

해외 음성 DB 구축 동향

○
*이용주,김봉완,김종진,양옥렬,**오원록,김종교
* 원광대학교 컴퓨터공학과, **전북대학교 전자공학과

Activities of Speech DB construction out of Countries

*Yong-Ju Lee, Bong-Wan Kim, Jong-Jin KIM, Ok-Yul Yang
**Won-Rok Oh, Chong-Kyo Kim
*Dept. of Computer Eng., WonKwang Univ.
**Dept. of Electronics, Chonbuk Univ.

요 약

음성정보처리 연구에 공통으로 이용 가능한 대량의 각종 음성 데이터를 수집, 편집, 배포하는 일은 연구 개발자의 입장에서는 분석, 합성, 인식등의 알고리즘 개발 평가에 이용 가능하며, 음성인식, 합성 시스템의 사용자 입장에서는 각종 시스템의 성능을 객관적으로 평가할 수 있다는 면에서 매우 중요하다. 본 논문에서는 국내 음성 DB의 효율적인 구축을 위한 방안 도출에 참고하기 위하여 해외 각국의 구축 동향을 기관별, 형태별, 분야별로 구체적으로 정리하여 소개한다.

1. 서론

한국어의 공학적 응용을 위해서는 그 기반이 되는 요소기술로서 음성인식 및 합성으로 대표되는 음성처리기술과 언어이해 및 기계번역으로 대표되는 언어처리기술의 연구가 필요하다. 이러한 음성 및 언어처리기술의 연구를 위해 가장 먼저 확보되어야 할 것이 음성, 언어 및 각종 사전 DB이다. 특히 한국어 음성을 대상으로한 음성 DB는 음성언어 연구의 기본으로서 개발초기 부터 확보되어야 할 연구자원이다.

선진각국에서는 자국어 음성DB에 대한 체계적인 구축이 음성 및 언어처리기술 확보를 위한 가장 기본적인 연구환경임을 깊이 인식하고 이에 대한 체계적인 확보가 공공연구기관을 중심으로 활발히 추진되고 있으며 이미 여러기관에서 이를 제작하여 CDROM 등에 수록하여 공동으로 이용할 수 있도록 하고 있다.

본 고에서는 이러한 외국의 구축 현황을 고찰함으로써 국내에서의 효율적인 음성 DB 구축에 참고하고자 한다.

2. 각국의 구축동향

2-1. 미국

미국은 국방부의 DARPA 프로젝트의 연구에 의해 구

축된 음성DB를 상무부 산하의 국립 연구소인 NIST(National Institute of Standards and Technologies)가 중심이 되어 종합적으로 보급하고 있고 그밖에도 각 연구소 및 대학, 기업등에서의 결과도 포함하여 LDC(Linguistic Data Consortium)을 통해 공동제작 및 보급이 이루어지고 있다.

1) 전화음성DB

가) MACROPHONE

- 1992년 COCOSDA(the Coordinating Committee for Speech Databases and Assessment)에서 시작한 합동 프로젝트인 POLYPHONE의 미국 영어판
- LDC(Linguistic Data Consortium)가 스폰서
- 디지털전화선(T1 혹은 ISDN)을 통해 직접수록

나) NYNEX NTIMIT 코피스

- TIMIT 코피스의 전화 회선을 통한 전화 음성 데이터

2) 단어 및 연속음성DB

가) TIMIT (DARPA Acoustic-Phonetic Continuous Speech Corpus)

- 음성인식시스템의 개발 및 평가와 acoustic-phonetic 연구용
- 대표적인 미국 방언 8개를 고려, 630명이 발성한 문장 DB
- CD-ROM 화 되었음

나) DARPA Resource Management Continuous Speech Corpora

- 연속 음성 시스템 설계 및 평가에 사용
- 미 해군 자원관리를 데스크로 한 100단어로 구성된 낭독 문장 음성 DB
- RM1
 - 화자독립적 자원관리 훈련 데이터
 - 160명의 화자, 25,000 문장
- RM2

백의 음성 DB 구축 동향

- 10 명이 참여한 3171 발성
- 화자중속적 인식기 훈련용
- ATIS2
 - 10대 도시에 대한 운항정보 DB
 - 450개의 주제에 대한 15,000 발성
- ATIS3
 - 46대 도시, 52 공항에 대한 운항정보 DB
 - 137개의 주제에 대한 774개의 시나리오에 대한 7,300 발성

라) Texas Instruments

- 고립단어 데이터베이스
 - 화자중속 인식알고리즘과 장치 테스트 목적
 - 20개의 단어 데이터베이스(0-9의 숫자와 10개의 제어된 단어들)
 - 16명의 성인 화자(남여 각 8명)
- Connected Digits 데이터베이스
 - 연결숫자의 화자독립인식을 위한 알고리즘 설계 평가 목적
 - 77개의 숫자열
 - 남자 111명, 여자 114명
 - 50명의 소년, 51명의 소녀
 - 20KHz sampling, 16 bits 양자화.
 - 방음실 환경

마) DARPA ROBUST 음성 데이터베이스

- Simulated Stress 데이터베이스
 - 내용(contents): 105개의 고립단어
 - 화자: 5명의 남성, 3명의 여성 화자
 - 환경(environment): fast, loud, shout, soft, 105dB noise 등 여러조건인 음성을 유도
- Advanced Fighter Technology Integrated connected word DB
 - 내용: 87개의 고립단어들과 153개의 구
 - 화자: 5명의 남성, 3명의 여성 화자
 - 녹음: 마이크로폰을 끼운 F-16 헬멧과 산소마스크
 - 환경(environment): 스트레스, 노이즈를 유도하는 조건
- Simulated Stress 데이터베이스
 - 내용: 50단어로 된 두개의 집합
 - 환경: LHX 헬리콥터들 시뮬레이션하는 진동조건
- Robust Recognition 데이터베이스
 - 내용(contents): 전투기에 적절한 207 단어
 - 539 training phrases와 219 test phrases
 - 화자: 10명 이상의 남자 공군직원

바) DARPA

- Acoustic-phonetic speech 데이터베이스
 - 미국의 모든 중요한 지역언어 대해서 연속음성의 음향적-음소적 특성을 나타내기 위한
 - 630명이 다양한 문장 발성
 - 20KHz sampling, 16 bits 양자화.
 - 방음실 환경
- Acoustic-phonetic speech 데이터베이스
 - 1000개의 단어에 의한 2000개 이상의 문장
 - 화자: 160명이 발성
 - 방음실 환경

2-2. 유럽

각국이 단독 또는 EC여러나라간의 음성 및 언어연구 공동프로젝트하에서 구축되고 있고 지역특성상 다언어(multilingual) 음성DB가 많다.

1) 전화음성DB

가) SIRVA (이탈리아)

- 음성 인식과 훈련용 DB
- 68개의 고립단어와 4개의 연속된 디지털로 구성
- PCM 16 Bit, 8kHz 샘플링 주파수들 이용
- 전국적 26개의 방언으로 부터 음성 데이터 수집
- 68개의 고립단어와 6개의 리스트로 구성되고, 4개의 연속된 디지털로 구성

나) CSELT (Telephone Application)

- 이탈리아어에 대한 전화망에 인식응용을 테스트하고 개발하기 위한
- 숫자와 5개의 명령단어로 구성
- 1100명의 화자 (남 550, 여 550)
- Sampling 주파수: 8kHz
- Speaker interface: 목소리로 알림

다) CSELT(DIGITS)

- 언어: 이탈리아어
- 목적: 화자 독립 인식응용의 훈련과 테스트용
- 13개의 단어
- 80명의 화자 (남 40, 여 40)
- Sampling 주파수: 12.8 kHz
- 환경: 조용한 방

2) 단어 및 연속음성 DB

가) "SPICOS"를 위한 연속음성 DB

- 독일어에 대한 훈련용 100문장, 평가용 200문장, 평가용 914단어를 대상
- 화자중속적 자원관리 연속음성 훈련 데이터
- 남녀 각 2명, 각각 2,652 문장
- 모두 CDROM화 되어 있음

다) ATIS (Air Travel Information Service)

- 연속음성인식과 자연언어처리를 위한 연속음성 공동 코퍼스
- WOZ 방법으로 수집
- 비행기에 의한 여행서비스를 데스크로 한 대화형식의 자유발화 음성 DB
- ATIS0 조종사 연속음성 코퍼스
 - WOZ 방법으로 수집
 - 30 명이 참여한 912 발성
 - 16 KHz 16 Bits 샘플링
- ATIS0 연속음성 데이터의 낭독 버전
 - 20 명이 참여한 478 낭독버전
 - 16 KHz 16 Bits 샘플링
- ATIS0 화자중속 훈련 데이터
 - 남성 3명, 여성 2명이 발성

나) BREF

- 프랑스어에 대한 화자변량 연구, 화자 독립

- 인식 시스템 개발을 위한 데이터
- 발생 : 개인당 650 문장 (9500 단어)
- 전체 167,539문장 (4,244,810단어) 중에서 11,002 문장 (115,746단어)를 선택
- 목표 : 개인당 10,000단어, 100명분 계획

다) GRECO(BDSONS)

- 프랑스어에 대한 인식알고리즘 평가,인식,합성에 새로운 모델 연구
- 단독숫자,3연및 4연숫자,단어및 음소열
- 화자
- 6명의 언어학자가 표준 프랑스어 발음을 대표하는 12명의 화자를 선택.

라) NATO(RSG10)

- 다국어(영국 영어, 프랑스어, 미국 영어, 네덜란드어, 독일어)에 대한 인식성능에 영향을 미치는 non-native 효과및 노이즈 보상(compensation)알고리즘을 연구
- 내용: 고표준자및 연결숫자
- 화자
- 자국어언어로 9회 발생
- 자국어언어와 비자국어언어로 10회 발생
- 다섯 나라(영,미,프,독,네)에서 19명의 화자를 선정 (각 나라의 여성화자 1명씩이 포함된다.)

마) COST 209

- 다국어(multilingual)에 대한 음성인식형 화자 중속 전환기를 등급화하고 비교 연구하기 위한
- 내용(contents) : 10개 숫자(0-9),30개의 문
- 각 5명의 남,여 화자

바) RSRE

- 영국 영어에대한 음성 분석 방법을 평가하고 인식알고리즘을 기술하고 테스트하는데 사용
- 상업적인 인식기를 평가
- 화자 변동과 근원을 관찰하기 위한,
- 내용(contents)
- DRT(Diagnostic Rhyme Test)의 영국
- 영어식 버전 - 192 단어
- 숫자 - 10개 단어, 알파벳 - 26 단어
- 전화 회화에 보통 사용하는 단어 60개
- 화자 : 각 10명의 남,여 화자.

사) Edinborough Corpus of Dictated Speech

- 영국 영어에 대한 자연스러운 환경에서 취한 task-specific 음성시료를 제공 목적
- 내용(contents)
- 4 spontaneous dictations
- 2 spontaneous dictations
- 화자
- 25명의 전문 화자(14명의 여성,11명의 남성)
- 35명의 비전문 화자

아) Edinborough 연속음성 데이터베이스

- 영국 영어에 대한 연속음성의 분석 데이터용
- 98개의 문장을 4번 반복
- 화자 : 10명의 화자(각 5명의 남,여 화자)

자) CSELT

- Training Lexicon DB
- 이탈리아어에 대한 subword unit 마르코프 모델에 기초한 인식 프로그램의 training을 위한 사전 작성 이 목적
- 음소적으로 균형된 1105 고립단어
- 9명의 화자(4명은 여자, 5명은 남자) 참여
- Test Lexicon DB
- 이탈리아어에 대한 subword unit 마르코프 모델에 기초한 인식 프로그램의 training을 위한 사전 작성이 목적
- 1021 고립단어 대상
- 5명의 화자(2명은 여자, 3명은 남자)

차) EUROM.0 (ESPRIT Project I54L,SAM)

- 덴마크어(DK), 네덜란드어(NL), 영어(UK), 프랑스어(F), 이탈리아어(I)에 대한 음성의 입력평가에 사용하고, 음성원에 cross-language 분석 데이터와 음향-음소 특징을 제공하기 위한
- 내용(contents)
- 단독숫자,digit triples,많은 수를 포함한 연속음성
- 화자 : 화자 20명 : 언어당 4명(각 2명의 남,여)

2-3. 일본

전자공업진흥회, 일본음향학회 등이 중심이 되어 각 대학, 연구소 및 기업이 공동으로 구축한 음성DB를 학회들 중심으로 보급 활용하고 있다.

1) 전화음성DB

가) VAJ (Voice Across Japan)

- TI 의 Tsukuba Research and Development laboratory에서 수행
- 목표 : 10,000화자 전화음성 DB 구축
- 주제당 4개의 PBS수집
- 503 ATR PBS 세트

2) 단어및 연속음성DB

가) ATR음성 DB

- 단어음성DB
- 국어사전의 중요어 5229단어, 음소 벨런스 216단어, 숫자, 알파벳, 110음절, 국제회의 예약 텍스트를 대상으로 한 모의회화 115문장,
- 남녀 각 10명씩 20명의 어나운서 및 나레이터
- 모의회화는 어절단위 끊어읽기 방법에 따라 3종류로 발생
- 문장음성DB
- 신문 잡지 소설 편지 교과서 등의 문헌으로부터 무작위 추출한 약 10000개의 문장중 음소환경의 벨런스를 고려하여 작성한 503 문장을 대상.
- 남녀 각5명씩 10명의 나레이터가 발생.
- 음성세그먼트 레이블 및 음운정보,언어정보 부여
- 다수화자 음성 DB
- 불특정화자 음성인식용
- 중요어 중에서 750단어, 음소벨런스문장 150 문장

백의 음성 DB 구축 동향

- 대부분 음성레이블링 되어 있음
- 250명이 발성
- 텍스트음성DB
 - 음성합성연구용 DB
 - 중학교 교과서 및 NHK TV 교재에서 뽑은 11개 글중에서 300문장을 대상 텍스트 낭독문장
 - 남녀 각1명의 나레이터가 발성한 음성
 - 레이블링,운용정보(피치, 액센트구 등) 및 언어정보(품사, 활용, 의존관계 등) 부여
- 영미어 음성DB
 - 영국영어의 경우
 - 영국 Edinborough 대학의 CSTR 작성
 - 일본어 단어 및 문장DB와 대응
 - 5000고빈도 단어, 음소벨런스 200 단문의 연속발성 및 단어구절 발성, 연숫자 및 알파벳
 - 남녀 각2명씩 4명
 - 레이블링 되어 있음
 - 미국영어의 경우
 - CMU가 작성한 12세선(235문장)으로 된 국제회의 예약테스크의 모의회화 음성.
 - 50명 발성
 - 연속발성 및 단어구절발성
 - 일부메타타 레이블링

- 323개의 발성, 각 4번 발성
- 단음절, 이산(discrete)단어, 연결 숫자들
- 화자
 - 각 75명의 남,여 화자.

마) 전자기술총합연구소(ETL)

- 대상
 - 100개의 단음절, 부의미단어(VCV,CVCVCV 등)
 - 54도시명, 100도시명(JEIDA와 동일)
- 문장(Short story, 뉴스 등)
- 음소벨런스 단어세트(WD-I 492단어, WD-II 1542 단어)
- 문장세트
- 대화문 등
- 20KHz, 12Bit
- 10명에서 수10명정도의 화자

바) 동북대 음성 DB

- 음성 데이터와 함께 12GB 음성 파형 데이터
- 12 Bit, 20 kHz, CDROM에 수록
- 내용
 - 212단어 세트에서 선택한 16,800단어
 - 발성자 : 남 32, 여 40
 - 300단어 세트에서 선택한 113,480단어
 - 발성자: 남 15, 여 20명

나)일본음향학회 연구용 연속음성DB

- ATR음소벨런스문장에 의한 연속음성코퍼스
 - ATR음소벨런스 문장에 대해 15기관이 협력
 - 64명의 화자(남30명, 여 34명) 총 9600문장 수집 (화자별 약 150문장)
 - 16khz샘플링
- 안내테스크 연속음성코퍼스
 - 목적지향형 대화문장의 낭독음성을 대상.
 - 16세대의 대화문을 1대화당 10명,16명분 수록
 - 남자 18명, 여자 18명 계 36명이 발성
 - 수집된 전체문장은 11500문장.

다) 일본 문부성 중점영역연구에 의한 음성 DB

- 연속음성DB
 - 남녀각 10명, 4회발성
 - 20~60세대(각 년대별 2명)
 - 각 기관 분담 수록
 - PCM processor에 의한 VTR에 녹음, 나중에 DAT판을 만들. 일부를 CDROM화 하였음.
 - 16KHz, 16bit
 - 대상음성목록: ATR단어, 단음절, 외래음절, 일본어교육용 문장, 외문문, 음성품질평가용 문장 (일상회화중에서 출현빈도가 가장 높은 음절 30개를 이용하여 만든 문장), 이습우화 (바람과 햇님), 일기에보문, 뉴스 및 나레이션.
- 방언음성DB
 - 일본 각지의 방언음성 수록
 - DAT판, CD판, CDROM판이 있음.

라) JEIDA

- 언어 : 일본어
- 목적 : 인성기의 성능 평가용

3. 결론

이상과 같은 기관별 구축현황을 주요항목별로 표-1 과 같이 정리하였다.

용도별로 분류하면 인식 시스템의 훈련및 평가용이 주종을 이루고 음향음성학적인 분석용 DB 가 언어별로 공통적으로 구축되었다.

제작기관은 DARPA,ESPRIT 등 대형의 음성관련 프로젝트에 의해 구축된 DB 는 기관별로 분업화 되어 있고, 예를들어 수집,레이블링,분석,배포 등이 대량의 데이터 수집을지역적으로 분할 수집이 눈에 띈다.

대상목록은 고립숫자,연결숫자,음소열,단어(PBW 포함),문장(PBS 문장) 및 대화문이 언어별로 다양하게 구축되어 있다. 발성량은 전화음성의 경우 대상어휘는 적으나 발성화자는 많은것이 특징적이다. 천여명이상이 보통이며 만명이상을 목표로 하는 추세이다.

녹음환경에서는 전화음성과 사무실음성, 사무실음성도 방음실과 일반사무실 음성으로 대별된다. 특히 군사적 목적으로 비행기등 혹독한 조건의 복수환경에서의 음성도 모아지고 있다.

발성목록의 제시방법은 낭독음성의 경우 CRT 터미널에 발성목록을 자동으로 제시하거나, 리스트 시트를 이용하는 경우로 대별되며, 대화음성의 경우 자동응답 시스템을 이용하거나 Wizard of OZ 방법을 이용하여 인간과의 대화를 모의하는 형태로 발성을 유도하고 있다.

레이블링의 경우, 음향음성학적인 분석을 목적으로 하는 경우는 대부분 음소단위로 레이블링 되어 있으며, 인식용의 경우 대량의 음성데이터중 일부만 레이블링 되어 있는것이 대부분이다.

다지탈화는 전화음성 DB 의 경우 주로 8Bits 에 8 KHz 로 샘플링하고 일반음성 DB 는 초기에는 12 Bits,12KHz 가 주를 이루었으나 최근에는 20KHz 또는 16 KHz 샘플링에 16 Bits 양자화로 통일되어가는 경향이다.

저장매체 또한 이남로그 테이프, 마그네틱 테이프, 또는 PCM 프로세서를 통한 비디오 테이프에서 DAT 테이프, 이제는 경제적인 보관, 배포, 이용을 위해 CD-ROM 등으로 통일되어 가는 것으로 보인다.

구축방법은 기관별로, 또는 LDC(Linguistic Data Consosium)형태로 대량의 음성 DB 를 구축하여 공동으로 이용하고 있다. 대상은 단어뿐만 아니라 남독음성, 그리고 최근에는 자동통역이나 음성대화시스템 개발을 위한 자유발화음성(Spontaneous Speech)의 확보에 주력하고 있다.

국내에서도 최근 통신망에서의 지능형 서비스를 위한 음성연구의 필요성 증대, 지능형 컴퓨터의 휴먼 인터페이스의 수단 제공 등과 관련하여 음성 번역 연구등 대규모의 음성공학 연구가 활발해지고 있으므로 이러한 연구의 기초기반기술 및 자료로서의 음성 DB 구축에 대한 노력이 절실히 요구되고 있다.

참 고 문 헌

[1] Shuichi Itahashi, "Recent Speech Database Project in Japan", Proc., 24.1.1, ICSLP 90.
 [2] NIST: Speech Copora Produced on CD-ROM Media by The National Institute of Standards and Technology(NIST), April, 1991
 [3] Seiichi Nakagawa, "Assessment and Database of Speech Recognition/Understanding Systems., (in Japanese), 일본전자정보통신학회, 12/'90
 [4] T. Ebara, et al, "ATR Dialogue Database", Proc., 24.5.1, [CSLP 9012] S. Itahashi, et al, "A Collection and Editing System of Speech Data for Research", 일본음향학회강연논문집, '82. 10
 [5] A. Kurematsu, "ATR Japanese speech database as a tool of speech recognition and synthesis", Speech Communication, 9, 1990
 [6] J. Gauvain, et al, "Design Considerations and Text Selection for BREF, a large French read-speech corpus", Proc., 24.6.1, ICSLP 90
 [7] K. Tanaka, "The Speech Database for Speech Analysis and Recognition Research", Proc., 24.7.1, ICSLP 90
 [8] S. Makino, et al, "A Distributed Speech Database with an Automatic Acquisition System of Speech Information", Proc. 24.91, ICSLP 90
 [9] 이용주, 이정철, 김경태, "음성 데이터베이스 구축에 관하여", 음향학회지, 제7권 제5호, 1988
 [10] 김경태, 최준혁, 이용주, "음성 데이터베이스 관리 시스템의 구축", Korea-Japan JointSymposium on Acoustics, 1991. 7
 [11] K. Shirai, et al, "Speech database projects in Japan - present and future", Proc. of ESCA Workshop, Speech I/O Assessment and Speech Database, Sep. 1989

[12] F. Guyote, et al. "A Speech database at the United States Air Force Academy", Proc., 7.2.1, ICASSP 86
 [13] K. Shikano, "Phonetically balanced word list based on information entropy", Preprints Autumn Meeting Acous. Soc. Japan, Paper 3-3-10, Mar. 1984
 [14] S. Hayamizu, et al. "Generation of VCV/CVC Balanced Word Sets for Speech Database", 일본 전자기술총합연구소, 제49권 제10호, 1985
 [15] G. Doddington, "The next generation DARPA speech recognition/natural language database", Proc. of the ESCA Workshop, Speech I/O Assessment and Speech Database, Sep. 1989
 [16] J.m. Baker, et al, "Speech Recognition Performance Assessments and Available databases", Proc., ICASSP 83
 [17] 정유현, 최준혁 외, "공통 음성 데이터베이스 구축을 위한 사전 조사 연구", 전자공학회 하계 학술 대회, 1992.6.
 [18] Michel Weintraub, et al, "Constructing Telephone Acoustic Models from a High-Quality Speech Corpus", ICASSP94, Vol, pp 185-188, 1994
 [19] Thomas Staples, et al, "The Voice Across Japan Database-The Japanese Language Contribution to POLYPHONE", ICASSP94, Vol, pp 189-192, 1994
 [20] Ronald Cole, et al, "Towards Automatic Collection of the U.S. Census", ICASSP94, Vol, pp 193-196, 1994
 [21] Bruce Milliar, et al, "The Australian National Database of Spoken Language", ICASSP94, Vol, pp 197-1100, 1994
 [22] John garofolo, et al, "The Development of File Formants for Very Large Speech Corpora: Speech Corpora: SPHERE and SHORTEN", ICASSP94, Vol, pp 1113-1116, 1994
 [23] Carlos Ribeiro, et al, "A Software Tool for Speech Collection, Recognition and Reproduction", EUROSPEECH-93, Vol, pp 179-182, Berlin 1993.
 [24] G.Castagneri, G.Di Fabbri, A.Massone, M.oreglia, "SIRVA-A Large Speech Database Collected on the Italian Telephone Network", EUROSPEECH-93, Vol, pp 199-201, Berlin 1993.
 [25] 이용주, "한국어 음성언어정보처리와 음성 데이터베이스", 한국어정보처리 소식 제2권 제 1,2호(94.10)
 [26] 조철우, "음성 DB 구축을 위한 국제간 활동 현황", 본 워크샵 발표 논문집(95.6)

[표-1] 주요 음성 DB 구축 현황

코퍼스명	종도	제작기관	대상목표	발성방법	녹음환경	화자인터페이스	발성 형태	메이킹 방법	샘플링	양자화비	파일 형식
MACROPHONE	인공기 울림 및 평가	COCOSDA	단어 숫자 문자 구	5000명의 200,000 발성 진행 음성	발성목표		남성	메이킹 방법	8 KHz	8 Bits	CD-ROM
NYNEX NTIMIT	음성인식시스템 개발용 평가	NYNE	TIMIT 와 동일	6300 TIMIT 데이터	진화음성		남성	메이킹 방법	8 KHz	8 Bits	CD-ROM
OCI Spelled & Spoken	OCI 인공음성	OCI	화자 이름, 지역, 성장장소	38650 편의 진화	진화음성	자동음성시스템	남성	메이킹 방법			
OCI 인공음성 코퍼스	OCI 인공음성	OCI	화자 ID, 연구 내용, 화자, 문장, 화자, 문장, 화자, 문장	1927편의 진화	진화음성	자동음성시스템	남성	메이킹 방법			
SPIDRE 평가 지점 코퍼스	OCI 인공음성	OCI	화자 ID, 연구 내용, 화자, 문장, 화자, 문장, 화자, 문장	3700 명, 5000 문	진화음성	자동음성시스템	남성	메이킹 방법	8 KHz	16 Bits	2-CD
VAA	OCI 인공음성	OCI	화자 ID, 연구 내용, 화자, 문장, 화자, 문장, 화자, 문장	3700 명, 5000 문	진화음성	자동음성시스템	남성	메이킹 방법	8 KHz	16 Bits	2-CD
TIMIT	음성인식시스템개발용	DARPA	미국인사이드	630명 남성	noise-isolated	타이핑 프로그램	남성	메이킹 방법	20 KHz	16 Bits	CD-ROM
DARPA 자음 판리	연속음성시스템	DARPA	미국인사이드	168명의 27,685 문장	noise-isolated		남성	메이킹 방법	20 KHz	16 Bits	CD-ROM
연속음성 코퍼스	성체인 평가	DARPA	미국인사이드	168명의 27,685 문장	noise-isolated		남성	메이킹 방법	16 KHz	16 Bits	CD-ROM
VOYAGER	대화 시스템 개발	DARPA	미국인사이드	남녀 50 명의 4646 문장	noise-isolated		남성	메이킹 방법	16 KHz	16 Bits	CD-ROM
ATIS	기류발음평가시스템 개발용	DARPA	여러 서비스 발음	ATISGAITSI 에는 총 Woz 방법의 451 발음, PNAMBIIC 방법, ATIS3 에는 22300 이용 발음	noise-isolated	WOZ 자유발화	자유발화	메이킹 방법	16 KHz	16 Bits	CD-ROM 6장
VERBEX	단어 인식기 개발	VERBEX	11 개의 숫자	남녀 각 5 명의 사무실 10210의 발음	noise-isolated	발성 단어 리스트	남성	메이킹 방법			
US Air Force Academy	음성 분석		숫자 문자 음장	1000명 남성			남성	메이킹 방법			
T1 코퍼스	화자 음성인식평가시스템과 강자 테스트	TI	숫자 문자 단어	남녀 각 6 명의			남성	메이킹 방법	12.5KHz	12 Bits	
TIDIGITS	연결숫자열의 화자 특성 인식을 위한 알고리즘 설계 및 평가	TI	연결 숫자	남 161명, 여 165명 발음			남성	메이킹 방법	0	16 Bits	M-Tape
DARPA Simulated Stress DB	목적한 조건에서 인식기 개발	DARPA	고밀도 단어	남 5명, 여 3명의 발성	목적한 환경 조건	발성 단어 리스트	남성	메이킹 방법			
DARPA Fighter Technology Integrated connected word DB	목적한 조건에서 인식기 개발	DARPA	고밀도 단어	남 5명, 여 3명의 발성	F-16 비행과 산소 마스크		남성	메이킹 방법			
DARPA Simulated Stress DB	목적한 조건에서 인식기 개발	DARPA	단어	10명의 남자가 발음한 Wright-Patterson Air Force Base			남성	메이킹 방법			
DARPA Robust 인공 DB	목적한 조건에서 인식기 개발	DARPA	단어	868 단어			남성	메이킹 방법			
DARPA Acoustic-phonetic 음성 DB	지연된 연속음성의 음향과 음소의 특징 분석	TIMIT, SRL, NBS	지연된 문장	6300명의 6300 발성			남성	메이킹 방법	20KHz	16 Bits	
DARPA Acoustic-Boneclic 음성 DB	인식기의 개발 및 평가	DARPA	단어 음장	1600명의 760 문장			남성	메이킹 방법			
Boneclic 음성 DB	Word-Spotting 시스템의 개발 및 테스트		문장	남 80명, 여 86 명의 발음한 미국인 사무 지도			남성	메이킹 방법			CD-ROM

ATCO (영문본체 코퍼스) YOH0	강원권 음성인식 자원 회사 인의 목소리	TI	70시간의 비평사의 판 제발신자의 통신내용	사부설 편집	자주 별과	8 KHz	16 Bits	CD-ROM
SIRVA	음성인식의 훈련용 CSELT 진로법	ITT	남 550명, 여 550명의 발성	사부설 편집	남 남	8 KHz 8 KHz	16 Bits	
우송 DB	발성인식의 인식이 개발된 음성인식용	CSELT	남 40명, 여 40명의 발 성	TESSOS 이용 목소리로 합성	남 남	8 KHz 8 KHz	16 Bits	
CSELT 숫자 DB	발성인식의 인식이 개발된 음성인식용	CSELT	남 3명, 여 2명의 300 문장과 914 단어	테리널 프로그램	남	12.8KHz		
SPICOS 용언용성 DB	SPICOS 평가용		남 5 명, 여 5명의 조음한 방 음		남			
Edinburgh 연속용성 DB	연속용성본의	해인버러	15명의 303문장 4689 단어 발성	발성 리스트	남			
CTH		KTH	모음, 단어, 문장		남			
KTH		KTH	모음, 자음, 음절, 단어		남			
BREF	표지연말연구 표지음절의 인식이 개발 인식음절의 음 평가, 인, 합성제 새로운 모델 연구		11,002 문장		남			
GRECO(BDSONS)	NON-NATIVE 표현연구 합음보상임코디언 연구		32명 이 발성	스튜디오	남			Disk
ANATORSG10	음성인식용 화자 음성 진화기 음절별 비교		5 개국 19명이 1400개 의 코원숫자와 25000 개의 연결숫자 발성	타라널 프로그램 발성리스트	남 남	16 Bits		DAT
COST 209	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape
RSRE	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape
Edinburgh Corpus of Dictated Speech Edinburgh 연속 용성 DB	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape
STC	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape
RAESI 진로용성 DB	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape
STA-(UCL/NPL)Avey Speech Technology Assessment Normative DB	음성인식용 화자 음성 진화기 음절별 비교		남 4, 여 1명 발성	테리널 프로그램	남			Video-Tape

CSELT(Training Lesson)	본당이 MM에 기초한 인식기의 훈련을 위한 사전	CSELT	교필단어	9명의 화자가 1106 교필단어 발성	조용한 방	발성 리스트	남쪽		
CSELT(Test Lesson)	본당이 MM 에 기초한 인식기의 훈련을 위한 사전	CSELT	교필단어	9명의 화자가 1021 교필단어 발성	조용한 방	발성 리스트	남쪽		
CSELT (Telephone Application)	전화특성인식용의 훈련과 테스트	CSELT	숫자5개의 명동어 단어	남여 550명의 숫자5개의 명동어 발성	전화용실	음소리로 알린	남쪽	8 KHz	
DENSIPPT (DR. Nether Laboratories)	음향해결간의 관계연구	CSELT	단어	남여 각 40명이 13 개의 단어 발성	조용한 방	타이널 프릴프르	남쪽	12.8KHz	
EUROMO (ESPRIT Project)	음성학적 평가 cross-language와 음성소통		단어,숫자, 연속음성	남여 4 명이 545 이음절,28 문장 발성	무합실		남쪽	10KHz	16 Bits DAT
오스트레일리아 음성 DB	기본적 민족,교육수준 기준으로 다양한 민족,교육수준 기준으로 분포된 데이터 수집		교필단어, 단어	남여 각 55명의 숫자,단어 33 개, 19 방음실 개의 중모음이중모음, 33 개의 연속수			남쪽	44.1KHz,16 Bits	Video-Tape
VAI 단어용성 DB	음성DB구축	TI	PBS 단어,입력,숫자, 문장	503 ATR, PBS 발성 5445단어,110음절,교과과외 115 문장을	우형실/조용한 사무실	VDU 프릴프르	남쪽	48 KHz	DAT
ATR 문장용성 DB		ATR	단어,문장	남여 각 10명이 발성			남쪽		
ATR 단수화자	발음정확차용성인식	ATR	단어,문장	750단어,150문장을 250명이 발성			남쪽		
남성 DB							남쪽		
ATR 텍스트 음성 DB	음성합성연구용	ATR	문장	300문장을 남여 각 17명이 발성			남쪽		
ATR 명동어 음성 DB		ATR	단어,연음,연속수	58명이 5000단어,200문장 발성			남쪽		
일본 음향학회 연구용 연속성 DB-ATR			문장	총 64명이 9600 문장 발성			각수 발라	16KHz	
홍소벨린스문장에 의한 연속성성 코러스			문장	남여 각 18명이 11500 문장 발성			남쪽		
일본 음향학회 연구용 연속성성 DB-한내			단어,문장용성, 이음,문장	남여 각 10명이 발성			남쪽		
벨스크 연속음성 코러스			이음,문장	남여 각 75명이 323 발성			남쪽		
일본 문자성 중립 영역 연구에 의한 연속음성 DB			교필단어	남여 각 107 명이 136280 단어 발성			남쪽		
CSELT 40 단어 DB			교필단어				남쪽		
CSELT 24 단어 DB			연음 숫자				남쪽		
CSELT 40 연속 숫자 단어			연음 숫자				남쪽		
JEIDA	인식기의 성능 평가용		단말어,인단어, 연속숫자	남여 각 75명이 323 발성	15개의 관찰조건	발성 리스트 타이널 프릴프르	남쪽		
ETL			연속문장	107명 단음절,문미단어				20KHz, 12Bits	CD-ROM
홍소벨린스문장 영역 일본 연속성 중립 영역			문장용성	107 명이 136280 단어 발성				20 KHz, 12 Bits	CD, DAT, CD-RO
일본 연속성 DB			문장용성						M