# 통신망환경 한국어 공통음성 DB 구축

김상훈, 박문환, 김현숙
한국전자통신연구원 컴퓨터소프트웨어연구소 음성/언어정보연구센터

# Common Speech Database Collection for Telecommunications

**Sanghun Kim, Moonwhan Park and Hyunsuk Kim**
Speech /Language Technology Research Center, Computer Software Research Laboratory, ETRI
E-mail : {ksh, moonhpark, hyskim}@etri.re.kr

## Abstract

This paper presents common speech database collection for telecommunication applications. During 3 year project, we will construct very large scale speech and text databases for speech recognition, speech synthesis, and speaker identification. The common speech database has been considered various communication environments, distribution of speakers' sex, distribution of speakers' age, and distribution of speakers' region. It consists of Korean continuous digit, isolated words, and sentences which reflects Korean phonetic coverage. In addition, it consists of various pronunciation style such as read speech, dialogue speech, and semi-spontaneous speech. Thanks to the common speech databases, the duplicated resources of Korean speech industries are prohibited. It encourages domestic speech industries and activate speech technology domestic market.

## I. 서론

음성처리기술을 개발하기 위해서는 대규모의 음성/텍스트 DB가 필수적으로 요구된다. 이에 대다수의 국내 음성처리업체들은 개별적으로 DB를 구축하여 자체 엔진개발에 활용하고 있으나, 일반적으로 DB 구축에 많은 시간과 비용이 소요되어 외국업체에 비해 기술개발 경쟁력이 떨어지고 있다. 이미 미국, 유럽 등의 선진국에서는 산/학/연 공동으로 연구기관을 설립하여 공통음성 DB를 구축 보급하고 있다[1]. 특히 국내업체간 중복 DB 구축으로 인해 국가적으로 자원이 비효율적으로 활용되고 있어 음성정보처리 관련 사업자들의 공동 이익을 도모할 수 있는 공통음성 DB 구축이 시급하다. 국내에서는 음성정보기술산업지원센터(SITEC, 원광대)에서 자동차 산업 등 전통산업분야에 특화된 대규모 음성 DB를 구축하고 있으나 여전히 통신망환경에서는 음성처리업계의 요구사항을 충분히 반영하지 못하고 있다[2].

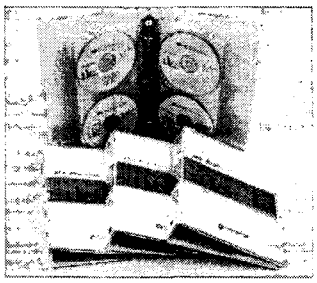

그림 1. 1차년도 공통음성 DB 구축 결과물

이에 한국전자통신연구원 음성/언어정보연구센터에서는 "언어정보처리기술 개발(2003.1-2005.12)" 사업을 통해 다양한 통신망환경에서 대규모 음성 DB를 구축 배포하여 국내 음성처리업체의 국제경쟁력을 강화할 수 있는 기반을 마련하고자 한다[3][4].

## II. 1차년도 DB 구축 내용

| 순번 | 목적 | 통신망 환경 | 화자수/발화수 | 발성내용/발성조건 | 비고 |
|---|---|---|---|---|---|
| 1 | 음성인식용 단어 | 휴대폰 | 1,000명/10만 발화 | -주식상장회사명, 지명, 인명, 상호명, 제품명, PC명령어, PDA명령어, 일반명사<br>-전화망인 경우, 전화망 인터페이스 보드는 NMS 계열 및 Dialogic JCT 계열을 이용."디지털보드:아날로그보드"="50:50" 비율로 수집. 유선전화기 사용을 유도하고, 무선전화기의 사용은 10% 미만이 되도록 함. 전화기 모델은 제한 두지 않음<br>-남/녀 비율은 50:50 으로 하며 최대 5%까지의 차이 허용. 연령별 구성은 "10대 : 20대 : 30대: 40대 이상" 의 구성비를 "20 : 30 : 30 : 20"으로 하고 오차는 각 5% 이하 허용. 지역별 구성은 "서울/경기: 경상: 충청: 전라: 제주 강원"의 구성 비율을 "40 : 20 : 15 : 15 : 10" 으로 하고 최대 2%까지 차이 허용 | |
| 2 | | 유선망 | | | |
| 3 | | VoIP | | | |
| 4 | | 마이크(중가) | | | |
| 5 | | 마이크(저가) | | | |
| 6 | | 헤드셋 | | | |
| 7 | 음성인식용 숫자 | 휴대폰 | 1,000명/10만 발화 | -번호독식 방식과 봉독식 방식에 대해 수집<br>-전화번호, 주민등록번호, 계좌번호 등으로 구성된 1-16연 숫자<br>-전화번호, 계좌번호 중 '-'를 '에', 다시', '국'으로 발성<br>-일부는 한자식과 우리말 숫자 혼합형으로 발성<br>-봉독식 방식은 99,999까지의 무작위 숫자를 한자식으로 발성하고, 일부는 우리숫자로 발성 | 성별/연령/지역별 구성 비율은 단어와 동일 |
| 8 | | 유선망 | | | |
| 9 | | VoIP | | | |
| 10 | | 마이크(중가) | | | |
| 11 | | 마이크(저가) | | | |
| 12 | | 헤드셋 | | | |
| 13 | 음성인식용 낭독체/준낭독체 문장 | VoIP | 1,000명/10만 발화 | -낭독체 문장은 발성목록은 방송뉴스에서 추출<br>-준낭독체 문장은 발성목록 없이 화자가 즉흥적으로 주어진 주제에 대해 발성(예: 자기 소개하기, 가장 친한 친구 이야기 하기, 자신의 학교 소개하기 등) | " |
| 14 | | 마이크(중가) | | | |
| 15 | | 마이크(저가) | | | |
| 16 | | 헤드셋 | | | |
| 17 | 음성인식용 대화체 문장 | 유선/휴대폰 전화망 | 250명/2,500 대화 | -예약, 은행, 증권, 관광안내, 텔레쇼핑 등 최소 30개의 시나리오를 작성하여 그 중 10개를 선택<br>-각 화자당 10개의 상황에 대한 대화음성을 실제 call center에서 수집<br>-1대화당 평균 20문장 또는 5분 이상이 되도록 구성 | " |
| 18 | 언어모델링용 문장 | 텍스트 | 2,000만 어절 수동/4,000만 어절 자동 | -신문기사 대상 띄어쓰기 및 철자오류 검증<br>-심볼, 영문 등을 한글로 변환 | |
| 19 | 음성합성용 문장 | 정보전달용 낭독체 | 남녀 1인/1만 문장 | -방송뉴스에서 추출한 낭독체 문장 발성<br>-트라이폰 분포 고려 | |
| 20 | 화자인식용 단어/문장 | 휴대폰 | 250명/45만 발화 | -2연, 4연 숫자음 및 10개의 질문에 대한 단답형 대답과 10개의 단문을 수집<br>-화자는 정해진 시차 간격에 따라 4차례 발성에 참가<br>-시차 간격은 1주, 1달, 3달임. 1주 간격의 경우, 2일의 오차 허용. 1달 간격의 경우, 5일의 오차 허용. 3달 간격의 경우, 10일의 오차 허용<br>-각 시차별 1명당 1차례 발성량은 2연 숫자 100개, 4연 숫자 250개, 10개의 단답형 대답 및 10개의 단문을 각 5회씩 한번 발성시 총 450개 발성 | 성별/연령/지역별 구성 비율은 단어와 동일 |
| 21 | | 유선망 | | | |
| 22 | | VoIP | | | |
| 23 | | 마이크(중가) | | | |
| 24 | | 마이크(저가) | | | |
| 25 | | 헤드셋 | | | |

## III. 1차년도 DB 구축 후 애로사항 및 검토사항

1차년도 대량의 DB를 구축하면서 DB구축 담당자의 애로사항을 정리하면 다음과 같다.

- 준낭독체 녹음: 녹음작업진행 중 가장 화자들이 어려워하는 부분임. 녹음 진행시간도 상당히 소요됨. 전사 작업의 난이도가 높았으며, 전사 관련 지출비용 중 상당부분을 차지한다
- 화자섭외: 기존 1,000명 외에 새로운 1,000명에 대한 화자녹음이 어렵다
- 마이크 숫자: 동시에 녹음할 수 있는 마이크 수량은 2-3개 적당하다
- 발화개수: 1회에 480토큰 발성(80토큰*6회 반복발성)으로 화자와 모니터 요원이 공통으로 지루해 함. 토큰의 갯수를 줄이는 방향으로 검토 요망된다
- 녹음장비: PC 3대와 믹서기 2대 등 각종 장비기 비대하고, 파일의 갯수가 너무 많아 관리가 어려움. PC 한대로 다채널 동시 녹음하는 툴이 필요하다
- 발성목록: 예를 들어 "당신의 초등학교 이름을 말씀해주세요."와 같은 질문을 텍스트로 보고 답변에 대한 음성을 DB화 할 경우, 녹음에 소요되는 시간이 3배가 걸리게 된다

최초 공통음성 DB 규격이 현실에 맞지 않는 경우가 발생하여 이를 검토, 2차년도에 보완하고자 한다. 다음은 2차년도 DB 수집시 보완내용을 정리한 것이다.

- 화자 및 환경분포
  - 환경별로 모으는 것을 현실적으로 수정하자. 지하철 같은 곳에서는 모을 수 없다
  - 10대는 중/고등학생인데 카드번호, 계좌번호, 주식 등 현실적으로 맞지 않는 것을 수집했기 때문에 어색하다
- DB 수집방법
  - 마이크를 현실화하자. 업체에서 실질적으로 사용하는 저가마이크(1만원대 이하)를 사용하자
  - PDA와 PC 모니터에 부착한 마이크도 사용하자
  - 유선전화의 경우, 철저한 품질관리가 필요하다
  - 유선전화는 수집해서 사후 관리하는데 드는 비용이 너무 크다
  - 수집 즉시 검토할 수 있게 하자. 수집하면서 모니터링 할 수 있게 하자
- 발성목록
  - 단어, 낭독체문장에 철자나 띄어쓰기 오류가 많다
  - 발음전사는 예외발음만 괄호를 사용하여 처리하자
  - 숫자는 철저한 발음전사가 필요하다
  - 여러개의 음성파일과 하나의 전사파일로 구성하는 것은 문제가 많음. 하나의 음성에 하나의 전사파일로 하자
  - 음성구간 마킹, 잡음 마킹 등의 정보를 없애자

2차년도 DB 구축은 1차년도에 비해 화자섭외가 어려운 점이 있으나 1차년도 DB 구축시 발생한 문제점을 이미 파악하고 있고, 수동/자동 검증툴을 확보하고 있어 1차년도 보다 효율적으로 고품질 DB를 수집할 수 있을 것으로 보인다.

## III. 2차년도 DB 구축 계획

ETRI에서는 2차년도 DB 구축을 위해 업체 의견을 수렴하였다. 수렴한 결과를 요약하면 다음과 같다.

- ◎ 원격음성 DB 구축 필요
- ◉ 잡음 DB 필요
- ◎ 외국어 합성 DB 필요
- ◉ 외국어 인식용 단어 DB 필요
- ◎ 숫자음 seed 모델용 수동 음소분할 필요
- ◎ 휴대폰/PDA/텔레메틱스 등 모발환경DB 구축 필요
- ◎ 공통DB 보다 응용서비스에 적합한 DB 수집 필요
- ◎ EU의 AURORA 표준 trace 필요
- ◎ 콘덴서 마이크 사용 필요
- ◎ 화자인식 DB의 경우, 시차별 DB는 매우 적절
- ◉ 화자인식 DB의 경우, 5-10음절 키워드가 적당
- ◎ 화자인식 DB의 경우 DB, 수집은 어렵고 DB 수요는 작음
- ◉ 합성 DB의 경우, 운율레이블링 필요
- ◉ 한국인의 영어발음 DB 필요
- ◉ 어린이용 대화체 운율/어투 모델링용 합성 DB필요
- ◉ 마이크 환경, 일상 대화 DB 필요

수렴결과 중 일부는 2차년도 DB 구축에 반영하였다. 2차년도 DB 구축계획은 (표 2) 와 같다.

표 2. 2차년도 DB 구축계획

| 목적 | 통신망 환경 | 화자수/ 발화수 | 발성내용/ 발성조건 |
|---|---|---|---|
| 음성인식용 단어 | 휴대폰 | 1,000명/ 10만 발화 | 주식상장회사명, 지명, 인명, 제품명, PC명령어, PDA 명령어, 일반명사로 구성. 성별, 연령별, 지역별 분포 고려 |
| | 유선망 | | |
| | VoIP | | |
| | 마이크(증가) | | |
| | PDA | | |
| | 헤드셋 | | |
| 음성인식용 숫자 | 휴대폰 | 1,000명/ 10만 발화 | 1-16연숫자. 번호독식/ 봉독식 발성, 계좌번호, 단위, 전화번호, 주민등록번호로 구성. 성별, 연령별, 지역별 분포 고려 |
| | 유선망 | | |
| | 마이크(증가) | | |
| | PDA | | |
| | 헤드셋 | | |
| 음성인식용 낭독체문장 | VoIP | 1,000명/ 10만 발화 | 각 화자가 100 문장씩 발성한 낭독체 방송뉴스 문장 |
| | 마이크(증가) | | |
| | PDA | | |
| | 헤드셋 | | |
| 음성인식용 대화체 문장 | 유선/휴대폰 전화망 | 500명/ 5,000대화 | 가상 Call center에서 고객과 상담원 대화 녹취(시나리오 사용) |
| 언어모델링 용 문장 | 텍스트 | 2,000만 어절 수동/ 4,000만 어절 자동 | 일간지 신문 3,000만 어절 수동 철자/띄어쓰기 수정. 9,000만 어절 자동 철자/띄어쓰기 수정 |
| 음성합성용 문장 | 정보전달용 낭독체 | 남녀 1인/ 1만 문장 | 남녀 성우 20여명 후보에서 선호도 평가 후 2명 선정 |
| | 대화체 문장 | 남녀 2인/ 1만 문장 | 2인이 서로 대화하는 음성 녹음 |

2차년도 DB 구축 일정은 다음과 같다.

- 2003년 2월: 공통음성 DB 규격 보완 완료
- 2003년 3월-7월: DB 구축
- 2003년 8월-9월: DB 검증
- 2003년 10월: DB 배포

## IV. 결론

이번에 구축한 대규모 한국어 공통음성 DB는 음성인식/음성합성 엔진을 개발하기 위한 기본 자료로 매우 유용할 뿐만 아니라 한국어 공통음성 DB 구축을 통해 국내 산업체간 음성/텍스트 DB 구축의 중복투자 방지하고 국내 음성정보산업 경쟁력 강화 및 음성서비스 시장을 활성하는데 크게 기여할 것으로 판단된다. 본 DB는 다양한 통신망 환경, 성별, 연령별, 지역별 화자 분포가 고려된 국내 최대의 한국어 공통음성 DB로써 다양한 영역(숫자, 단어, 문장)의 발성음성으로 구성되었으며, HCI (Human Computer Interface), CTI (Computer Telephony Interface), 텔레매틱스 (Telematics), 생체정보인식, 시각장애인용 음성응용, 자동통역 등 각종 음성인식 및 합성엔진 개발에 활용될 것이다.

ETRI 음성/언어정보연구센터에서는 지속적으로 음성기술의 발전방향에 따라 요구되는 DB를 시기적절하게 공급하여 국내업체의 경쟁력을 강화하고자 하며, 향후 각종 음성언어정보의 체계적인 표준화작업을 수행하여 DB의 활용성을 높이는데 최선을 다하고자 한다.

## 감사의 글

본 연구는 정보통신부 출연 2002년도 "음성정보처리 기반기술" 과제의 일환으로 수행되었습니다.

## 참고문헌

[1] ETRI 음성/언어정보연구센터: http://voice.etri.re.kr

[2] 김봉완, 이용주, "음성정보기술산업지원센터의 음성코퍼스 구축 현황 및 계획," *한국음향학회 하계학술대회논문집*, 제21권, 제1(s)호, pp.49-52, 2002

[3] 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성 DB 구축," *한국음향학회 하계학술대회논문집*, 제21권, 제1(s)호, pp.21-24, 2002

[4] 오승신, "공통음성 DB 구축을 위한 발성목록 설계," *한국음향학회 하계학술대회논문집*, 제21권, 제1(s)호, pp.29-32, 2002