국어공학센터의 한국어 음성DB 구축계획

○
이용주, 김봉완, 김종진, 양옥렬, 임선영
*원광대학교 컴퓨터공학과

Construction of Korean Speech DB at KLE

Yong-Ju Lee, Bong-Wan Kim, Jong-Jin Kim, Ok-Yul Yang, Seon-Young Lim*
*Dept. of Computer Engineering, Wonkwang Univ.

요 약

본 논문에서는 국어공학센터에서 국어정보베이스 구축의 일환으로 추진되고 있는 한국어 음성DB에 대하여 구축현황 및 향후 계획을 소개한다.

1. 서 론

지금까지 한국어의 음성데이타베이스는 각 연구자가 필요에 따라 음성데이타를 만들어 보관하고 이용해 왔다. 음성연구가 진보되어감에 따라 처리가능한 데이타수는 많아져 가고, 따라서 준비해야할 데이타량도 대폭적으로 증가되었다. 최근에는 음성인식의 경우, HMM(Hidden Markov Model)이나 Bigram/ Trigram등 언어모델로 대표되는 통계적수법의 발달에 따라 대량의 음성데이타가 시스템의 학습에 필요하게 되었다. 또한 음성정보처리시스템의 연구개발을 위해서는 분석, 합성, 인식의 각종 알고리즘을 적절하게 비교 평가할 필요가 있지만 이를 위한 방법으로는 현재까지는 공통음성데이타를 이용하여 알고리즘을 수행하고 그 결과를 비교하는 방법 이외에는 알려져 있지 않다. 따라서 공통이용가능한 각종 대량 음성 데이타를 수록, 보관, 공개하는 것은 연구 개발과정에서의 이용 및 인식장치의 성능평가 양면에서 필요하다. 이러한 목적으로 이용하는 음성데이타를 일반적으로 음성데이타베이스, 음성코퍼스 또는 음성사전이라고도 부른다. 합성의 경우에도 지금까지 다이폰, 반음절 등 각종단위에 의한 접속방식이 주류를 이루고 있고 최근에는 대형의 음성DB로 부터 임의 길이의 음성부분을 골라내어 접속하므로써 좋은 합성품질을 얻고 있다. 이를 위해서는 잘 정비된 대형의 음성DB가 필요하다. 또한 인식 및 합성 알고리즘의 개발을 위해서는 다양한 환경의 음성언어학적 분석이 필연적으로 요구되는데 이를 위해서도 음성DB 개발 초기에 시급히 확보 되어야 한다.[1][2][3]

본고에서는 국어정보베이스 구축의 일환으로 연구하고 있는 한국어 음성DB를 위해 필요한 음성 데이타의 녹음, 저장 및 국어정보베이스 통합 시스템과의 인터페이스 등에 관한 구축계획 및 현황에 대하여 기술한다. [4]

2. 국어공학센터의 국어정보베이스

최근에 들어 국내에서도 한국어의 텍스트, 음성, 문자, 사전 등의 통합적인 데이타 베이스 구축을 위한 체계적이고 지속적인 계획을 마련하여, 각 요소기술간의 기술 연계와 대량의 데이타의 공동이용 및 객관적인 평가와 검증을 위하여 과기처와 문화체육부가 지원하는 국어공학센터가 설립되었다. 이를 통해 산발적이고 소규모의 개별적 연구체제를 지양하고 계획적인 산,학,연 협동연구를 지향하고 있다. 국어공학센터에서는 한글정보처리 표준화기술, 국어정보 베이스 구축, 지능형 처리기 개발 등에 대해서 연구하고 있다.

국어정보베이스의 역할은 이러한 기존의 텍스트, 사전, 음성, 문자데이타와 사용과 처리상의 문제점을 해결하기 위해 각 요소기술들을 통합하여 확보하고 통합 개발환경 지원과 관리 시스템 개발하여 연구와 개발을 담당하는 주 대상자들과 자원을 공유하고 여기서 만들어진 요소기술들을 통해 한국어의 말과 글의 포괄적인 응용시스템 개발에 필요한 요소기술들을 제공하는 역할을 하려 하고 있다. 이 중 음성DB는 한국어 정보베이스의 일환으로 연구가 수행되고 있다.

3. 공통음성DB 확보를 위한 구축계획[1]

한국어 정보처리기술 개발[1]이 제 1단계('94~'96)에서는 어절기반의 국어 정보 처리 기술 환경 정비 및 개발, 제 2단계('97~'99)에서는 구문기반의 국어 정보 처리 기술 개발, 제 3단계('2000~'3)에서는 지식기반의 우리말 컴퓨터 시제품 개발을 계획하고 있는데 발맞추어 음성DB도 이와 상응하는 장기계획하에 지속적이고 체계적인 구축을 계획하고 있다.

제 1단계의 구축계획은 한국어 음성정보처리를 위한 음성DB의 체계적이고 지속적인 구축과 공동이용 및 유지관리, 보급체계 확보에 그 중점을 두는 공통음성DB 구축을 위한 기술 환경 정비 및 환경 개발을 그 목표로 하고 있다.

1차년도에서는 공통음성DB의 연차별 확보를 위한 장기적인 확보계획을 작성하면서, 개발환경의 구축을 위해 우선 공통적인 대상인 PBW, 단독숫자, 4연숫자, 이야기문 등의 음성분석용 기본세트 음성DB를 시험적으로 구축한다. 또한 수요자의 요구분석을 통해 대상목록 추가와 우선순위를 조정할 예정이다.

2차년도에서는 단어레벨 음성DB를 확장하기 위한 레이블링 작업과 함께 발성자의 숫자를 100명으로 확대하고 이에 따른 단어음성 CDROM 형태의 시제품을 개발하고 문장레벨음성DB 설계를 위한 문장음성 발성목록을 작성할 것이다.

3차년도에는PBS(Phonetically Balanced Sentences)의 문장음성 대상으로 100명분의 데이타를 CDROM형태로 구축할 예정이다.

목록의 선정은 본 1차년도 연구결과인 장기 구축계획에 의해 세부적으로 확정할 것이나 1차년도 연구기간 중에는 대상단어를 확정하지 않는 단계이므로 대상단어 중 기본이 되는 것을 우선으로 하며, 단독숫자, 4연숫자, PBW, 이야기문등을 1차년도의 발성 대상으로 한다. 이 데이타는 국어의 음운현상의 연구, 음성 합성을 위한 데이타 파일 작성 등에 유용하다. 2차년도 이후에는 장기구축 계획의 검토결과에 따라 단어, 문장순으로 확대한다. PBW의 발성목록은 약 120만어절 규모의 텍스트 코퍼스로부터 고빈도 5000어절을 추출하고 이 중에서 다양한 한국어 음소환경이 포함되도록 선정하였다. [5]

4. 구축현황

지금까지 선정된 발성목록을 기준으로 확정된 PBW, 단독숫자, 4연숫자, 이야기문을 이용해 만들어진 음성DB의 구축현황은 다음과 같다.

4.1 PBW(PhoneticallyBalancedWords) : 452종

[그림-1] PBW Samples

4.2 단독숫자음 : 41종

영, 공, 일, 이, 삼, 사, 오, 육, 륙, 칠, 팔, 구, 십, 백, 천, 만, 억, 조, 경, 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟(여덜), 아홉, 열, 스물, 서른, 마흔, 쉰, 예순, 일흔, 여든, 아흔, 다시, 예, 네, 아니오

4.3 4연숫자음 : 35종

0287, 5732, 9601, 4156, 1199, 1398, 6843, 0712, 5267, 6633, 2409, 7954, 1823, 6378, 8877, 3510, 8065, 2934, 7489, 2244, 4621, 9176, 3045, 8590, 5500, 6872, 5861, 3649, 0316, 7083, 8194, 9205, 1427, 2538, 4750

4.4 이야기문 : 1종

(국제음성학회 한국어 발음 예문 "바람과 햇님"중에서)

바람과 햇님이 서로 힘이 더 세다고 다투고 있을 때, 한 나그네가 따뜻한 외투를 입고 길어 왔습니다. 그들은 누구든지 나그네의 외투를 먼저 벗기는 이가 힘이 더 세다고 하기로 결정했습니다.

4.5 발성자 선정

표준어를 사용하는 20 ~ 50세의 일반인(또는 아마추어 아나운서) 남녀 각 2인이 2회씩 발성하였다. 2차년도의 발성자 확장시 지역별,연령별로 고른 분포가 되도록 할 예정이다.

4.6 음성데이타 녹음

4.6.1 녹음장소 : 방음부스를 사용하였으며 암소음 레벨은 20dB 이하로 하였다.

4.6.2 입력장치 : Senheizer HMD 224X 마이크를 사용하였으며, 마이크는 입에서 왼쪽으로 대각선으로 약 10cm의 거리에 두고 녹음하였다.

4.6.3 녹음 : 디지탈 오디오 테이프(DAT)를 이용하여 녹음하였다.

4.6.4 A/D과정 : 16KHz로 샘플링하고, 16Bit로 양자화 하였다.

4.7 음성 데이타의 편집

모든 음성 데이타는 A/D하여 컴퓨터의 하드 디스크에 저장한 후 단어 또는 문장단위로 편집하였으며, 이때 음성 데이타의 시작점과 끝점으로부터 300ms의 무음구간을 두고 편집한다. 편집은 끝점 검출 알고리즘에 의해 단어를 자동 분할하고 컴퓨터 화면상에서 음성 파형을 디스플레이 하여 수정한다.

4.8 음성 데이타의 저장

편집된 음성데이타 파일을 저장하는 방법은 편집된 데이타를 하나의 큰 데이타로 묶어서 필요한 파일만을 찾아낼 수 있는 Dictionary 파일의 형태로 하면 데이타 관리는 용이하나 개개의 파일을 복사할 때 번거로움이 따른다. 반대로 음성파일을 개개의 하나씩 이름을 붙여 사용하면 파일을 관리하기는 어렵지만 필요한 파일을 복사 또는 이용하기가 용이하다. 본 연구에서는 각각의 데이타를 계층적 디렉토리 구조로 저장하였으므로 CDROM화에도 용이하며 효율적으로 데이타를 검색할 수 있도록 하였고 그림-2는 디렉토리 구조를 나타낸 것이다.[6]

/<CORPUS>/<USAGE>/<DILECT>/<SEX>/<SPEAKER_ID>/<TEXT_ID>.<FILE_TYPE>

- CORPUS :== kbessdb
- USAGE :== train | test
- DIALECT :== 특별시, 직할시, 각 도 등의 코드

이 름	코 드
서울	1
부산	2
대구	3
인천	4
광주	5

- SEX :== m | f
- SPEAKER_ID :== <INITIALS><DIGIT>
  - INITIALS :== 화자 이니셜, 3문자
  - DIGIT :== 0 ~ 9, 동일이니셜일 경우의 구분
- TEXT_ID :==<TEXT-TYPE><TEXT_NUMBER>
  - TEXT_TYPE :== pbw | pbs
  - TEXT_NUMBER :==1, 2, 3, ...//지표번호
- FILE_TYPE :== wav | txt | wrd | phn

파 일 타 입	기 술
.wav	음성 waveform파일
.txt	발성단어의 orthographic
.wrd	transaction
.phn	Time alligned word transaction Time alligned phonetic transaction

[그림-2] Speech DB Directory Structure

4.9 관리 및 배포 방법

4.9.1 음성 데이타의 CDROM화

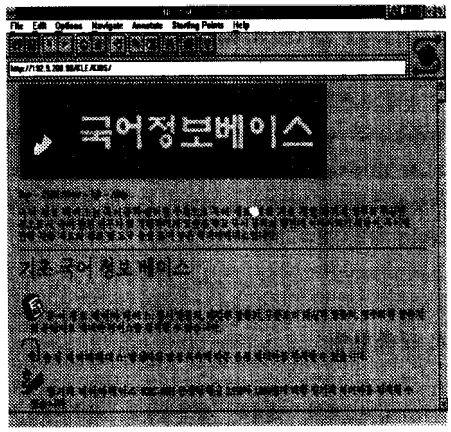
일반적으로 음성데이타는 그 양이 매우 많으므로 대용량의 음성 데이타를 저장하기 위해서는 CDROM을 사용하는 것이 일반적이다. 또한 시판되고 있는 CDROM 드라이버는 비교적 저가이고 컴퓨터에 부착하는데도 어려움이 없다. 따라서 본 연구결과는 2차년도에 양적으로 대폭 확대하며 음성 데이타 베이스를 ISO 9660표준 포맷에 의거하여 CDROM으로 제작할 수 있도록 할 예정이다.

4.9.2 음성 데이타의 확장

음성 데이타는 초기에 단어레벨 음성DB를 기준으로 하여 확장할 예정이며, 이후 문장레벨의 음성DB를 구축, 점차적으로는 자유발화음성(Spontaneous Speech)으로 확대하도록 계획하고 있다.

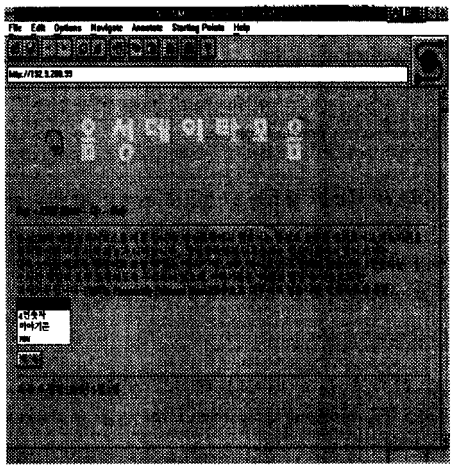
5. 국어 정보 베이스 통합 시스템과의 인터페이스

향후 구축된 음성DB는 통합된 국어 정보베이스로 Network상에서 검색이 가능하도록 할 예정이며, 이를 위한 인터페이스 방법으로 인터넷(InterNet) 상에서 모자익(Mosaic)을 사용하고 있다. 국어공학센터에서 인터넷상의 모자익 서버(Mosaic Server)를 운영하고 모자익 브라우지(Mosaic Browser)를 통해 음성DB 자료를 네트워크를 이용하여 검색이 가능하도록 하였다. 그림-3은 모자익 서버를 상에서 HTML(Hyper Text Markup Language)를 이용하여 만든 국어 정보 베이스에 관한 Home Page이다.



[그림-3] 국어정보베이스 Mosaic Home Page

Browser가 음성DB 파일에 대한 Viewer로서 제공되어야 할 기능으로 음성DB 파일에 대한 음성 입출력기능과 파형 및 스펙트로그램을 보여주는 기능, 그리고 음성DB 파일에 기록된 음소단위 세그먼테이션, 레이블링등의 정보를 나타내는 기능이 포함될 예정이다. 그림-4는 모자익에서의 음성DB에 관한 홈 페이지이다.



[그림-4] Speech DB Home Page

6. 결 론

지금까지 국어공학센터에서의 공통음성DB 구축 현황 소개 및 계획을 하였다. 본 사업이 성공적으로 수행되어 음성연구자들이 조기에 공동으로 사용 가능한 음성DB가 구축되도록 노력할 것이며, 관련연구자 여러분들의 조언을 부탁드린다.

본 연구는 과기처의 연구비 지원으로 이루어진 것으로, 연구수행을 지원한 국어공학센터 박동인 센터장, 국어정보 베이스팀의 최거선 박사께 감사드린다.

참 고 문 헌

[1] 이용주, "한국어 음성언어정보처리와 음성 데이타베이스", 한국어정보처리 소식 제2권 특별기고, 1994.10

[2] 이용주, 정유현 외, "보급형 음성 데이타베이스 구축에 관한 연구", 한국전자통신연구소 최종보고서, 1992.7

[3] 정유현, 최준혁 외, "공통 음성 데이타베이스 구축을 위한 사전 조사 연구", 전자공학회 하계 학술 대회, 1992.6.

[4] 이용주, 김봉완 외, "국어정보베이스-한국어 음성DB의 구축에 관한연구", 한국과학기술원 중간보고서, 1995.4

[5] 이용주, 김봉완 외, "음성DB용 PBW에 관한 검토",한국음향학회 제12회 음성통신 및 신호처리워크샵 논문집, 1995.6

[6] NIST: Acoustic-Phonetic Continous Speech Coppus CD-ROM by The National Institute of Standards and Technology(NIST), Feb, 1990