공통음성 DB 구축

김상훈, 오승신, 정호영, 전형배, 김정세
한국전자통신연구원 네트워크연구소 음성정보연구센타

Common Speech Database Collection

Sanghun Kim, Seungshin Oh, Ho-Young Jung, Hyung-Bae Jeong, and Jeong-Se Kim
SpeechTechnology Research Center, Network Laboratory, ETRI
E-mail :{ksh, oss63354, hbjeon, hjung, jungskim}@etri.re.kr

요약

본 논문은 ETRI 음성정보연구센터에서 추진하고 있는 공통음성 DB 구축에 관하여 기술한다. 총 3 년(2001.11~2004.10) 동안 음성인식, 음성합성, 화자인식 등 다양한 용도의 음성 DB 를 수집할 예정이며, 1 년차인 2002 년에는 총 14 종의 음성 DB 를 수집할 계획이다. 공통 음성 DB 는 다양한 통신망(마이크, 헤드셋, VoIP, 유무선 전화망), 지역, 성별, 발성환경(사무실, 지하철, 도로 등)을 고려하여 설계하였으며, 발성대상은 숫자, 단어, 문장이고, 발성방법은 자유발화, 대화체, 낭독체 등 다양한 스타일의 음성 DB 로 구성되어 있다. 이에 본 논문에서는 총 14 종에 해당하는 공통음성 DB 의 구축내역과 구축방안 및 DB 구축 일정에 관해 기술하고자 한다.

1. 서론

음성정보처리기술을 개발하기 위해서 일반적으로 음성 DB 구축에 많은 시간과 비용이 투여되고 있다. 특히 국내업체간 중복 구축된 음성 DB 로 인해 국가적으로 자원이 비효율적으로 활용되고 있으며, 외국업체에 비해 경쟁력을 약화시키는 주요 원인이 되고 있다. 이에 따라 음성정보처리 관련 사업자들의 공동 이익을 도모할 수 있는 공통음성 DB 구축이 절실해 지고 있다. 특히 미국, 유럽 등의 선진국은 음성정보처리 기술개발에 필수적인 음성 DB 의 보급을 위해서 국가적인 차원에서 연구소, 대학, 업체가 공동으로 연구기관을 설립하여 공통 음성 DB 를 구축/보급하고 있으며 미국 LDC, 유럽 ELRA 등이 대표적인 기관이다. 국내에서는 ETRI, KT 등에서 소규모 음성 DB 를 구축하여 배포한 적이 있으나 국내 음성정보처리 업계에 기여도가 미미한 실정이었고 음성정보기술산업지원센터(SITEC, 원광대)에서 자동차 산업 등 전통산업분야에 대한 지원을 위해 자동차 내에서의 환경이나 제조현장등의 소음환경에 특화된 대규모 음성 DB 를 구축하고 있으나 여전히 통신망환경에서는 음성정보처리업계의 요구사항을 충분히 반영하고 있지는 못하고 있는 실정이다. 이에 ETRI 음성정보연구센터에서는 다양한 통신망환경에서 대규모 음성 DB 를 구축하여 국내업체에 경쟁력을 강화할 수 있는 기반을 마련하고자 한다.

2. DB 구축 계획

ETRI 음성정보연구센터에서 추진하고 있는 공통음성 DB 구축계획은 표 1 과 같다.

표 1: 공통음성 DB 구축계획

용도	환경	내역	1차	2차	3차
음성인식 단어	휴대폰	-10,000단어 10세트	0	0	
	유선망	-100단어씩 1,000명 발성	0	0	
	VoIP		0	0	
	마이크	-성별,연령별,지역별	0	0	
	헤드셋		0	0	
음성인식 숫자음	휴대폰	-(1,3,4,7,10,16)자리 숫자	0	0	
	유선망		0	0	
	VoIP	-주민번호, 계좌번호	0	0	
	마이크	-단위, 우리말 숫자 포함		0	0
	헤드셋	-단어DB와 동일한 수집환경		0	0
음성인식 문장	낭독체	-총 15만 낭독체 문장 -총 15만 준낭독체 문장 -총 3,000명 -성별,연령별,지역별	0	0	0
	대화체	-응용영역 선정(핸드폰,유선) -총 15,000대화 음성 -250명/250명/1,000명 총1,500명 -시나리오 기반	0	0	0
	텍스트	-375만/375만/750만 문장씩 총 1,500만 문장 -띄어쓰기 및 철자오류 검수	0	0	0
음성합성	정보전달	-매년 4,000문장 x 남녀 각 1인 총 4인	0	0	
	대화체	-총 10,000문장 x 4인		0	
	특정발성	-총 2,000문장 x 10인			0
화자인식	2연숫자	-매년 성,연령별 250명 발성 -시차별(주,달,계절) 4회 수집 -유선,핸드폰, 마이크,헤드셋 환경	0	0	
	4연숫자		0	0	
	단문				0

3. DB 구축 내역

본 논문에서는 텍스트 DB 를 제외한 음성 DB 구축내역에 관해 상세히 기술한다. 1 차년도(2001.11~2002.10)에는 음성인식용 단어/숫자/문장 음성 DB, 대화체 음성인식용 DB, 화자인식용 DB 및 음성합성용 DB 를 수집한다.

3.1 음성인식용 단어/숫자/문장 음성 DB

음성인식용 DB 인 경우, 마이크/헤드셋/VoIP 망을 통하여 단어, 숫자, 문장음성 DB 를 총 1,000 명의 화자에 대하여 구축하며, 유/무선 전화망인 경우, 단어, 숫자 음성 DB 를 총 1,000 명의 화자에 대하여 구축한다.

가. 발성목록

(ㄱ) 단어인 경우, 상장회사명, 지명, 인명, 상호명, 제품명, PC 명령어, PDA 명령어, 그 외 일반명사로 구성된다.

(ㄴ) 숫자의 경우, 번호독식 방식과 봉독식 방식에 대해 수집한다. 번호독식방식은 1, 3, 4, 6, 7, 10 자리 일련숫자와 전화번호, 은행계좌번호로 구성된다. 6, 7, 10 자리 일련숫자 중 일부는 우리말 숫자로 수집한다. 전화번호, 은행계좌번호 일련숫자 중 '-'를 '에', '다시', '국' 으로 발성하고, 일부는 한자식과 우리말 숫자 혼합형으로 발성한다. 봉독식 방식은 99,999 까지의 무작위 숫자를 한자식으로 발성하고, 일부는 우리숫자로 발성한다.

(ㄷ) 문장의 경우, 낭독체문장 50,000 문장과 준낭독체문장 50,000 문장 총 100,000 문장을 수집한다. 낭독체문장 발성목록은 방송뉴스에서 추출한다. 준낭독체 문장은 발성목록 없이 화자가 즉흥적으로 어떤 주제에 대해 발성하는 것을 말한다(예: 자기 소개하기, 가장 친한 친구 이야기 하기, 자신의 학교 소개하기, 친구에게 e-mail 쓰기, 어제 하루 일과를 정리하기 등).

나) 수집방법

(ㄱ) 펜티엄 III 이상의 데스크탑 PC 에서 녹음하며, 노트북은 허용하지 않는다.

(ㄴ) 사운드 카드는 SoundBlaster 128 이상을 사용하며, On-board 형태의 사운드카드는 허용하지 않는다.

(ㄷ) 수집 시스템은 중가의 마이크와 저가의 마이크, 헤드셋으로 구성된다. 중가 마이크는 2 종류의 마이크 중 한 개를 선택하고, 저가 마이크는 5 종류의 마이크를 20%씩 비율로 사용한다. 중가의 헤드셋에 대해 80%인 800 명의 음성을 녹음하고, 저가의 헤드셋에 대해 20%인 200 명의 음성을 녹음한다.

(ㄹ) 마이크는 PC 모니터 양 옆에서 화자의 입을 향하여 20cm 거리에 위치시키며, 헤드셋의 마이크는 입과 동일한 높이 입 가장자리에서 1.5cm 떨어진 곳에 둔다.

(ㅁ) VoIP DB 를 수집하기 위해 서로 다른 건물에 설치한 PC 를 초고속 통신망에 연결하고, 한쪽 PC 에서 헤드셋을 통해 녹음한 음성인식용 단어, 숫자음성을 H.323 프로토콜을 이용한 VoIP 망을 통해 전송한다. 다른 쪽 PC 에서는 VoIP 로 전송된 음성을 저장한다. 다양한 트래픽 상황을 고려하기 위해 "10 시~12 시 : 1 시~5 시 : 7 시~10 시" 시간대별로 약 "30 : 40 : 30" 의 비율로 수집한다.

(ㅂ) 전화망인 경우, 전화망 인터페이스 보드는 NMS 계열 및 Dialogic JCT 계열을 이용한다. "디지털보드:아날로그보드" = "50:50" 비율로 수집한다. 유선전화기 사용을 유도하고, 무선전화기의 사용은 10% 미만이 되도록 한다. 전화기 모델은 제한을 두지 않는다.

(ㅅ) 핸즈프리의 사용은 음성의 품질을 들어보고 결정하며, 핸즈프리 사용시에는 사용 핸즈프리 제품명을 기입한다. 핸즈프리 사용자는 총 5%를 넘지 않도록 한다.

(ㅇ) 전화망 구성에 관해서는 유선망의 경우 시내, 시외의 제한은 두지 않는다. 무선망의 경우 사업자별 분포가 (011, 017) : (016, 018, 019) 의 비율이 "60 : 40" 이 되도록 한다. 수집비율에서 ±10% 오차를 허용한다.

다) 화자분포

(ㄱ) 남/녀 성별비율은 50:50 으로 하며 최대 ±5%까지의 차이를 허용한다.

(ㄴ) 연령별 구성은 "10 대 : 20 대 : 30 대 : 40 대 이상" 의 구성비를 "20 : 30 : 30 : 20" 으로 한다. 오차는 각 ±5% 이하만을 허용한다. 10 대는 중고생 재학생을 의미한다.

(ㄷ) 지역별 구성은 "서울/경기: 경상: 충청: 전라: 제주/강원" 의 구성 비율을 "40 : 20 : 15 : 15 : 10" 으로 한다. 최대 ±2%까지의 차이를 허용한다. 지역의 기준은 화자의 초등학교 재학시 거주지로 한다. 기본적으로 서울 거주자를 대상으로 수집하며, 지방 거주자에 대해서는 표준어로 발음할 것을 유의시킨다.

라) 수집환경

(ㄱ) PC 마이크/헤드셋 환경인 경우, 조용한 사무실 환경에서 수집하며, 최소 SNR 25dB 이상 되어야 한다. 무향실 또는 방음실에서는 녹음하지 않도록 한다.

(ㄴ) 유선망인 경우, 사무실, 집, 공중전화 환경을 포함한다. 각 환경별 최소 10% 이상이 포함 되어야 하며, SNR 은 최소 15dB 이상 되어야 한다.

(ㄷ) 무선망인 경우, 사무실, 집, 거리, 지하철 역사내, 백화점 환경을 포함한다. 그리고 각 환경별 최소 10% 이상이 포함 되어야 하며, SNR 은 최소 15dB 이상 되어야 한다.

3.2 대화체 음성인식용 DB

대화체 문장음성 인식용 DB 는 크게 2 가지의 DB 를 수

집한다. 첫째로 음향모델링을 위한 시나리오 기반 call center 고객/상담원과의 대화음성과 대화체 언어모델링을 위한 call center 고객/상담원과의 7,500 대화문장을 수집한다. 시나리오 기반 대화음성은 1 인당 10 대화씩 250 명이 발성한 총 2,500 대화음성을 수집할 계획이며, 대화체 언어모델링용 텍스트 DB 는 실제 서비스되고 있는 상황을 전사한다.

가) 화자분포
음성인식용 단어/숫자/문장 음성 DB 의 화자분포 조건과 동일함.

나) 대화영역
(ㄱ) 시나리오 기반일 경우, 예약(호텔, 식당, 기차, 고속버스, 항공기, 렌터카, 콘도, 영화, 연극, 음악회, 스포츠 등의 예약, 확인, 취소), 은행(계좌조회, 자금이체), 증권(매도/매수주문, 조회, 시황), 관광안내(여행상품, 요금, 추천 관광지, 교통편 문의, 시간), 텔레쇼핑(상품특징, 가격, 수량, 구입방법, 결재방법 등 홈쇼핑 주문) 등 최소 30 개의 시나리오를 작성하여 그 중 10 개를 선택, 각 화자당 10 개의 상황에 대한 대화음성을 수집한다.
(ㄴ) 언어모델용 텍스트 DB 인 경우, 실제 call center 상담내용 중 증권/은행 영역을 최소 30%이상 반영한다.

다) 수집환경
음성인식용 단어/숫자/문장 음성 DB 의 수집환경과 동일함.

라) 수집방법
(ㄱ) Call center 실제 서비스환경을 그대로 이용하되 NMS AG4000 2E1 또는 Dialogic JCT series(Digital) 보드를 사용해야 한다.
(ㄴ) 유선전화 및 휴대폰 제조사의 모델은 특정하지 않는다.
(ㄷ) 시나리오 방식인 경우, 상담원과 고객(화자)은 시나리오 및 고유명사를 미리 숙지하도록 한다. 특히 상담원은 화자의 자연성을 유도할 수 있도록 대화를 유도한다.
(ㄹ) 상황에서 날짜와 시간, 번호들의 다양한 숫자의 선택을 주지시킨다.
(ㅁ) 각 상담원이 최대 10 명의 화자와 대화하는 음성을 수집한다. 따라서 상담원은 최소 25 명 이상이 참여해야 한다.
(ㅂ) 시나리오는 다양한 고유명사(사용자이름, 은행사명, 증권사명, 호텔명, 주식 상장사명, 대학교명, 부서명 등), 숫자(주민번호, 패스워드, 계좌번호, 주식수, 이체금액 등)를 포함해야 하며, 문제해결이 되도록 유도하여야 한다.
(ㅅ) 시나리오 중 화자의 주문을 확인하는 /예, 아니오/ 질문을 최소 1 개 이상은 포함하여야 한다.
(ㅇ) 1 대화당 평균 20 문장 또는 5 분 이상이 되도록 구성하고, 휴대폰의 특성인 음성의 끊김, echo 등은 오류로 처리하고 재발성시킨다.
(ㅈ) 상담원과 화자간 대화가 겹치거나 외부잡음으로 잘 들리지 않을 때, 다시 물어서 대화가 자연스럽게 진행되도록 한다.
(ㅊ) 잘못발성, 파형잘림, 이해할 수 없는 소리, 혀짧은 소리 등은 오류이므로 재발성하도록 하고, 지방색, 망설임, 화자잡음(간투사, 입술소리, 기침소리 등)은 허용한다.
(ㅋ) 음성신호레벨은 16bit 로 변환하였을 때, 최대 피크 amplitude 가 10,000 정도가 되어야 하며, SNR 은 15dB 이상이 되도록 한다.

3.3 화자인식용 DB
250 명을 대상으로 마이크, 헤드셋, VoIP, 유/무선 전화망 환경에서 화자인식용 2 연, 4 연 숫자음 및 10 개의 질문에 대한 단답형 대답과 10 개의 단문을 수집한다.

가) 발성목록
(ㄱ) 0~9 사이의 숫자로 이루어진 2 연 숫자 100 개를 대상으로 하며, 각 화자는 임의로 추출된 20 개를 발성한다.
(ㄴ) 0~9 사이의 숫자로 이루어진 4 연 숫자 1000 개를 대상으로 하며, 각 화자는 임의로 추출된 50 개를 발성한다.
(ㄷ) 발성목록 없이 개인정보와 관련된 10 개의 질문이 주어지고, 각 화자가 이 질문에 단어나 짧은 문장으로 답하는 형식이다.
(ㄹ) 추가로 3 어절 이내로 구성된 단문 10 개를 발성한다.

나) 화자 관리
(ㄱ) 각 화자는 정해진 시차 간격에 따라 4 차례 DB 구축에 참가한다.
(ㄴ) 시차 간격은 1 주, 1 달, 3 달이다. 1 주 간격의 경우 ±2 일의 오차를 허용한다. 1 달 간격의 경우 ±5 일의 오차를 허용한다. 3 달 간격의 경우 ±10 일의 오차를 허용한다.
(ㄷ) 전체 250 명을 4:4:2 의 비율로 3 그룹으로 나누어 관리한다. 첫번째 그룹은 100 명이며, 1 주 간격으로 4 차례 정해진 목록을 발성한다. 두번째 그룹도 100 명이며, 1 달 간격으로 4 차례 정해진 목록을 발성한다. 세번째 그룹은 50 명이며, 3 달 간격으로 4 차례 정해진 목록을 발성한다.
(ㄹ) 화자는 임의의 한 그룹에 소속되어 주어진 시차 간격대로 4 차례 발성한다. 각 시차별 1 명당 1 차례 발성량은 2 연 숫자 20 개 * 5 회 =100 개, 4 연 숫자 50 개 * 5 회=250 개 및 10 개의 단답형 대답과 10 개의 단문을 각 5 회씩 한번 발성시 총 450 개를 발성하게 된다. 전체적으로는 시차별 4 차례 반복하여 발성하므로 1 명당 1800 번 발성한다.

다) 수집 방법
(ㄱ) 펜티엄 III 이상의 PC 를 이용하기를 권장하

며 on-board 형태는 허용하지 않는다. 또한 노트북 사용은 허용하지 않는다. 사운드카드는 SoundBlaster 128 이상을 권장한다.

(ㄴ) 한번 발성시 마이크 2 개와 헤드셋으로 동시에 수집한다. 마이크는 중가와 저가의 2 종류로 수집하며 250 명 화자 모두 중가와 저가 마이크 DB 수집에 참여한다. 헤드셋의 경우도 중가와 저가의 2 종류로 수집한다. 250 명의 화자를 125 명씩 분류해서 한쪽은 중가, 다른 한쪽은 저가의 헤드셋으로 DB 를 수집한다.

(ㄷ) 중가 헤드셋을 위한 125 명은 1 주 간격 시차 그룹에서 50 명, 1 달 간격 시차 그룹에서 50 명, 3 달 간격 시차 그룹에서 25 명 선발한다. 저가 헤드셋을 위한 125 명은 이미 선발한 인원을 제외한 나머지를 대상으로 한다.

(ㄹ) 헤드셋은 입과 동일한 높이로 입 가장자리에서 1.5cm 떨어진 곳에 위치하도록 한다. 마이크는 화자 입과 약 20cm 떨어진 곳에 위치하도록 한다.

(ㅁ) 전화망의 경우, 전화망 인터페이스 보드는 NMS 계열 및 Dialogic JCT 계열을 이용한다. 보드별 비율은 디지털 보드 아날로그 보드 =50:50 으로 나누어 수집한다.

(ㅂ) 디지털 보드는 1 주 간격 시차 그룹에서 50 명, 1 달 간격 시차 그룹에서 50 명, 3 달 간격 시차 그룹에서 25 명, 총 125 명을 대상으로 한다. 아날로그 보드는 이미 선발한 인원을 제외한 나머지 125 명을 대상으로 한다.

라) 수집 환경

음성인식용 단어/숫자/문장 음성 DB 의 수집환경과 동일함.

마) 화자분포

음성인식용 단어/숫자/문장 음성 DB 의 화자분포 조건과 동일함.

3.4 음성합성용 DB

정보전달용 낭독체 음성합성 DB 는 남/녀 아나운서가 발성한 낭독체 각 10,000 문장씩 총 20,000 문장 남/녀 아나운서가 발성한 한국어 다이폰 세트(2,000 여 단어), 남/녀 아나운서가 발성한 한국어 반음절 세트(2,000 여 단어), 남/녀 문장음성 발성시 래링고 신호를 동시 녹취한다.

가) 수집방법

(ㄱ) 마이크의 주파수 응답특성이 중음대역과 고음대역의 특성이 좋은 것을 사용하도록 한다

(ㄴ) 음성과 래링고신호는 DAT reccrder 를 이용하여 스테레오로 녹취하여야 한다.

(ㄷ) DAT recorder 가 아닌 수집시스템의 경우, 44.1kHz 또는 48kHz 로 sampling 한 데이터를 PC Hard disk 로 저장해야 하며, 이로부터 16kHz 데이터를 down-sampling 하여 수집하도록 한다.

나) 수집환경

(ㄱ) 반사음 및 외부잡음이 없도록 soundproof booth 에서 녹음해야 한다.

(ㄴ) 테이블의 떨림을 최소화 해야 하며 데스크형 스탠드로 사용을 할 경우는 데이블 위에 두꺼운 천을 깔아서 사용한다.

(ㄷ) 파열음(예:ㅋ,ㅌ,ㅍ,ㅊ)이 포함된 단어를 발음할 때, 팝 노이즈가 발생하지 않도록 마이크를 입술의 위 또는 옆에서 향하도록 위치를 조정하거나 팝 필터(Pop Filter)나 윈드스크린(Wind Screen)을 사용하도록 한다.

다) 발성방법

(ㄱ) 목소리의 음질과 평균 녹음레벨, 평균 피치 그리고 평균 템포 등이 일정하게 유지되도록 녹음해야 한다.

(ㄴ) 정확한 내용과 정보를 전달하기 위해서는 단어마다 정확한 발음을 해야 한다.

(ㄷ) 래링고신호를 깨끗하게 받기 위해서 발성자의 목부위에서 가장 떨림이 큰 부분에 센서를 대어야 한다.

(ㄹ) 문장음성의 경우, 문장간 휴지길이(pause length)를 약 1 초 이상 두면서 발성한다.

(ㅁ) 다이폰/반음절의 경우, 발성단위간 휴지길이(pause length)를 약 1 초 이상 두면서 발성한다.

(ㅂ) 다이폰/반음절의 경우, 발음하기에 어려운 발음이 많이 있으므로 명료한 발음이 나올때까지 발성시킨다.

4. DB 구축 일정

ETRI 음성정보연구센터의 DB 구축일정(잠정안)은 다음과 같다.

- □ 2002. 3: 음성 DB 구축안 완료, 발성목록 정리
- □ 2002. 4: 발성목록 최종 검증, DB 구축 시작
- □ 2002. 5: 5% DB 구축, 자동/수동검증 수행
- □ 2002. 7: 40% DB 구축, 자동/수동검증 수행
- □ 2002. 8: 100 % DB 구축. beta version 구축
- □ 2002. 9: Release version 구축, 배포 및 관리

5. 결론

ETRI 음성정보연구센터에서는 지속적으로 음성기술의 발전방향에 따라 요구되는 DB 를 시기적절하게 공급하여 국내업체의 경쟁력을 강화하고자 하며, 향후 각종 음성언어정보와 체계적인 표준화작업을 수행하여 DB 의 활용성을 높이는데 최선을 다하고자 한다.

감사의 글

본 연구는 정보통신부 출연 "음성정보처리기반기술" 과제의 일환으로 수행되었습니다.

참고문헌

1] ETRI 음성정보연구센터 홈페이지: http://voice.etri.re.kr