인식 및 합성용 음성 코퍼스의 발성 목록 설계

김형주*, 김봉완**, 이용주*
*원광대학교 컴퓨터공학과
**원광대 음성정보기술산업지원센터

Design of the Linguistic Contents of Speech Corpus for Speech Recognition and Synthesis

Hyoung-Ju Kim*, Bong-Wan Kim, Yong-Ju Lee***
* Dept. of Computer Engineering, Wonkwang Univ.
** Speech Information Technology & Industry Promotion Center, Wonkwang Univ.

E-mail : snowbomb@korea.com, bwkim@sitec.or.kr, yjlee@wonkwang.ac.kr

요 약

최근 컴퓨터와 인간간의 대화 수단으로 음성을 활용하는 기술인 음성정보기술이 발달함에 따라 대어휘 연속 음성 인식 및 무제한 어휘 음성 합성의 고도화를 위한 연구가 진행되고 있다. 음성 인식의 경우 HMM으로 대표되는 통계적 수법의 발달에 따라 시스템의 학습을 위해 대량의 음성데이터가 필요하며, 음성 합성의 경우에도 최근 대형의 음성 데이터 베이스로부터 임의 길이의 음성 부분을 골라내어 접속함으로써 좋은 합성 품질을 얻고 있다. 본 논문에서는 이러한 음성 인식 및 합성을 위해 공동으로 사용하기 위한 음성 데이터베이스의 발성 목록을 설계하고 설계된 결과에 대하여 논의한다.

1. 서론

한국어의 공학적인 응용을 위해서는 그 기반이 되는 요소기술로써 음성인식 및 합성으로 대표되는 음성 처리기술과 언어 이해 및 기계번역으로 대표되는 언어 처리 기술의 연구가 필요하다. 이러한 음성 및 언어 처리기술의 연구를 위해 가장 먼저 확보되어야 할 것이 음성, 언어 및 각종 사전 DB 등 국어 정보베이스이다. 이들의 체계적인 조기확보 여하에 따라 음성 및 언어처리연구의 성패를 좌우한다고 해도 과언이 아니다. 특히 한국어 음성을 대상으로 한 음성 DB는 음성 언어 연구의 기본으로서 개발초기부터 확보되어야 할 연구자원이다.

음성을 맨 머신 인터페이스의 수단으로 활용하기 위한 음성 정보처리 연구는 관련 기술의 진보에 따라 한정 어휘의 인식 및 합성 시스템들의 실용화에 접어들고 있고, 이제 임의 어휘를 대상으로 하는 음소 단위 인식기술 개발이 필수적이지만, 연속 음성중의 음소는 발성자에 따른 개인차는 물론이고 전후에 발성되는 음소의 영향에 의한 조음 결합에 따라 그 특성이 크게 변화한다. 이러한 개인차 및 조음 결합의 현상을 분석하기 위해서는 많은 사람이 발성한 다양한 음성 데이터가 필요하다. 또한 시스템의 객관적인 평가를 위해서 표준적인 음성 데이터도 필요하다. 이 음성 데이터는 다종 다양(성별, 연령, 발성지수, 방언)한 것이 필요한데, 지금까지는 각 연구자가 필요에 따라 음성 데이터를 만들어 보관하고 이용해 왔다.

음성연구가 진보되어감에 따라 처리 가능한 데이터 수는 많아져 가고, 따라서 준비해야 할 데이터 량도 대폭적으로 증가되었다. 최근에는 음성인식의 경우, HMM을 이용한 음향 모델, bigram/ trigram등의 언어모델 등 통계적 수법의 발달에 따라 대량의 음성 데이터가 시스템의 학습에 필요하게 되었다.

한편 음성 정보처리 시스템의 연구개발을 위해서는 분석, 합성, 인식의 각종 알고리즘을 적절하게 비교

평가할 필요가 있지만 이를 위한 방법으로 현재까지는 공통 음성 데이터를 이용하여 알고리즘을 수행하고 그 결과를 비교하는 방법이외에는 알려져 있지 않다. 따라서 공동으로 이용 가능한 각종 대량의 음성 데이터를 수록, 보관, 공개하는 것은 연구개발 과정에서의 이용 및 인식장치의 성능평가 양면에서 필요하다. 이러한 목적으로 이용하는 음성 데이터를 일반적으로 음성 데이터베이스, 음성 코퍼스 또는 음성 사전이라고 부른다. 합성의 경우에도 지금까지 다이폰, 반음절 등 각종 단위에 의한 접속 방식이 주류를 이루고 있고 최근에는 대형의 음성 데이터 베이스로부터 임의 길이의 음성 부분을 골라내어 접속함으로써 좋은 합성 품질을 얻고 있다. 이를 위해서는 잘 정비된 대형의 음성 데이터베이스가 필요하다. 또한 인식 및 합성 알고리즘의 개발을 위해서는 다양한 환경의 음성 언어학적 분석이 필연적으로 요구되는데 이를 위해서도 음성 데이터베이스의 중요성은 크다[1].

따라서 음성 및 언어의 연구를 위해 데이터가 중요하다는 것은 연구자간에 이론이 없으나 공동으로 사용하기 위한 대규모의 데이터베이스를 갖추는 일은 간단한 문제가 아니다. 본질적으로 데이터는 연구 그 자체의 일부이며 어떤 데이터를 어떻게 모아 가공하는가는 연구 내용에 크게 의존한다. 따라서 다른 연구 목적간에 두루 쓰일 수 있도록 노력하여야하나 완벽하게 범용일 수는 없다. 그러나 대량의 데이터가 공통의 포맷으로 제공되는 것만으로도 그 의의는 매우 크다.

본 논문에서는 이러한 공동 이용을 위한 음성 데이터베이스를 구축하기 위하여 대어휘 연속 음성 인식을 위한 Dictation용 음성 DB의 발성 목록과 무제한 음성 합성용 DB의 발성 목록을 설계하고 그 결과를 논의한다.

2장에서는 대어휘 연속 음성 인식을 위한 Dictation용 음성 DB의 설계에 대하여 기술하며 3장에서는 무제한 어휘 합성용 음성 DB의 설계에 대하여 기술한다. 마지막으로 4장에서 결론을 기술한다.

2. Dictation용 낭독 음성 DB를 위한 발성 목록 설계

가. Dictation용 음성 DB의 사례 연구

최근의 음성인식연구의 방향을 크게 두 가지로 나눈다면 하나는 자연스러운 대화를 인식하는 대화시스템(dialog system)과 낭독문을 중심으로 대어휘 연속음성을 인식하는 구술시스템(dictation system)의 연구가 될 것이다. 특히 구술시스템은 이른바 받아쓰기 기계 또는 음성타이프라이터와 같은 개념으로 대어휘의 문장을 연속적으로 발성하여 이를 인식해 내는 기술이 근간이 된다.

문장단위로 발성된 음성을 연속적으로 인식하는 "연속음성인식"기술은 또한 음성인식기술의 중심적 과제이다. 특히 수만 단어 이상의 어휘를 다루는 대어휘 연속음성인식은 일상적으로 우리 인간이 발성하는 음성언어의 거의 모두를 처리대상으로 하므로 음성인식기술의 응용 분야 확대에 매우 중요한 과제이기도 하다.

최근, 미국이나 유럽에서는 음성/언어의 데이터 및 모델을 공통화시켜 대어휘 연속음성인식시스템의 성능을 비교하고 이를 통해서 각각의 요소기술을 평가함으로서 시스템 성능을 비약적으로 향상시키고 있다. 최근의 보고에 의하면 2만 단어의 어휘를 갖는 신문낭독문의 인식을 범용 워크스테이션을 이용하여 실시간의 수배 정도로 실행하여 단어 에러율 10%이하의 인식정확도를 달성하는 것이 표준적인 시스템의 성능이다.

미국에서는 1992년경에 DARPA(Defence Advanced Research Project Agency)의 주도 하에 신문기가(Wall Street Journal)를 대상으로 한 대어휘 연속음성인식 연구가 시작되었다. 어휘사이즈를 5K, 20K로 설정하여 기본적인 평가조건을 Hub, 음향모델/ 언어모델의 적응화에 의한 평가 등 부대적인 평가조건을 Spoke 라고 부르며 여러 평가 조건아래에서 활발한 연구가 진행되고 있다. 1992년에는 "5K 어휘/미지어 불포함"의 신문기사 태스크 상에서 16.6%의 단어 에러율이 보고되었다. 그 후 2년 간 에러율이 1/3에서 1/4까지 줄였다. 최근에는 어휘제한이 없는 (20K어휘에 1단어 이상의 미지어가 포함된) 태스크에서 6.6%의 에러율이 보고되고 있다. 이들 결과는 N-gram이라고 하는 단순한 것임에도 언어모델을 도입함으로서 수만 단어의 어휘를 가진 연속음성을 인식할 수 있다는 것을 보인 것이다.

한편, 유럽에서도 신문낭독문의 대어휘 연속음성인식(dictation)을 통한 기술평가가 최근 수년동안 활발히 이루어져왔고 이들 기술개발 및 평가에 있어서 여러 종류의 텍스트코퍼스가 담당한 역할은 크다. 또한 SQALE(Speech recognition Quality Assessment for Linguistic Engineering) 프로젝트에서는 영어(미국식,

영국식), 불어, 독어에 관한 연속음성인식을 평가하였다. 이 프로젝트에서는 언어모델 및 음소모델의 학습 데이터 량을 언어간에 거의 균등하게 하여 상호비교를 쉽게 함과 동시에 연구기관간에도 단어의 음소사전을 공통으로 사용되도록 하였다. SQALE의 결과는 영어 이외의 유럽어에 있어서도 영어와 똑같은 방법에 의해 대규모의 연속음성인식이 가능하다는 것을 보이고 있다.

일본의 경우는 최근 음성인식의 대규모화에 대응하여 일본음향학회에서 연속음성 데이터베이스가 구축되어 음소모델 학습용의 데이터베이스로서 널리 이용되고 있다. 그리고 연속 음성인식용 텍스트 데이터에 관해서는 정비가 이루어지고 있지 않다가 일본 정보처리학회의 음성 언어 연구 연락회가 중심이 되어 "대어휘 연속음성 인식연구를 위한 데이터베이스 정비 워킹그룹"이 발족되어 1995년 11월부터 1997년 10월까지를 목표로 대어휘 텍스트코퍼스, dictation을 위한 음성 데이터베이스 그리고 Dictation을 위한 기본모델 및 개발도구의 구축을 목표로 활발한 활동을 보였다. 특히 이 그룹에서는 마이니치신문 4년분의 기사를 이용하여 선정된 문장을 대상으로 150명 정도의 화자가 총 15000문장 정도의 신문기사 낭독음성을 수록하는 것을 목표로 추진하였으며 낭독 대상문은 대어휘용 (2만단어급), 중어휘용(5000단어급)의 단어세트를 기본으로 문장의 길이나 복잡성을 고려하여 선택하였다. 아울러서 신문기사의 텍스트를 이용하여 bi-gram, tri-gram과 같은 기본적인 언어모델도 정비하고 있다[2].

국내의 경우, 대어휘 연속 음성의 연구 예나 이를 위한 평가방법, 그리고 이를 평가하기 위한 공통의 데이터에 대한 연구가 아직 없다. 특히 텍스트 및 음성 코퍼스는 이러한 기술 개발에 있어서 초기의 훈련과정에서뿐만 아니라 객관적인 성능평가를 위해서 초기부터 확보하여야 할 연구환경이다.

여기에서는 대어휘 연속 음성 인식을 위한 문장음성 DB를 구축하기 위하여 그 대상이 되는 발성목록을 설계한 결과와 설계된 발성목록을 토대로 구축된 Dictation용 음성 DB에 대하여 기술한다.

나. 발성 목록 설계

(1) 발성 목록 선정을 위한 텍스트 전처리

발성 목록 선정을 위한 모집단으로는 KAIST에서 구축된 4,300만 어절의 KAIST Corpus[3]를 사용하였다. 그러나 이러한 텍스트 코퍼스는 그 내용이 문자언어이므로 자연스럽게 낭독하는데는 곤란한 표현이 존재할 수 있다. 따라서 이러한 문장은 발성목록 후보에서는 제외하여야 한다. 발성 목록의 설계를 위해 제외 또는 수정된 사항은 다음과 같다.

- ○ 문장부호의 삭제
  - '<'과 '>'의 삭제
    예) .... 나고 일어난다는 것은 <철학의 빈곤>을 말하고 ...
    수정 후) ... 나고 일어난다는 것은 철학의 빈곤을 말하고 ...
  - '['과 ']' 의 삭제
    예) 물론 그 전에도 롱기누스가가 처음으로 시의 쾌락론을 내놓았다는 [숭엄론]이 있다
    수정 후) 물론 그 전에도 롱기누스가가 처음으로 시의 쾌락론을 내놓았다는 숭엄론이 있다
  - '"', '"', '/', '~', '-', '=' 등 의 삭제
    예) 田成子 = 성이 田, 이름 常. ...
    수정 후) 田成子 성이 田, 이름 常. ...
- ○ 괄호를 포함하고 있는 단어 (예를 들면, 한자어) 혹은 문장의 삭제
  - 한자어의 경우
    예) ... 삼엄한 망루(望樓)가 서 있고, 굳게 ...
    수정 후) ... 삼엄한 망루가 서 있고, 굳게 ...
- ○ 기타 특수 기호, 외국어로만 이루어진 문장 등의 삭제
- ○ 신문 기사의 제목 등 문장 발성 목록으로 부적절한 것 등의 삭제
- ○ 한자가 직접 사용된 것 삭제
  예) 勸君獅子吼 莫學野干鳴 若能香象起 感得鳳凰道
  예) 經體本無名 受待無色聲 心依無相理 眞是金剛經
- ○ 중복된 문장
  - 중복된 문장이 자주 출현할 수 있으므로 전체 문장을 정렬하여 중복된 문장은 하나의 문장만을 두고 나머지는 삭제

위의 과정을 통하여 발성 목록의 후보로 177만 문장을 선정하였다.

(2) 발성 목록 선정을 위한 고빈도 어절의 선정

어휘의 경우, 외국의 예에서는 주로 고빈도 5,000단

어수준의 것과 20,000단어수준의 것을 고려하는 경우가 많다. 이는 실제적으로 대어휘 연속음성 인식의 훈련 및 평가목적으로 쓰이므로 이 정도가 일반적이다. 여기에 미지어의 문제를 고려하여 5,000단어세트나 20,000단어세트에 해당 단어세트에 포함되지 않은 한두개 정도의 단어를 포함한 문장세트도 구성하고 있다. 우리말의 경우에도 이러한 방법을 취하는 것이 바람직하다고 판단되어 모집단 4,300만 어절에 대해 고빈도 어절 분포 조사를 실시하였으며 그 결과는 다음 그림과 같다.

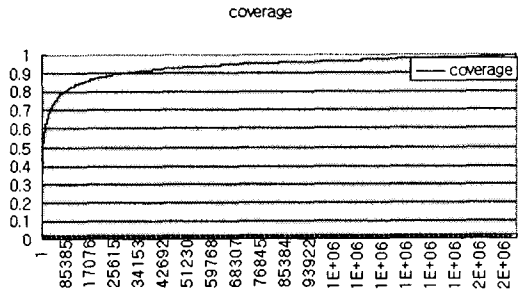

그림1. 4,300만 어절코퍼스에서의 어절 출현 빈도에 대한 누적분포

분석을 통해 상위 고빈도 5,000어절이 전체 어절에 대해 50.6%의 coverage를 가지며 상위 10,000어절의 경우 58.4%, 20,000어절의 경우 66.2%의 coverage를 나타내는 것으로 분석됐다. 상위 고빈도 77,000어절의 경우 약 80%의 coverage를 갖게되나 이러한 경우 어휘의 수가 너무 많아지게 된다.

따라서, 본 연구의 경우, 발성 목록 선정을 위해 고빈도 5,000어절, 8,000어절 및 10,000어절을 발성 목록 선정을 위한 대상어휘로 선정하였다.

(3) 발성 목록의 선정

위에서 선정된 문장에는 너무 길거나 짧은 문장이 다수 포함되어 있을 수 있으므로 전체 문장의 길이에 대한 조사를 한 후, 문장길이가 6어절 이상 25어절에 해당하는 문장만을 발성 목록의 대상으로 고빈도 10,000어절만으로 구성된 문장을 추출하였다. 추출된 문장의 총 수는 20,833문장으로 문장의 평균 길이는 문장 당 7.43어절이다.

또한 인식 대상 어휘에 포함되지 않은 단어가 발성된 경우에 대처하기 위한 OOV(Out of vocabulary) 테스트를 위해 다음과 같이 문장 목록을 구성하였다.

○ 5K 문장 세트 (8,608 문장)
- 고빈도 5,000어절에 포함된 어휘만으로 구성된 문장 세트

○ 8K-5K 문장 세트 (7,301 문장)
- 고빈도 8,000어절에 포함된 어휘만으로 구성된 문장 세트을 구성하고, 여기에서 5K 문장 세트에 포함된 문장은 중복되므로 이를 삭제한 것
- 즉, 고빈도 8,000어절에 포함된 어휘만으로 구성되어 있으며 고빈도 5,001 ~ 8,000 순위에 포함된 어휘가 1개 이상 포함된 문장 세트
- 5K 문장 세트와 8K-5K 문장 셋을 합하면 8K 문장 세트를 구성할 수 있음
- 8K-5K 문장 세트는 5K 문장 셋에 대한 OOV 테스트용으로 사용 가능

○ 10K-8K 문장 세트 (4,924 문장)
- 고빈도 10,000어절에 포함된 어휘만으로 구성된 문장 셋을 구성하고, 여기에서 5K 문장 세트와 8K-5K 문장 세트에 포함된 문장은 중복이므로 이를 삭제한 것
- 즉, 고빈도 10,000어절에 포함된 어휘만으로 구성되어 있으며 고빈도 8,001 ~ 10,000 순위에 포함된 어휘가 1개 이상 포함된 문장 세트
- 5K 문장 세트, 8K-5K 문장 세트 그리고 10K-8K 문장 세트를 합하면 10K 문장 세트를 구성할 수 있음
- 10K-8K 문장세트는 8K 문장 세트에 대한 OOV 테스트용으로 사용가능

위와 같은 과정을 거쳐 선정된 문장 목록의 예는 다음과 같다.

sent001 :
가게 안에서 사람이 나오는 바람에 여옥은 다시 걸음을 옮겼다.

sent002 :
그 결과 두 사람은 다음과 같은 결론을 내렸다.

sent003 :
그 사람은 정말 이름 그대로, 못된 사람입니다.

sent004 :
그가 내 이름으로 말할 때에, 내 말을 듣지 않는 사람은, 내가 벌을 줄 것이다.

sent005 :
그날 밤, 남편은 좀처럼 돌아오지 않았다.

3. 무제한 어휘 합성용 음성 DB를 위한 발성 목록 설계

가. 발성 목록 설계

여기에서는 음성 합성 시스템의 개발에 사용가능한 다양한 음소환경을 반영한 발성 목록을 선정하는 절차와 선정된 목록에 대하여 기술한다.

(1) 발성 목록 선정의 모집단

음성 합성 시스템을 위한 발성 목록 선정의 모집단으로 사용된 텍스트 코퍼스는 설명문, 수필문, 사회학, 방송 3社(KBS, MBC, SBS 등)의 뉴스, 신문(조선일보, 한국일보), 경제학, 전산학, 기계학, 생물학 등의 장르별 균형 텍스트로 구성된 KAIST Taged Corpus 100만 어절을 사용하였다. 사용된 텍스트 코퍼스의 특성은 다음과 같다.

■ 어절수: 1996년도 약 424,300 어절
1997년도 약 276,562 어절
1998년도 신문 코퍼스 약 196,000 어절
1999년도 코퍼스 약 100,000 어절

(2) 발성 목록 선정을 위한 텍스트 전처리

텍스트 코퍼스는 그 내용이 문자언어이므로 자연스럽게 낭독하는데는 곤란한 표현이 존재할 수 있다. 따라서 이러한 문장은 발성 목록의 후보에서는 제외하여야 한다. 발성 목록의 설계를 위해 제외 또는 수정된 사항은 Dictation 용 음성 DB의 경우와 동일하며, 이러한 과정을 통하여 발성 목록의 후보로 6만 문장을 선정하였다.

(3) 글자/음운 변환

발성 목록 선정 알고리즘을 적용하기 위한 전 단계로 모든 문장을 읽기 규칙을 적용하여 Triphone 단위로 글자/ 음운 변환(Grapheme to phoneme conversion)을 하였다. 변환된 예는 다음과 같다.

변환 전 :

어쨌든 이 책의 의도는 다위니즘의 일반적 옹호에 있는 것이 아니다

변환 후 :

sil-ʊ+Z ʊ-Z+E Z-E+d' E-d'+D d'-D+U D-U+n U-n+i n-i+c i-c+E c-E+g E-g+E g-E+Wi E-Wi+d Wi-d+o d-o+n o-n+U n-U+n U-n+d n-d+a d-a+wi a-wi+n wi-n+i n-i+z i-z+U z-U+m U-m+E m-E+i E-i+l i-l+b l-b+a b-a+n a-n+z n-z+ʊ z-ʊ+g' ʊ-g'+o g'-o+N o-N+h N-h+o h-o+E o-E+i E-i+n i-n+n n-n+U n-U+n U-n+g n-g+ʊ g-ʊ+s ʊ-s+i s-i+a i-a+n a-n+i n-i+d i-d+a d-a+sil

(4) 발성 목록 선정 알고리즘

위에서 선정된 모든 후보 문장을 발성 목록으로 사용하기에는 그 양이 너무 많으므로 본 연구에서는 Greedy Algorithm을 적용하여 발성 목록을 선정하였다. 상세한 절차는 다음과 같다.

① 모집단에서 유일하게 출현한 Triphone을 갖는 문장은 모두 발성 목록에 추가

② 나머지 모집단 문장 중 발성 목록에 나타나지 않은 Triphone 종류를 가장 많이 갖는 문장을 찾아 발성 목록에 추가

③ 발성 목록에 나타난 Triphone의 종류수가 모집단의 Triphone의 종류수와 같으면 발성 목록 선정 절차 종료, 그렇지 않을 경우 ②의 과정 반복

위와 같은 과정을 거쳐 최종적으로 선정된 문장 발성 목록은 4,360문장이며 이 문장에 포함된 Triphone의 총 종류수는 모집단과 같이 18,025 종류이다. 선정된 발성목록의 예는 다음과 같다.

sent0001 :
진화와 다위니즘 어떤 행성에서 지적 생물이 성숙했다고 말할 수 있는 것은 그 생물이 자기의 존재 이유를 처음으로 발견했을 때이다.

sent0002 :
어쨌든 이 책의 의도는 다위니즘의 일반적 옹호에 있는 것이 아니다.

sent0003 :
그것은 우리의 사회 생활의 여러 면 예를 들어 사랑과 미움, 싸움과 협력, 줌과 훔침, 탐욕과 관대에 관한 것이다.

sent0004 :
진화에 관한 로렌츠의 견해를 내가 이해한 바로는 그는 테니슨의 이 유명한 어구가 의미하는 것을 배척한다는 점에서 몬태규와 같다.

sent0005 :
만약 어떤 남자가 시카고에 갱단에서 오랫동안 순조롭게 살아왔다고 할 때 그 사람이 어떤 종류의 사람일까에 대해 어느 정도 짐작이 가능하다.

4. 결론

최근 컴퓨터와 인간간의 대화 수단으로 음성을 활용하는 기술인 음성정보기술이 발달되고 이를 상품화하는 움직임도 활발하다. 기술개발 측면에서도 대어휘 연속 음성 인식 및 무제한 어휘 음성 합성의 고도화를 비롯하여 대화음성의 인식 및 합성을 위한 연구도 진행되고 있고 이에 필요한 잘 정비된 음성코퍼스는 필수적인 요소이다. 본 연구에서는 이러한 코퍼스를 구축하는데 필요한 발성목록을 설계함에 있어서 다양한 발성환경을 포함하여 가급적이면 적은 양의 데이터로 풍부한 음운환경을 포함하도록 설계된 결과를 보고하였다.

이러한 연구의 결과로 얻어진 음성코퍼스는 음성정보기술개발을 체계적으로 지원하는 바탕이 될 것이다.

[참고문헌]

[1] 이용주, "음성언어코퍼스," 한국정보과학회지, 1998.2.

[2] 이용주, *한국어 음성 DB 구축에 관한 연구 제2차년도 최종보고서*, 한국과학기술원, 1996. 7

[3] 최기선, *KAIST 언어자원 2001년도판*, 과학기술부 핵심 소프트웨어 과제 결과물 1995-2000 (http://kibs.kaist.ac.kr)

[4] Dafydd Gibbon, Roger Moore, Rchard Winski, *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter 1997

[5] 이용주, 김봉완, "음성연구 및 음성데이타베이스," 대한음성학회 음성학학술대회자료집, 1996. 2

[6] Bong-wan Kim, Sun-Tae Kim, Tae-Hwan Kim, Young-Il Kim, "Design and construction of Korean speech database for common use," ICSP 97 Aug. 1997