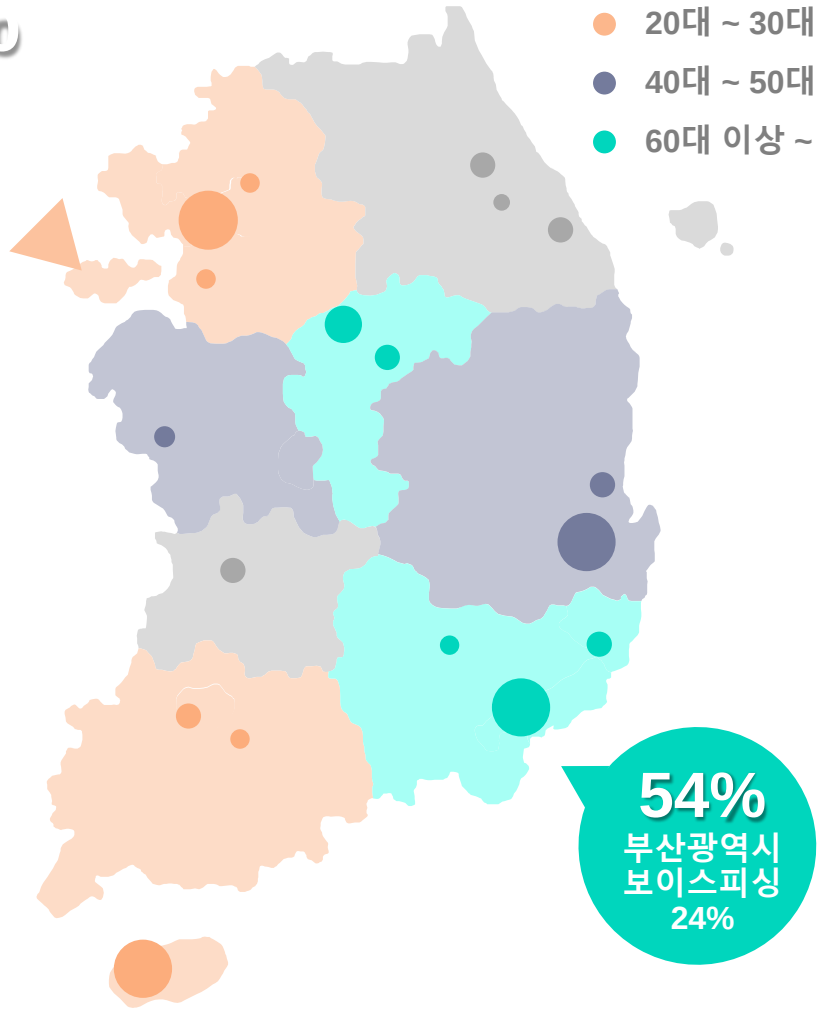
서비스 고도화로 실시간 보이스피싱 탐지를 넘어, SMS 스미싱까지 개발하여 Platform으로 발전



IV 향후 발전 방향

, , 등 통계를 활용하여 서비스 개발 및 제공

82%



01 사용자 연령대별 보이스피싱 방지 대안 마련

02 지역별 보이스피싱 발생 확률을 통해 방지 대안 마련

03 직업군별 분석하여 그에 맞는 보이스피싱 방지 대안 마련

Contents

Q&A

감사합니다

2018. 11. 21

과학기술정보통신부 NIA 한국정보화진흥원 IBK기업은행