

# 2025 가명정보 활용 경진대회

가명정보 처리 가이드라인에 기반한  
정형 및 비정형(음성) 심리 상담 데이터의 가명처리  
흑석동 물주먹

총괄주최 |



개인정보보호위원회



과학기술정보통신부

공동주최 |



보건복지부



금융위원회

주관 |



한국인터넷진흥원

후원 |



건강보험심사평가원



국민건강보험  
National Health Insurance Corporation



금융결제원



금융보안원  
FINANCIAL SECURITY INSTITUTE

DOUZONE



한국신용정보원  
Korea Credit Information Services



한국데이터산업진흥원



한국사회보장정보원



한국지능정보사회진흥원



한국지역정보개발원

# Contents

**1 가명처리 방향성**

**2 정형 데이터 가명처리**

**3 비정형(음성) 데이터 가명처리**

**4 결론**

# 가명처리 방향성



## 가명처리 방향성



### 1. 시나리오 목적을 고려한 가명처리 기법

상담 주기/빈도/내용을 통한 효과 분석 및 상담 가이드 제작



### 2. STT시 온프레미스 환경 가정

AI 모델 적용시 개인정보 유출 및 탈취 우려 고려

✓ 정형 데이터 + 비정형(음성) 데이터 결합 시에 식별 가능 부분 최소화 + 데이터 유용성 최대화

# 정형 데이터 가명처리



## 항목별 개인정보유형 및 가명처리 유무

순번	항목명	개인정보유형	가명처리 유무
1	이름	개인식별정보	O
2	주소	개인식별가능정보	O
3	성별	개인식별가능정보	X
4	연령	개인식별가능정보	O
5	생년월일	개인식별가능정보	O
6	예정 면담 회차	개인식별가능정보	X
7	실제 면담 회차	개인식별가능정보	X
8, 9, 10, 11, 12	1차/2차/3차/4차/5차 면담일	개인식별가능정보	O
13	면담 종료 코드	개인식별가능정보	X

\*분석 목적: 상담 주기 및 상담 내용에 따른 상담 효과 분석

## 가명처리 미적용 이유

- 성별  
→ 비슷한 비율의 낮은 식별 가능성  
 $M : F = 57,696 : 42,304$
- 예정 면담 회차, 실제 면담 회차, 면담 종료 코드  
→ 원본 형태를 유지하는 것이 분석 목적\*에 적합하다 판단

# 정형 데이터 가명처리



## 이름

1. 마스킹: 성을 제외한 이름을 '00'으로 대체

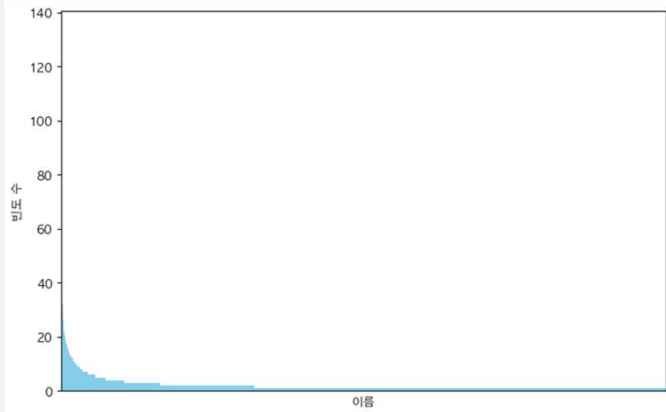
2. 치환:

2.1) 빈도 수가 낮은 성 선택 (전체 데이터 크기의 5%(1,000개) 이하; 114개 카테고리)

2.2) 빈도 수가 낮은 성을 빈도 수가 높은 5개의 성(김, 이, 박, 최, 정) 중 랜덤 치환

### [처리 전]

이름
김민준
김승주
...
방인하

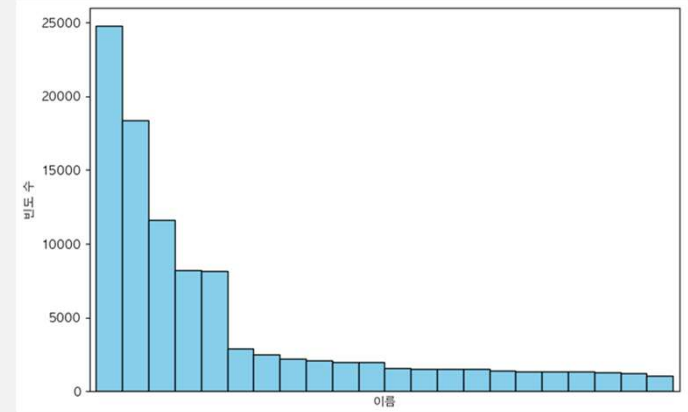


48,060개 카테고리

- 최대 빈도 수: 134
- 최소 빈도 수: 1 (동일 최소값 개수: 32,711)

### [처리 후]

이름
김00
김00
...
방00



22개 카테고리

- 최대 빈도 수: 24,747
- 최소 빈도 수: 1,086

# 정형 데이터 가명처리



## 주소

### 1. 부분삭제: 시 · 군 · 구 단위 이하 삭제

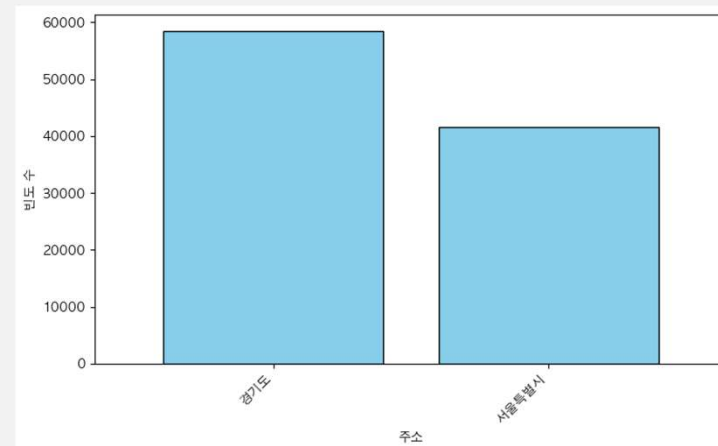
#### [처리 전]

주소
서울특별시 동대문구 무학로 49길 43
경기도 파주시 적성면 국사로 119-40
...
서울특별시 중랑구 동일로 161길 8-8

100,000개 카테고리

#### [처리 후]

주소
서울특별시
경기도
...
서울특별시



2개 카테고리

- 경기도: 58,424
- 서울특별시: 41,576

## 정형 데이터 가명처리



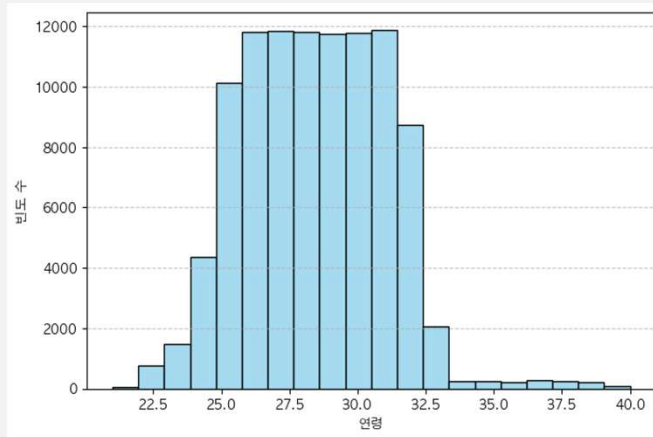
## 연령

## 1. 범위 방법

- 25세 미만: [0, 25)
- 25세 이상 30세 미만: [25, 30)
- 30세 이상 35세 미만: [30, 35)
- 35세 이상: [35, 100)

## [처리 전]

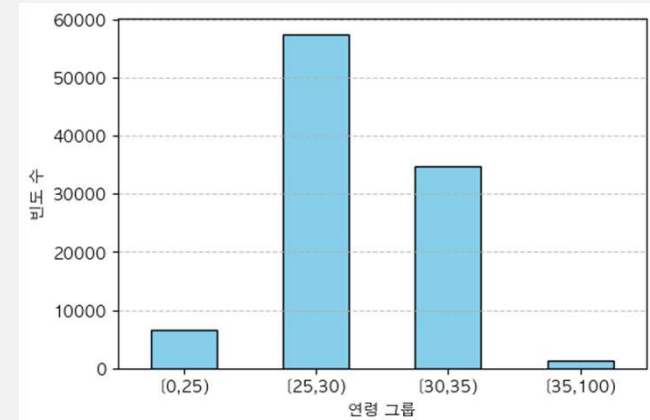
연령
27
23
...
31



- 평균: 28.336
- 표준편차: 2.74
- 최댓값: 40 (빈도 수: 72)
- 최솟값: 21 (빈도 수: 54)

## [처리 후]

연령
[25,30)
[0,25)
...
[30,35)



## 4개 카테고리

- 최대 빈도 수: 57,347
- 최소 빈도 수: 1,275

# 정형 데이터 가명처리

## 💡 생년월일

1. 부분삭제: 출생 월과 일은 삭제

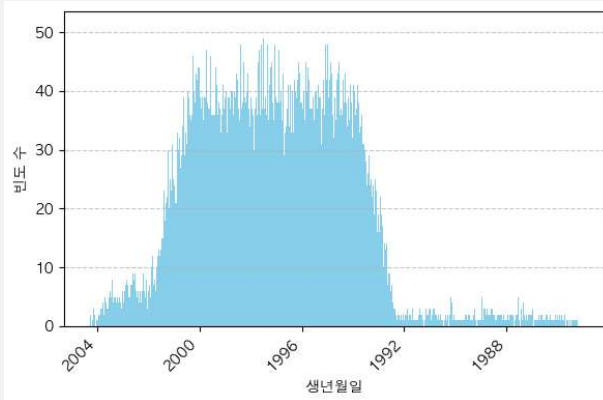
2. 범위방법

- 2000년 초과 2005년 이하: (2000, 2005]
- 1990년 초과 1995년 이하: (1990, 1995]

- 1995년 초과 2000년 이하: (1995, 2000]
- 1990년 이하: (0, 1990]

### [처리 전]

생년월일
1998.12.16
2002.12.10
...
1994.4.28

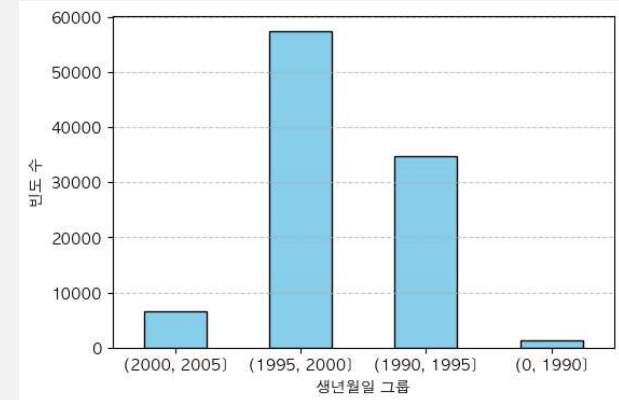


5,447개 카테고리

- 최대 빈도 수: 51 (동일 최댓값 개수: 2개)
- 최소 빈도 수: 1 (동일 최솟값 개수: 1,027개)

### [처리 후]

생년월일
(1995, 2000]
(2000, 2005]
...
(1990, 1995]



4개 카테고리

- 최대 빈도 수: 57,347
- 최소 빈도 수: 1,275



# 정형 데이터 가명처리



## 1차 / 2차 / 3차 / 4차 / 5차 면담일

시나리오 목적인 '상담 주기' 반영

### 1. 1차 면담일: 난수 생성기

→ 2025년 1월 1일부터 2025년 8월 1일 전까지의 날짜들 중 랜덤값 적용

### 2. 2차 / 3차 / 4차 / 5차 면담일: 잡음추가

→ 기존 1차 면담일과의 동일한 일자수를 고려한 잡음 추가.

### [처리 전]

1차	2차	3차	4차	5차
2025.2.6 (기준 일)	2025.2.16 (+10일)	2025.2.24 (+18일)	2025.3.7 (+29일)	2025.3.31 (+53일)
...	...	...	...	...
2025.2.3 (기준 일)	2025.2.12 (+9일)	2025.2.19 (+16일)	2025.3.2 (+27일)	2025.4.9 (+65일)

### [처리 후]

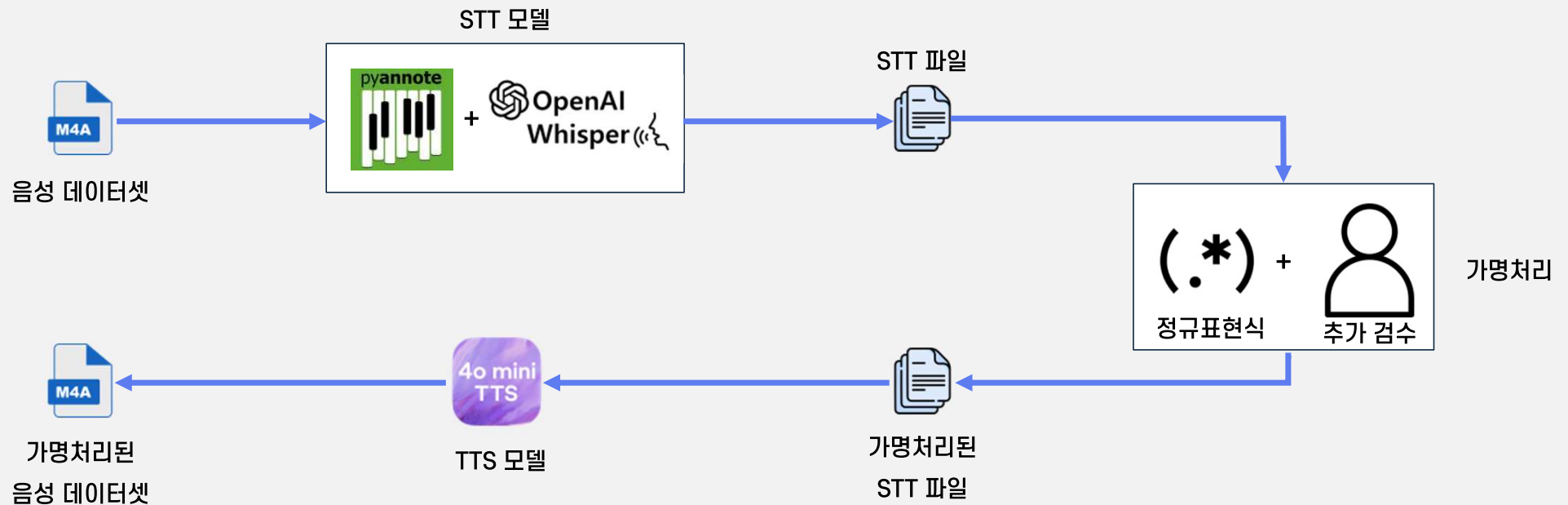
1차	2차	3차	4차	5차
2025.1.11 (기준 일)	2025.1.21 (+10일)	2025.1.29 (+18일)	2025.2.9 (+29일)	2025.3.5 (+53일)
...	...	...	...	...
2025.1.10 (기준 일)	2025.1.19 (+9일)	2025.1.26 (+16일)	2025.2.6 (+27일)	2025.3.16 (+65일)

# 비정형(음성) 데이터 가명처리



## 비정형(음성) 데이터 가명처리 프레임워크

STT-TTS ⇒ 추가 검수 과정에서의 시간 단축, 가명처리한 데이터가 문맥을 더 잘 살릴 수 있는 방향(AI 모델에 활용)



# 비정형(음성) 데이터 가명처리

## 💡 STT 시 주의할 점

### STT API Service



NAVER CLOVA Speech



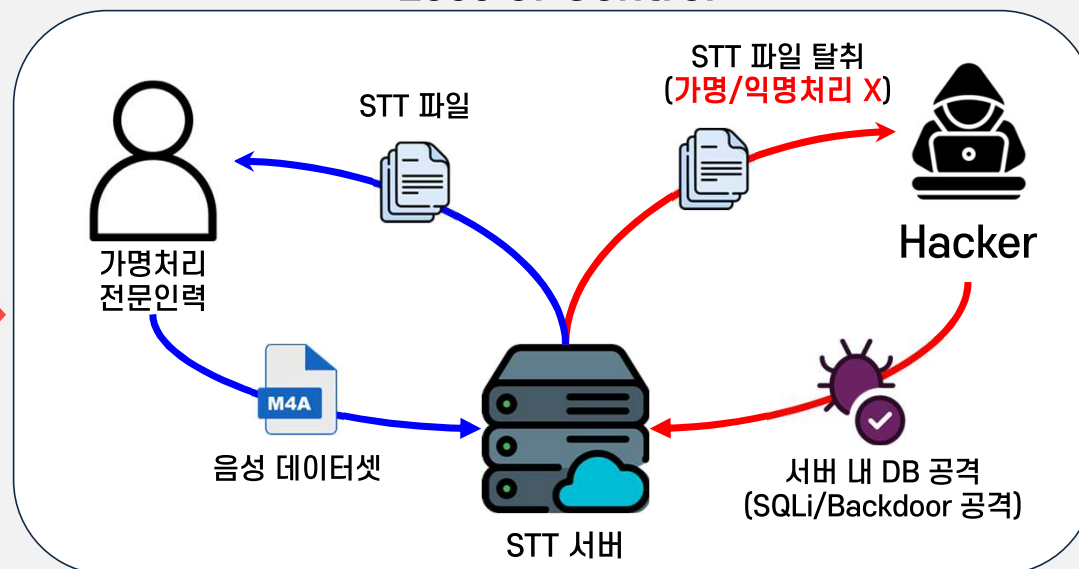
MS Azure Speech



Google Cloud Speech API



### 가명처리할 데이터에 대한 “Loss of Control”



✓ STT 시, 외부 API 호출 없이 로컬 AI 모델로 처리하는 것이 안전  
⇒ **온프레미스 환경**에서 수행

# 비정형(음성) 데이터 가명처리

## 💡 STT (Speech To Text)



Pyannote-audio  
: 화자 분리(Speaker Diarization)

- Local 환경에서 운영 가능
- 2인 이상(3인, 4인, ...) 화자 인식 및 분리 가능
- 소리가 겹치는 구간도 인식하여 화자 별 발화시간 분리
- 파인튜닝 지원



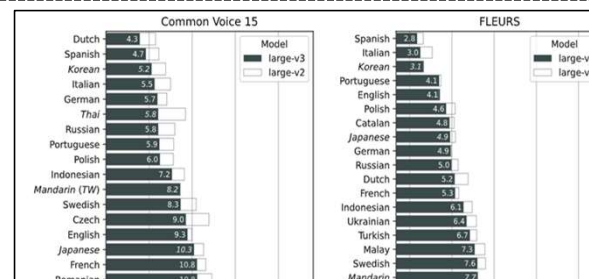
Whisper-large-v3 (Local)  
: STT(Speech-To-Text)

- Local 환경에서 운영 가능
- 낮은 CER\*
- 다국어(Multilingual) 지원
- 파인튜닝 지원

\* CER(Character Error Rate) → 한국어 STT 성능 평가 지표

$$CER = \frac{S + D + I}{N} \times 100(\%)$$

- 대치 (Substitution, S): 정답 문자가 다른 문자로 잘못 변환된 경우 (예: '사과' -> '사고')
- 삭제 (Deletion, D): 정답에 있는 문자가 누락된 경우 (예: '사과' -> '사')
- 삽입 (Insertion, I): 정답에 없는 불필요한 문자가 추가된 경우 (예: '사과' -> '사과과')



[오픈소스 whisper-large-v3의 언어 간 CER/WER 비교]

\*출처: <https://github.com/openai/whisper/discussions/1762>

# 비정형(음성) 데이터 가명처리

## 💡 STT (Speech To Text)



Pyannote-audio  
: 화자 분리(Speaker Diarization)

발화자 및 시간대



발화 내용

Data  
Cleansing



Whisper-large-v3 (Local)  
: STT(Speech-To-Text)

```
"file": "사례1",
"result": {
  "text": "안녕하세요. 저는 윤미숙이라고 하고요. 서울사이버대학교에서 학생들을 가르치고 있어요. 본인 소",
  "chunks": [
    {
      "timestamp": [
        0.0,
        6.82
      ],
      "text": "안녕하세요. 저는 윤미숙이라고 하고요. 서울사이버대학교에서 학생들을 가르치고 있어요.",
      "speaker": "SPEAKER_01"
    },
    {
      "timestamp": [
        7.5,
        9.0
      ],
      "text": "본인 소개 좀 해주시겠어요?",
      "speaker": "SPEAKER_01"
    },
    {
      "timestamp": [
        9.78,
        14.1
      ],
      "text": "저는 24살 이서연이라고 하고요. 동덕여대 다니고 있어요.",
      "speaker": "SPEAKER_00"
    },
    {
      "timestamp": [
        14.94,
        17.58
      ],
      "text": "혹시 전공이 뭔지 물어봐도 될까요?",
      "speaker": "SPEAKER_01"
    }
  ]
}
```

가명처리할 최종 json 파일 생성

## 비정형(음성) 데이터 가명처리

## 💡 가명처리 기법 1: 정규표현식을 활용한 규칙 기반 개인정보 삭제/대체/마스킹

## 사례 1) 이름/학교명/주소(지명)/나이

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
이름	윤미숙	대체	홍길동	0:02
이름	이서연(서연)	대체	김행복(행복)	0:11
학교명	서울사이버 대학교	대체	대한민국 대학교	0:04
학교명	동덕여대	대체	한국대	0:12
학교명	평택	삭제/대체	- / 경기도	1:39
나이	24살	대체	20살	0:10
나이	8살	대체	4살	1:32
전공	회화	대체	미술	0:18

```

surnames = {
    '김', '이', '박', '최', '정', '강', '조', '윤', '장', '임', '한', '오', '서', '신', '권', '황', '안',
    '송', '류', '전', '홍', '고', '문', '양', '손', '배', '백', '허', '유', '남', '심', '노', '하',
    '곽', '성', '차', '주', '우', '구', '라', '민', '진', '지', '엄', '채', '원', '천', '방', '공', '현',
    '함', '변', '염', '여', '주', '도', '석', '소', '선', '설', '마', '길', '연', '위', '표',
    '명', '기', '반', '왕', '금', '옥', '인', '맹', '제', '모', '남궁', '황보', '제갈', '사공',
    '선우', '서문', '독고', '동방'
}

# 개선된 이름 패턴: 성씨 + 이름 + 명확한 문맥
surnames_pattern = '|'.join(surnames)
self.name_pattern = re.compile(
    rf'({surnames_pattern})[가-힣]{{1,2}}(?:\s*(?:입니다|이고요|이라고|씨|님|선생님|입니다만|이에요|예요))?'
)

```

```

# 기존 패턴들은 그대로 유지
self.rn_pattern = re.compile(r'(\d{6}[-]\s?[1-4]\d{6})')
self.phone_pattern = re.compile(r'(01[016789])[-.\s]?(\d{3,4})[-.\s]?(\d{4})')
self.email_pattern = re.compile(r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}')

# 개선된 주소 패턴: 실제 주소 구조만 탐지
self.address_pattern = re.compile(
    r'('
    r'([가-힣]+(시|도|특별자치시|특별시|광역시))\s*'
    r'([가-힣]+(구|군|시))\s*'
    r'([가-힣]+(동|읍|면|리|가))\s*'
    r'(\d{1,5}번지)?'
    r'|'
    r'([가-힣]+(로|길|대로))\s*(\d{1,4})(-\d{1,4})?')?'
)

```

## 비정형(음성) 데이터 가명처리

## 💡 가명처리 기법 1: 정규표현식을 활용한 규칙 기반 개인정보 삭제/대체/마스킹

사례 2) 이름/학교명/주소(지명)/나이

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
이름	윤미숙	대체	홍길동	0:04
이름	안정태(정태)	대체	안행복(행복)	0:09
학교명	한양대	대체	한국대	2:09
학교명	서울대	대체	대한대	10:59
주소(지명)	울산	대체(마스킹)	땡땡	2:17
주소(지명)	서울에	대체(마스킹)	땡땡에	14:04
주소(지명)	왕십리역	대체(마스킹)	땡땡역	14:34

cf. 사례 2의 디자인 전공은 현재 전공과 관련 없는 정보로 가명처리 X

사례 3) 이름/주소(지명)

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
이름	윤미숙	삭제	-	0:05
이름	서민정	대체	김행복	0:28
주소(지명)	부산시	대체 (마스킹)	땡땡시	3:08

## 비정형(음성) 데이터 가명처리

## 💡 가명처리 기법 2: 추가 검수를 통한 필터링 되지 않은 개인정보 및 민감(특이)정보 처리

## 사례 1) 학년 정보 삭제 및 민감(특이) 정보 삭제

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
학년	지금 4학년 다니고 있어요.	삭제	-	0:19
민감(특이) 정보	열 달 전	대체	일곱 달 전	1:06
민감(특이) 정보	자취하고/자취하면서	삭제	-	1:42 /2:18

## 사례 2) 필터링 되지 않은 정보 대체 / 민감(특이) 정보 삭제

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
나이	스물 한두 살	대체(범주화)	20대	2:06
학년	3학년	대체 (마스킹)	뽕학년	2:09
민감(특이) 정보	울산시청의 좀 높 은 자리에 계셨거 든요. 아버지는 지 금 울산시 부시장 이세요.	삭제	-	9:43~ 9:38

## 사례 3) 민감(특이) 정보와 관련된 수치형 데이터 대체

카테고리	처리대상 정보 유형	처리 방법	처리 결과	시간
민감(특이) 정보	4년	대체	5년	0:31

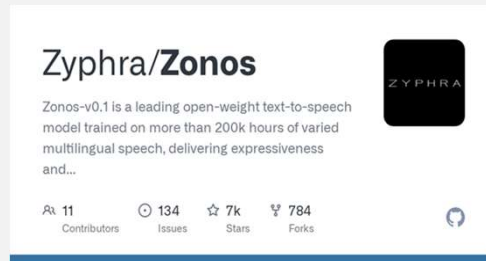
cf. 피부와 상담 코디네이터 퇴사 및 다른 곳에서의 1~2회 상담 시기  
⇒ 특정 가능



# 비정형(음성) 데이터 가명처리

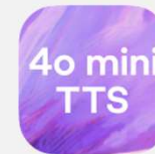
## 💡 TTS (Text to Speech)

### Zonos(오픈소스 한국어 지원 TTS)



- Local 환경에서 운영 가능
- zero-shot voice cloning

### GPT-4o mini TTS



- 빠르고 다양한 내장된 voice
- instruction으로 조건 설정 가능

사례 2  
(숫자 텍스트 음성 변환 X)



내담자: 하루 육시간씩 아침 시부터 사시까지 해요.



내담자: 하루 여섯 시간씩 아침 열시부터 네시까지 해요.

사례 3  
(음성 변환 X)



내담자: ???



내담자: 남들 앞에 서는 걸 진짜 못하거든요.

✓ 음성 합성의 퀄리티를 고려하여 GPT-4o mini TTS 적용



## 결론

### ✓ 가명처리 기법 및 기술

#### 정형데이터

- 시나리오 목적을 고려한 가명처리 기법 (e.g. 범위방법, 잡음 추가 등)

#### 비정형데이터

- 온프레미스 환경에서의 STT 고려
- 대화자 인식
- STT 및 TTS 모델 비교 선정
- 정규표현식 및 추가 검수를 통한 가명처리

### ✓ 향후 보완 가능성

#### 비정형데이터에서의 다양한 PII 개체명 인식

- NER(Named Entity Recognition) 적용 (e.g. KoELECTRA KLUE-NER)

#### 화자인식모델(Pyannote)와 STT모델(Whisper)

- 파인튜닝 적용을 통한 인식 성능 향상 및 CER 감소

#### 엄격한 가명처리 환경

- 온프레미스 환경에서도 가능한 TTS 모델 적용 (e.g. 파인튜닝한 Zonos 모델)

# 감사합니다.

총괄주최 |  개인정보보호위원회  과학기술정보통신부

공동주최 |  보건복지부  금융위원회

주관 |  한국인터넷진흥원

후원 |  건강보험심사평가원  국민건강보험  금융결제원  금융보안원  DOUZONE  한국신용정보원  Kdata 한국데이터산업진흥원  SSiS 한국사회보장정보원  NIA 한국지능정보사회진흥원  KLID 한국지역정보개발원