공통 음성 DB 구축을 위한 발성목록의 설계

오승신
한국정보통신연구원 음성정보연구센터 음성 DB 기술팀

Creation of scripts for building common speech database

OH, Seung-Shin
Speech Database Team, Speech Information Technology Center
Electronics and Telecommunications Research Institute
e-mail: oss63354@etri.re.kr

요약: 본 논문에서는 ETRI 의 음성정보연구센터에서 추진하고 있는 공통 음성 DB 구축을 위한 발성목록의 설계 방법 및 그 내용에 대해 기술한다. 이 공통 음성 DB 는 국내의 학계와 연구기관, 산업체에 배포하여 음성기술개발에 활용될 수 있도록 하려는 목적으로 구축되는 대규모의 DB 인 만큼, 다양한 활용 분야를 고려하여 설계되었다. 따라서 내용적으로 중립성을 유지하면서도 효율성과 기능성을 고려하여 설계한 것이 이번 발성목록 설계의 특징이라고 할 수 있다. 이번 발성목록은 크게 음성 인식용 DB 와 낭독체 합성용 DB, 대화체 합성용 DB, 그리고 화자 인식용 DB 분야로 나뉘어진다. 이 논문에서는 각 DB 종류별로 발성 목록의 내용과 작성 방법, 그리고 이들을 작성함에 있어서 고려된 사항 등을 기술한다.

1. 서론

ETRI 에서는 음성기술개발에 필수적인 음성 DB 의 보급을 위해 대규모의 공통 음성 DB 의 구축 사업을 진행하고 있다. 이 공통 DB 란 말 그대로 여러 다른 기관이 음성기술의 연구나 제품 개발 등의 목적을 위해 사용할 수 있도록 구축되는 DB 를 말하므로 규모 면에서도 충분한 양이 수집되어야 할 뿐 아니라, 여러 목적에 사용될 수 있도록 내용면에서도 보편성과 동시에 다양성과 기능성을 갖도록 수집되어야 한다. 따라서 발성목록의 작성에 있어서도 이런 점들이 고려되어야 한다고 본다.

음소의 분포와 같은 기본적인 사항 외에도, 문장 발성목록 같은 경우에는 대상 코퍼스의 선정과 코퍼스의 정제와 검토가 중요하며, 단어 발성목록 같은 경우에는 영역 설정, 데이터의 선택, 단어 사용 빈도 등이 고려되어야 한다고 본다. 또한 숫자음의 경우에는 우리말 화자의 모든 가능한 숫자음의 발성 양상 등이 고려되어 작성되어야 한다.

또한 음성 DB 가 인식용이냐, 합성용이냐에 따라 발성의 난이도 등이 고려되어 작성되어야 할 것이다.

본 논문에서는 음성인식용 DB, 낭독체 음성합성용 DB, 대화체 음성합성용 DB, 화자인식용 DB 의 순서로 발성목록의 내용과 목적, 작성 방법, 작성시에 고려된 조건 등을 기술한다.

2. 음성인식용 DB 발성목록

음성인식용 DB 를 위한 발성목록은 기본적으로 인식기 개발에 필수적인 단어, 숫자, 문장으로 구성되어 있고, 여기에 받아쓰기(dictation)와 같은 응용 프로그램 개발에 필요한 명령어 세트나 단어의 철자 읽기(spelling), 인식기로 하여금 방언 화자를 구별할 수 있도록 하기 위한 방언화자 구별 문장 등을 추가로 작성하였다. 다음에 구체적인 내용을 살펴본다.

2.1. 녹음 시작 질문

녹음 시작 질문은 화자의 신상 정보를 얻기 위한 것으로 화자의 이름, 나이, 출신지, 주민등록번호 등을 묻는 질문 8 개로 이루어져 있다.

2.2. 단어

단어 목록은 총 10000 단어인데, 단어의 영역별 구성을 보면, 상장회사명과 지명이 20%, 인명이 20%, 상호명과 제품명, 일반명사가 60%를 차지한다. 단어 선정에 있어서는 음소의 분포나 사용빈도가 고려되었으며, 그 내용으로 상장 회사명이나 제품명, 인명 등을 포함시킨 것은 응용 분야에서의 활용성을 고려한 것이다.

2.2.1. 상장회사명

여기에는 국내의 상장 회사명이 모두 포함된다.

2.2.2. 지명

지명에는 도시명, 국가명, 산이름, 유명 관광지 이름 등이 포함된다.

2.2.3. 인명

인명으로는 한국인의 이름 2000 개를 작성하였는데, 이것은 한국인 인명 데이터를 바탕으로, 성은 한국인의 성이 모두 포함되도록 하였으며, 이름은 빈도순으로 2000 개를 뽑아 성과 이름을 조합하여 작성하였다.

2.2.4. 상호명

상호명은 국내의 상호명 목록을 바탕으로 작성하였다.

여기에서 1 차로 발성하기 어려운 단어를 제거하였는데, 단어의 길이는 2 음절에서 8 음절 이내로 제한을 두었고, 일반화되지 않은 로마자 상호라든가, 지나치게 읽기 어려운 단어는 제거하였다. 이렇게 1 차로 정제된 목록을 가지고 2 차로는 트라이폰(triphone)을 단위로 음소 분포를 고려하여 상호명 3000 개를 추출하였다.

2.2.5. 일반명사와 상품명

일반 명사는 사용빈도를 고려하여 고빈도 명사를 뽑고, 여기에 인터넷 쇼핑의 웹사이트에 나와 있는 상품명 가운데 비교적 일반화된 명사를 합쳐 3000 개의 명사를 추출하였다.

2.3. 숫자

숫자음은 크게 번호독식과 봉독식으로 나뉜다. 번호독식이란 일련의 숫자를 하나 하나의 숫자(digit)로 떼어서 읽는 방식이고, 봉독식이란 일련의 숫자를 하나의 수(number)로 읽는 방식이다. 번호독식과 봉독식을 합쳐 총 10420 개의 숫자 목록을 작성하였다 발성목록에는 발성의 편의를 고려하여 아라비아 숫자 뒤에 한글을 괄호 안에 병기하였다.

2.3.1. 번호독식(connected digits)

2.3.1.1. 1, 3, 4 자리 숫자

여기에는 1 자리 숫자 20 개(0~9 의 한자, 우리말 숫자)를 비롯하여, 3 자리 숫자 200 개, 4 자리 숫자 1000 개를 작성하였는데, 3, 4 자리 숫자에서는 숫자와 숫자가 이어지면서 나타낼 수 있는 모든 음운 환경을 커버하도록 작성하였다.

2.3.1.2. 7, 10 자리 숫자

7 자리 숫자와 10 자리 숫자는 각각 1000 개씩 작성하였는데, 이 때도 숫자열에서의 음운환경을 고려하여 작성하였다.

2.3.1.3. 주민등록번호 6 자리

주민등록번호의 앞 6 자리 숫자 1000 개를 추출하였는데, 숫자의 분포를 고르게 하면서 각 자리에는 가능한 숫자(연월일)만이 오도록 작성하였다.

2.3.1.4. 전화번호, 은행계좌번호

전화번호와 은행계좌번호 형식의 숫자는 읽는 방식에 따라 다음의 네 가지로 구분하여 모두 3500 개를 작성하였다. 이 때에도 역시 숫자열의 음운 환경이 고려되었다.

- '-'을 '에'로 읽는 경우
  예: 011-334-2321(공일일에 삼삼사에 이삼이일)
- '-'을 '다시'로 읽는 경우
  예: 134-245486-08736(일삼사 다시 이사오사팔육 다시 공팔칠삼육)
- 한자식과 우리말 숫자의 혼합형
  예: 585-5717(오 팔 오에 오 칠 하나 칠)
- '국' 포함 전화번호
  예: 321 국에 1342(삼백이십일국에 일삼사이)

위에서 한자식과 우리말 숫자의 혼합형이란 전화번호나 주민등록번호 등을 읽을 때 혼동을 피하기 위해 '1'과 '2'를 '하나'와 '둘'로 발성하는 경우를 말한다. 이 때 모든 자리의 1 과 2 가 우리말 숫자로 발성되는 것이 아니라는 점을 고려할 필요가 있다. 이를테면 '185-5717'에서 처음 '1'은 '하나'보다는 '일'로 발성하는 경우가 일반적이므로, 이런 경우는 그대로 한자식으로 읽도록 작성하였다.

2.3.2. 봉독식(natural number)

2.3.2.1. 한자식 자연수

한자식 자연수로는 0 과 99,999 사이의 자연수에서 랜덤(random)하게 생성된 2000 개 숫자를 추출하였다.

2.3.2.2. 우리말 숫자

1 에서 110 사이의 우리말 숫자 200 개가 작성되었는데, 여기에는 수관형사도 포함되었으며, 수관형사의 경우는 뒤에 올 수 있는 단위 명사도 삽입하여 화자가 자연스럽게 읽을 수 있도록 하였다(예: 두 살, 세 개, 네 명).

2.3.2.3. 십만 이상의 조, 억 단위 포함 자연수

십만 이상의 자연수로 조, 억 단위를 포함하는 숫자 500 개를 작성하였다.

2.4. 낭독체 문장

낭독체 문장으로는 방송 뉴스 코퍼스를 대상으로 50,000 문장을 추출하였다. 내용적으로는 정치, 경제, 사회, 문화 영역으로 제한하였는데, 그것은 스포츠 같은 영역은 외래어가 많이 포함되어 있어 일반 화자들이 발성하기에 용이하지 않기 때문이다.

코퍼스를 이용하여 문장을 추출할 경우에는 코퍼스의 정제가 중요하다. 코퍼스의 정제 기준은 다음과 같았다.

- 발성시 혼동을 일으킬 수 있는 요소인 영어, 한자, 아라비아 숫자 등은 한글로 변환시킨다.
- 쉼표, 마침표, 물음표가 아닌 문장부호가 들어간 문장 역시 발성시에 혼동을 줄 수 있으므로 제외한다.
- 비문법적 문장, 문장이 아닌 구, 낭독체 문장으로 볼 수 없는 문장은 제거한다.
- 화자의 발성 시에 심리적인 부담감을 줄 수 있는 비속어 표현이 들어있는 문장은 제거한다.
- 화자가 실수없이 발성할 수 있도록 문장 길이는 어절 수 4~8 사이로 제한하되 최소 9 음절 이상이 포함되도록 한다.

발성목록은 이러한 코퍼스의 정제를 거쳐 5 만 문장을 추출하였다.

2.5. 명령어 세트

응용 프로그램 개발에 활용될 수 있도록 PC 와 PDA 명령어 및 받아쓰기(dictation) 프로그램 명령어 358 개를 수집하여 명령어 세트를 작성하였다.

2.6. 단어 철자로 읽기(spelling)

단어를 철자로 읽은 음성데이타는 받아쓰기(dictation) 프로그램 등에서 발음이 혼동되기 쉬운 한국어 단어를 철자로 읽는 경우나 영어 단어를 철자로 읽는 경우가 있으므로 이를 위해 활용될 수 있다.

한국어 단어로는 발성만으로는 철자화가 혼동되기 쉬

운 2~4 음절의 단어(예: 은혜, 은해) 2500 개를 추출하여 자소로 분리하여 읽도록 목록을 작성하였다.

예: 이응 으 니은 히읗 여 이

또한 영어 단어도 철자로 읽기 목록 2500 개를 작성하였다.

2.7. 방언사용자 구별 문장

이번에 수집되는 음성 DB 는 각 방언 사용자들의 음성을 일정한 비율로 포함하도록 되어 있다. 방언사용자 구별 문장의 목적은 인식기로 하여금 이러한 방언 사용자의 발음을 구분해낼 수 있도록 하기 위한 것이다.

이를 위해서 각 방언의 발음 특성을 드러낼 수 있는 단어들을 포함하는 문장을 10 개씩 2 세트 작성하였다. 각 문장 세트 안에는 여러 방언의 발음 특성을 나타내는 음소들이 고르게 분포되도록 하였다. 다음의 문장들이 그 예이고, 밑줄 친 부분은 방언에 따라 다르게 발음되는 음소를 포함하는 음절들이다.

- 언어는 의사를 전달하기 위한 수단이다.
- 이번 경기에서 한국 팀이 우승을 했다고 한다.
- 오늘 슈퍼에서 사온 쌀이 맛이 없다.

3. 정보전달용 낭독체 음성합성 DB 발성목록

정보전달용 낭독체 음성합성 DB 는 낭독체 문장, 다이폰 세트, 반음절 세트로 구성된다.

3.1. 낭독체 문장

낭독체 문장은 총 10,000 문장을 추출하였는데, 코퍼스로는 방송뉴스코퍼스와 신문 코퍼스를 사용하였다. 이 역시 인식용 DB 의 문장에서와 같이 1 차로 일정한 정제과정을 거친 코퍼스를 사용하였다. 숫자나 기호는 한글로 바꾸어 발성 시에 혼동을 주지 않도록 하였고, 문장이 아닌 구나 비문법적인 문장, 대화체적인 문장은 제거를 하였다.

다만, 합성용 문장은 인식용 문장과는 달리 훈련된 화자가 발성하는 것이므로 문장 길이의 제한을 완화하여 4~12 어절 사이의 문장을 추출하였다.

2 차로는 트라이폰(triphone)을 단위로 하여 한 문장 내에 트라이폰을 최대한 많이 포함하고, 최소의 문장 개수에 최대의 트라이폰을 커버하는 문장 세트를 추출하였다.

기본적으로는 20 여만 문장의 방송뉴스 코퍼스를 이용하였는데, 여기서 누락된 트라이폰(missing triphone)들은 신문코퍼스를 이용하여 추가하였다. 문장세트는 단문과 장문으로 구분하여 추출하였다. 문장세트의 추출 과정은 다음과 같다.

- 방송뉴스 단문 코퍼스(4-8 어절)에서 트라이폰 13709 개를 포함하는 문장 5270 개를 추출.
- 방송뉴스 장문 코퍼스(9-12 어절)에서 트라이폰 14658 개를 모두 포함하는 문장 4301 개를 추출.
- 신문 코퍼스(4-10 어절)에서 위의 단문, 장문 코퍼스에 나타나지 않은 트라이폰 429 개를 포함하는 문장 391 개를 추출.

3.2. 휴지 앞(pre-pausal) 세그멘트를 위한 발성목록

대용량 음성 DB 를 기반으로 하는 합성 시스템은 휴지 앞 세그멘트(segment)를 그대로 가져다 쓰는데, 대체로 긴 휴지가 나타날 때는 그 앞의 세그멘트는 피치(pitch)가 내려가면서 모음의 길이가 길어지는 것이 자연스러우므로, 이렇게 휴지의 구현에 필요한 휴지 앞 세그멘트가 DB 에 충분하지 않으면 자연스런 합성음을 생성할 수 없다.

그런데 아무리 많은 문장을 녹음한다 해도 가능한 휴지 앞 세그멘트를 모두 갖추기는 불가능하다. 따라서 휴지 앞 세그멘트를 위한 데이터를 따로 만들 필요가 있다.

휴지 앞 세그멘트 데이터를 위한 발성목록은 두 가지 종류로 모두 555 문장을 작성하였다.

하나는 문장의 낭독시에 휴지가 발생할 수 있는 부분, 이를테면 연결어미 같은 것이 최대한 많이 포함된 문장 세트이다.

두 번째 목록으로는 명사 등을 열거할 때 생길 수 있는 휴지 앞 세그멘트를 확보하기 위해, 명사를 나열하고 명사 사이에 쉼표를 붙여, 발성 시에 휴지를 유도하는 문장을 작성하였다. 이 때 명사의 목록은 일반명사와 고유명사 데이터를 바탕으로 하여, 단어의 마지막에 올 수 있는 트라이폰이 최대한 포함되도록 단어 목록을 추출하여, 이들을 5 개씩 배열하고 단어 사이에 쉼표를 붙여 의미 없는 문장을 만들었다. 다음은 그 예이다.

예: 보기 일 번은 함유, 보기 이 번은 아내, 보기 삼 번은 자녀, 보기 사 번은 단위, 보기 오 번은 한우, 이상입니다.

3.3. 다이폰 세트 추출용 발성목록

다이폰을 단위로 하는 합성기 개발을 위해 다이폰 세트 추출용 발성목록을 작성하였다. 휴지(pause)와 모음 21 개, 초성 자음 18 개, 종성 자음 7 개의 가능한 다이폰 세트를 포함하는 무의미어의 단어목록 2037 개 작성하였다.

3.4. 반음절 세트 추출용 발성목록

반음절을 단위로 하는 합성기 개발을 위해 CV, VC, V 의 반음절 세트와 그 가능한 음운환경을 포함하는 무의미 단어 목록 5271 개를 작성하였다. 이때 고려된 반음절과 음운환경은 다음과 같다.

- C_CV_pause
- pause_CV_C
- pause_VC_C
- pause_V_C
- V_CV_V
- V_V
- V_VC_pause

4. 대화체 음성합성 DB 발성목록

대화체 음성합성 DB 의 발성목록 작성은 TV 대담

프로를 전사한 코퍼스를 이용하였다.[1]

총 79577 문장 가운데 4~15 어절의 문장을 추출하고, 이를 바탕으로 트라이폰을 단위로 음소분포를 고려한 문장 3139 개를 추출하였다. 구어 전사 코퍼스에는 더듬거림이나 간투사, 단어의 반복, 비표준어 등이 포함되어 있으므로 음소의 분포와 배열을 최대한 살리는 범위에서 문장 수정을 하였다.

5. 화자인식용 DB 발성목록

화자인식용 음성 DB 는 2 연 숫자 100 개와 4 연 숫자 1000 개, 단문 수집용 질문목록으로 구성된다.

5.1. 2 연 숫자

0~9 사이의 숫자로 음운 환경을 고려하여 2 연 숫자 100 개를 추출하였다.

5.2. 4 연 숫자

0~9 사이의 숫자로 음운 환경을 고려하여 4 연 숫자 1000 개를 추출하였다.

5.3. 단문 수집용 질문목록

화자로부터 단어나 짧은 문장의 답을 유도해내기 위해 개인정보와 관련된 10 개의 질문을 작성하였다.

6. 결론

본 논문에서는 공통 음성 DB 구축을 위한 발성목록의 내용과 작성 방법, 목록 작성시 고려된 사항들을 기술하였다. 이번에 작성된 발성목록은 음성 기술의 연구나 프로그램 개발을 위해 기본이 되는 음성 DB 를 수집하기 위한 목록이라고 할 수 있다.

DB 구축을 위한 발성목록을 작성할 때 문제가 될 수 있는 것은 활용할 수 있는 텍스트 데이터의 양과 질이다. 단어나 문장 목록 작성시에는 최대한 많은 수의 인식, 또는 합성 단위를 포함할 수 있으려면 충분한 양의 텍스트 데이터가 확보되어야 하며, 녹음 시에 문제가 없도록 정제 작업이 선행되어야 할 것이다. 현재, 낭독체 문장의 경우에는 활용할 수 있는 코퍼스가 많아 데이터 수집의 어려움이 덜한 편이나, 대화체 문장의 경우에는 국내에서 이용할 수 있는 구어의 전사 자료 양이 매우 제한되어 있고, 또 이를 정제하기 위해서는 낭독체 문장보다 더 많은 노력이 필요하다. 이러한 점을 고려하여 DB 구축 계획 시에 텍스트 데이터에 대한 대책이 마련되어야 할 것이다.

[1] TV 대담 전사 코퍼스는 21 세기 세종 말뭉치를 이용하였다.