방송 및 회의용 오디오 및 텍스트 정보의 시간 동기화를 위한 연속어 음성인식기술

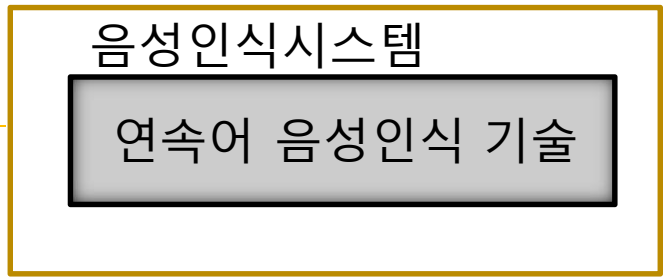
2016. 1.

한국전자통신연구원
Electronics and Telecommunications Research Institute

기술 개요

- 방송 및 회의용 오디오 및 텍스트 정보의 시간 동기화를 위한 연속어 음성인식 기술
  - 사용자는 텍스트 또는 음성 키워드를 사용하여 동영상 컨텐츠를 검색
  - 동영상 컨텐츠로부터 해당 자막에 대한 시간정보를 이용하여 원하는 위치로 이동

시간정보	자 막 스 크 립 트
00:00 - 00:01 00:02 - 00:50 00:51 - 00:51 01:00 - 01:30	안녕하세요? 오늘 배우는 문법은 제가 사실은 앞쪽에서 많이들 다뤄왔어요. 되고, (중간 생략) 쌤이 생각하기에 영어 문법중에 가장 중요한 문법은 무엇이라고 생각하나요? 그러면 저는 망설이지 않고, 그중에 하나를 오늘 배우게 될 전치사 라고 말할거에요. (이하 생략)



서비스 예시

- 방송 및 회의용 오디오 및 텍스트 정보의 시간 동기화를 위한 연속어 음성인식 기술

기술이전 내용 및 범위

- 기술명: 방송 및 회의용 오디오 및 텍스트 정보의 시간 동기화를 위한 연속어 음성인식 기술

- 기술이전의 범위
  - 리눅스 및 Windows 환경에서 실행 가능한 Library 형태의 오브젝트와 인식용 이미지파일 생성 도구
    - 한국어 음성인식 엔진 SDK
    - 런타임 이미지 생성 도구
    - 개발자용 지침서

※ 제약조건
- 본 기술은 불특정 음성에 대해 문자 정보를 생성하는 음성인식 기술이 아닌,
  오디오와 텍스트 정보 간을 시간적으로 동기화하여 정렬하는 기술임
- 따라서 적용 대상이 되는 방송 및 회의 콘텐츠에 대해서 개별 콘텐츠별로 미 리 속기사 등이 전사한 자막정보 또는 속기문서(텍스트)가 있어야 함
- 오디오 파일은 최소 16kHz 샘플링 주파수로 인코딩되어 있어야 함
- 방송 및 회의 콘텐츠에 한하여 적용함음성인식에 있어 유사어휘 및 어휘패
  턴을 사용하는 화자군에 기반하는 언어모델

기술료 제안 (예상기술료)

- 예상기술료 (공동연구 참여기업이나 본 기술개발에 기여가 없음)

구분		일반 기업		
		중소기업	중견기업	대기업
경상기술료	착수기본료(천원)	25,000	50,000	50,000
	매출정률사용료(%)	1.25%	3.75	5.0

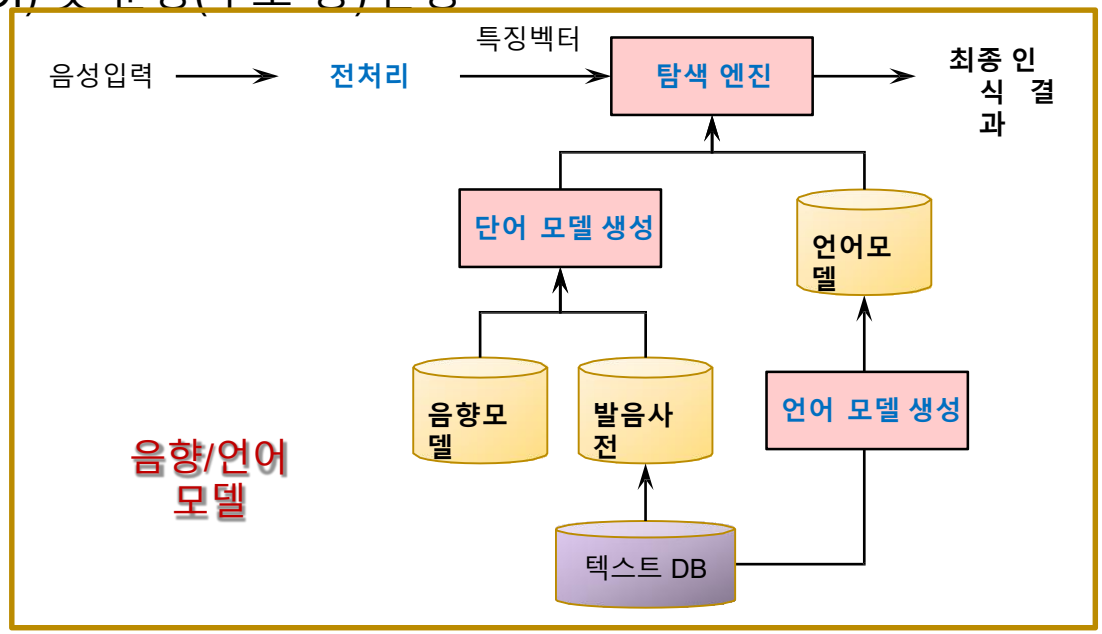
별첨

1. 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

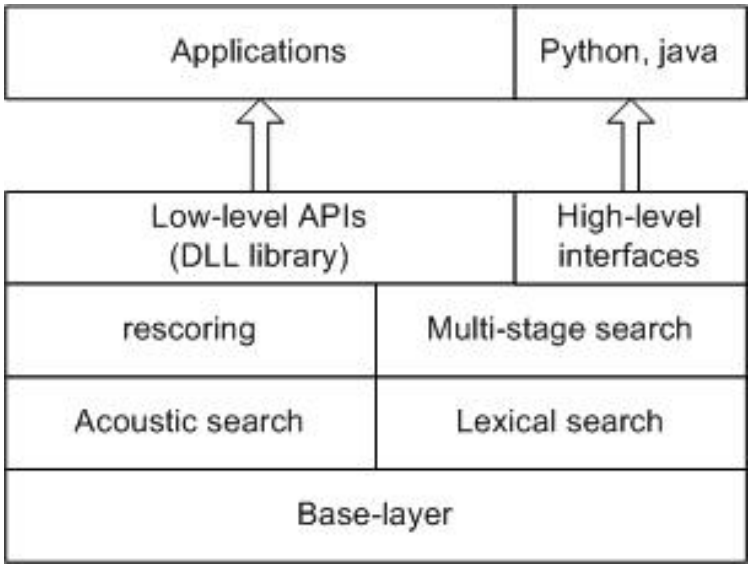
2. 음성인식 엔진의 구성

- 탐색 엔진 (search engine 또는 decoder)
  - 음향 및 언어 모델 등의 지식 베이스에 기반하여 고속/고성능 음성인식을 수행
- 음향 모델
  - 차량 환경의 잡음 환경 반영
- 언어 모델
  - 단어(POI) 및 문장(주소 등) 반영



3. LASER 구조

- 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



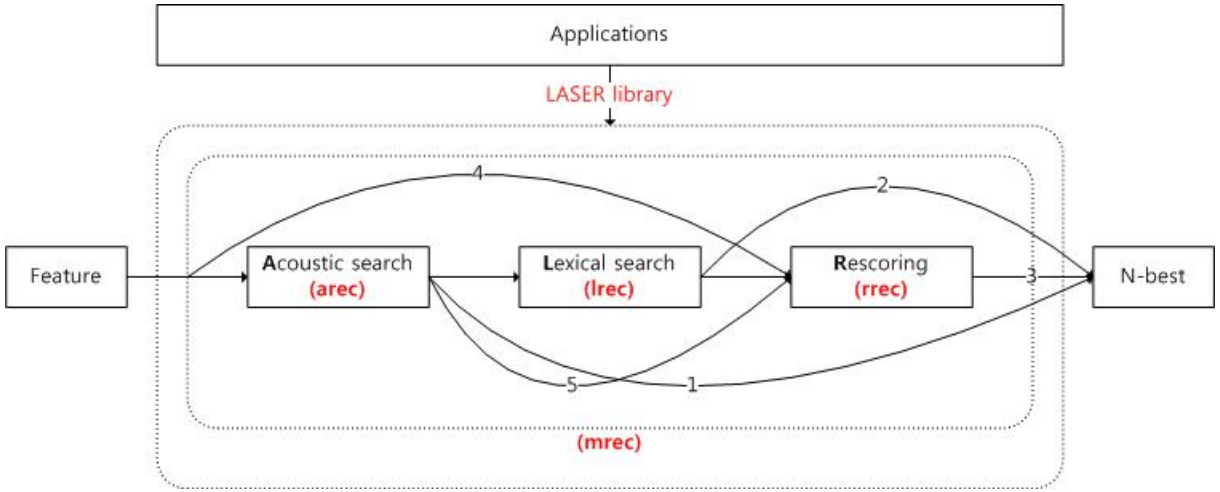
3. LASER 구조

Core 2 search components

- Acoustic search
  - Weighted finite state transducer-based speech recognizer
- Lexical search :
  - discrete HMM-based lexical level noisy channel decoder

Recognition modes

- Various recognition modes for different domain and system configurations



4. LASER 사양

Consideration			High-end device	Low-end embedded device
Language	Supporting languages		Korean, English	Korean, English
Platform	Supporting platforms		Linux, Windows	Windows, Android, iOS, Nucleus, etc.
Recognition Mode	Continuous	Vocabulary size	>100K (140M trigrams)	> 10k
		RTF	1.0xRT	1.0xRT
	One-shot	Vocabulary size	-	> 300k VDE entries
		RTF	-	1.0xRT
Minimum H/W requirements	CPU		2.6 GHz	1GHz
	Storage memory		30GB	50MB
	Running memory		40GB	100MB
etc	Grammar definition		ARPA, BNF, JSGF	ARPA, BNF, JSGF