콜센터 서비스를 위한 한국어 음성인식 기술 version 2.0

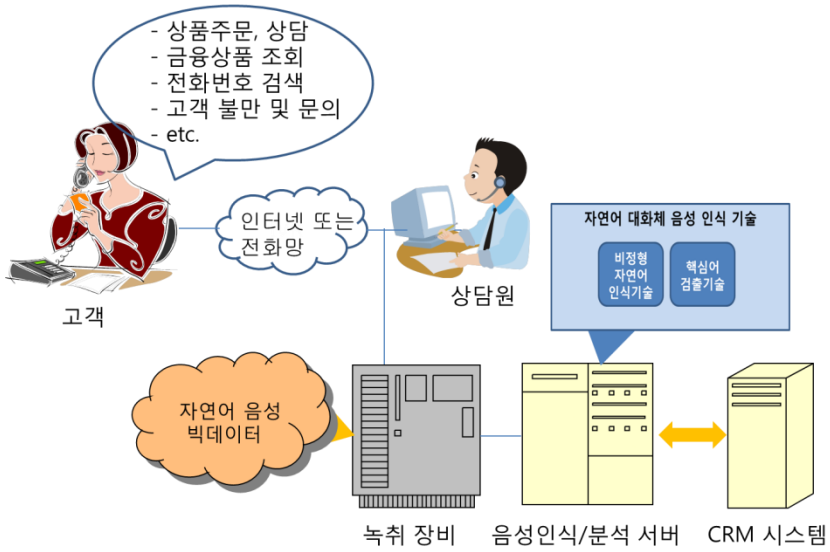
지능정보연구본부

제4차 산업혁명을 선도하는 ICT Innovator

ETRI

기술 개요

- **정의**
  - **"콜센터 서비스를 위한 한국어 음성인식 기술"은 콜센터의 상담원과 고객과의 통화 녹음 데이터를 자동으로 음성인식 하는 기술로써 자연어 대화체 음성 인식 기술이다.**
- **활용 방법**
  - **이 음성인식 기술은 종래 ARS 시스템을 일부 대체하거나 SA(Speech Analytics) 서비스에 적용 가능하다.**
  - **1차적으로 고객 센터로 걸려오는 전화에 대해 24시간 365일 무휴 고객응대서비스가 가능하여 고객 만족도를 제고할 수 있다.**
  - **2차적으로 콜센터 녹취 데이터를 음성 인식하여 문자정보로 자동 변환한 후, 이를 분석하여 상담원 대응의 타당성 여부, 고객 불만 사항 및 이슈 조기탐지 등을 수행하는데 이용한다.**
- **시스템 구성**



기술이전 내용 및 범위

- 기술명 : 콜센터 서비스를 위한 한국어 음성인식기술 version 2.0

- 기술이전의 범위
  - 음성인식 엔진
    - 음성인식 SDK (리눅스 라이브러리 및 그 API 제공) 및 사용자 지침서
    - 음성인식 SDK 적용 예제 코드
  - 한국어 음향 및 언어 모델, 이에 기반하는 적응 훈련 도구
    - 기본 음향 모델: 8kHz
    - 음향 모델 적응 훈련 도구 및 스크립트
    - 기본 언어 모델: 10만 단어급 지원
    - 언어 모델 적응 훈련 도구 및 스크립트

- 기술이전 특이 사항
  - 기존 “콜센터 녹취데이터 음성인식 기술” 및 “콜센터 서비스를 위한 한국어 음성인식기술” version 1.0의 업그레이드 기술임
  - DNN(Deep Neural Network)에 기반하는 딥러닝(Deep Learning)에 따른 고도화된 음향 모델, 그 훈련 도구 및 스크립트가 추가됨

- 상세한 기술 구성은 <첨부> 참조

기술료 제안

- 기술개발 기간 : 2013. 3. 1 ~ 2015. 7. 31.(29개월)
- 투입 연구비 : 24억원
- 경상 기술료 방식 적용 (공동연구 참여기업이 있으나 본 기술개발에 기여가 없음)

구분	중소기업	중견기업	대기업
착수기본료 (천원)	120,000	240,000	240,000
매출정률 사용료 조건 1	총 판매석수에 따라 변경적용 • 1만석이하 : 11,000원/석 • 1만석~3만석 : 10,000원/석 • 3만석이상 : 9,500원/석	총 판매석수에 따라 변경적용 • 1만석이하 : 33,000원/석 • 1만석~3만석 : 30,000원/석 • 3만석이상 : 28,500원/석	총 판매석수에 따라 변경적용 • 1만석이하 : 44,000원/석 • 1만석~3만석 : 40,000원/석 • 3만석이상 : 38,000원/석
매출정률 사용료 조건 2 (※ 석당 기준을 적용 불가할 경우에 한하여 적용함)	1.25%	3.75%	5%

※ 매출정률사용료 산정 기준 : 제품(Speech Analytics)의 예상 가격은 약 80만원 정도이며 중소기업기준 1.25%의 매출정률 사용료를 적용할 경우 1만원임

※ 착수기본료 차감 조건 1: 본 건의 평가용 기술을 전수받은 기관에 대해 착수기본료를 차감하고 계약을 추진함
- 기술명: "IVR서비스를 위한 연속어 음성인식 기술(평가용)" (계획번호: 7040-2012-0001)
- 기술료: 20,000천원/중소기업, 40,000천원/대기업 (매출정률사용료 없음)

※ 착수기본료 차감 조건 2: 본 건의 하위 버전을 전수받은 기관에 대해 착수기본료를 차감하고 계약을 추진함
- 기술명1: "콜센터 서비스를 위한 한국어 음성인식기술" (계획번호: 1122-2014-10226)
- 기술명2: "콜센터 녹취데이터 음성인식 기술" (계획번호: 1122-2014-0001)
- 기술료(공통): 100,000천원/중소기업, 200,000천원/대기업

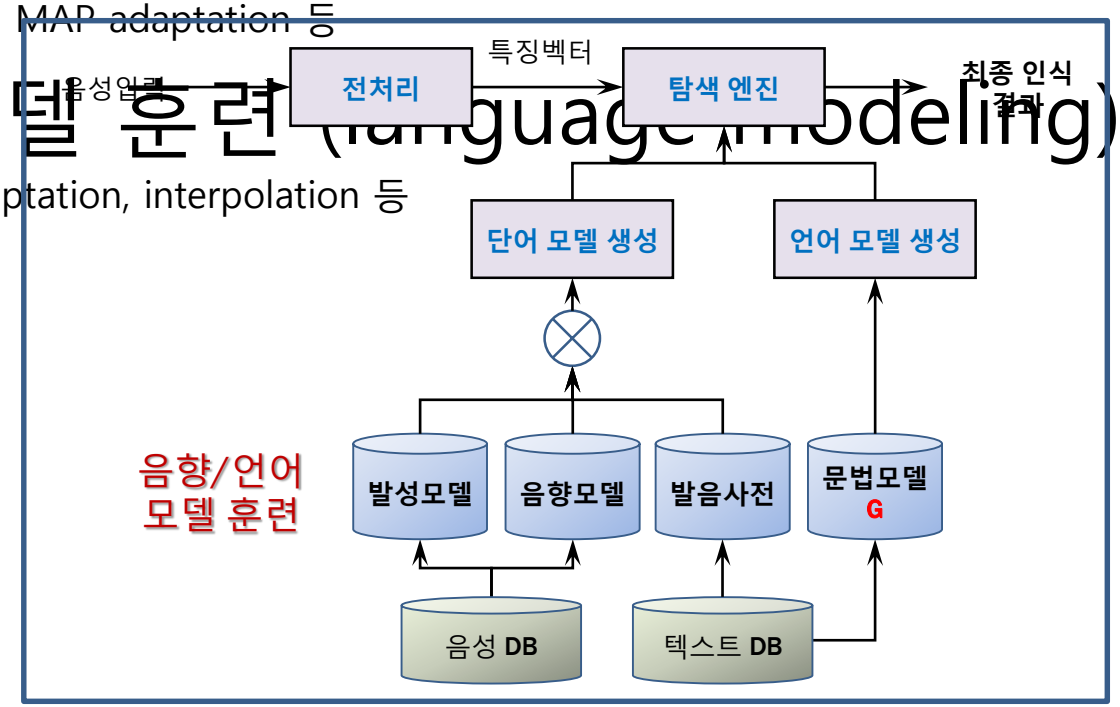
첨부

음성인식 엔진(ESTk-laser) 개요

- ESTk-laser: ETRI Speech Toolkit - Large Scale Speech Recognizer
- ESTk-laser is developed to recognize very large scale of recognition domain on both high-end servers and resource-limited embedded devices.
- Technical features
  - Language independency
  - Platform independency
  - Single channel speech enhancement
  - Noise-robust endpoint detection
  - Speaker and environment adaptation
  - Speaker and channel normalization
  - Deep Learning (deep neural network) support

음성인식 엔진의 구성

- 탐색 엔진 (search engine 또는 decoder)
  - 음향 및 언어 모델 등의 지식 베이스에 기반하여 고속/고성능 음성인식을 수행
- 음향 모델 훈련 (acoustic modeling)
  - full training, MAP adaptation 등
- 언어 모델 훈련 (language modeling)
  - domain adaptation, interpolation 등



LASER Specifications

Consideration			High-end device	Low-end embedded device
Language	Supporting languages		Korean, English	Korean, English
Platform	Supporting platforms		Linux, Windows	Windows, Android, iphone, nucleus
Recognition Mode	Continuous	Vocabulary size	>100K (140M trigrams)	> 10k
		RTF	1.0xRT	1.0xRT
	One-shot	Vocabulary size	-	> 450k VDE entries
		RTF	-	2.6
Minimum H/W requirements	CPU		2.6 GHz	620MHz
	Storage memory		30GB	50MB
	Running memory		40GB	14MB
etc	Grammar definition		ARPA, BNF, JSGF	ARPA, BNF, JSGF

LASER Architecture

- Base layer
  - Wrapper for platform independency
- Decoding layer
  - Acoustic search
  - Lexical search
  - Rescoring
  - Multi-stage search
- Interface layer
  - low-level APIs : DLLs
  - Script-level interfaces : python, java