研究活動

研究 및 評價用 音声 데이터베이스의 開發動向과 提案

李勇柱 · 金敬泰/ 音響研究室

〈要　約〉

음성인식 기술개발을 위해서는 음성 데이터베이스가 필요하다. 본고에서는 음성 입출력기술 표준화의 일환으로서의 공통음성 데이터 제정에 관한 각국의 현황을 소개하였고, 우리말을 대상으로 한 음운특성 연구용의 데이터베이스를 확보하기 위한 고려사항을 검토하였다.

Ⅰ. 서 론

음성인식 기술개발을 위한기본적인 도구로서 음성 데이터베이스가 필요하다. 이 음성 데이터 베이스는 주로 평가시험용 등에 이용되는 공통 데이터로서의 성격을 가진 것과 음성의 음향적인 특성을 연구하기 위한 것으로 나누어 생각할 수 있다. 공통 음성데이터는 음성인식 장치가 점차 실용화 되어감에 따라 입력장치의 성능을 평가하거나 각종의 분석방식을 비교할 때의 기준으로서, 공통으로 이용할 수 있는 음성 데이터를 의미한다. 공통 이용 가능한 데이터를 수집, 보관, 공개함으로써 연구개발의 입장에서는 분석방식 및 인식 알고리즘의 개발, 평가에 이용할 수 있고 사용자의 입장에서는 인식장치의 성능평가의 비교를 객관적으로 할 수 있게 된다.

한편 단어 단위 인식을 기본으로 한 특정화자 소용량 음성 인식장치가 각국에서 실용화 된 후 대 어휘를 다루는 음운단위 불특정화자용 음성 인식 장치 및 연속 음성을 대상으로 한 연구가 활발해지고 있는데, 단어 또는 연속 음성중의 음운은 화자에 따른 영향은 물론이고 전후에 발성된 음운의 영향(조음결합)에 의해서도 크게 변하게 되어 이러한 현상을 파악하는 것은 음운 인식에 필수불가결한 일이 되므로 대량의 음성 데이터의 수집 및 분석이 필요하게 된다. 따라서 임의의 음운 환경하에서 원하는 음운을 검색할 수 있도록 하여 음향 특성을 살피거나 음운단위 인식 실험에 이용할 수 있는 음소 단위로 labelling된 데이터 베이스도 요구된다.

본고에서는 먼저 음성입출력 기술 표준화의 일환으로서의 공통 음성데이터 제정에 관한 각국의 현황을 소개하였고, 우리말을 대상으로 한 음운특성 연구용의 데이터 베이스를 확보하기 위해 고려할 사항들을 기 구성된 시스팀을 참고로 하여 검토하였다.

Ⅱ. 음성입출력 기술의 표준화와 공통음성데이터

공통음성데이터는 음성입출력 기술 표준화의 일환으로 각국에서 연구되고 있다. 미국에서는 1982년에 음성입출력 기술 workshop이 NBS (National Bureau of Standards)에서 개최되었고 여기에는 음성 입출력기기의 제조회사, 음성 연구자 뿐만 아니라 이용자까지 참가하여 음성 인식 시스팀의 평가, 음성 입출력 시스팀의 성능 평가를 위한 데이터 베이스 등에 관한 토의가 있었다.

또한, 과학 아카데미는 많은 정부기관의 요청을 받아 컴퓨터 음성 인식기술위원회를 발족시켰다. 여기서는 현재의 음성 인식 기술 수준을 평가한 후, 검토해야 할 과제로 음성인식 기술과 알고리즘, 음성열화 및 인간요인의 3가지로 결정하고 이에 대한 연구를 진행하고 있다. 이 위원회는 특히, 소수의 실험적인 데이터 베이스 이외에 앞으로 널리쓰일 공통의 데이터 베이스와 평가 방법을 만들어 중앙에서 책임을 지고 보관 및 배포를 담당하도록 제안했다.

NBS는 음성입출력 기술 표준화에 관한 기본적인 문제로 연구실 및 실제의 사용 환경에서의 음성 데이터 베이스, 성능평가법 및 효과적인 운용을 위한 지침 등을 검토후, 토의를 거쳐 음성 인식 장치의 성능 평가 지침 초안을 마련하였다.[1]

IEEE ASSP의 음성처리기술위원회에서도 음성입출력 기술 평가를 위한 표준에 관한 working group이 발족되었다.

이밖에 기업, 연구소, 대학 등에서 미국음향학회, Speech Tech, Voice I/O등의 연구회에서 음성인식 장치의 여러가지 환경하에서의 성능평가, 평가를 위한 음성데이터 베이스에 관한 연구가 적지않게 보고되고있다.

현재 계획중이거나 진행중인 데이터 베이스는 1985년 부터 시작한 DARPA의 음성인식 프로젝트하에서 이루어지고 있는데 다음의 3종류가 고려되고 있다.[2] 첫번째로는 10～20명의 화자에 의한 100～200단어 데이터 베이스로서 녹음조건은 잡음포함, 긴장상태에 있는 화자의 발성도 포함된다. 이는 비행기 조종실내의 조종사의 발성을 감안한 것이다. 두번째는 특정한 task의 연속 음성으로 녹음조건은 잡음을 포함시키고 화자의 즉흥적인 회화를 수록한다.

세번째 음성학적인 것을 대상으로 한 데이터 베이스는 미국내의 4～6지역에서 선발된 1,000명의 화자에 의한 음성을 수록한다. 각 화자에 대해서 30초간의 음성을 기본으로 하고 있지만 자유로운 회화음성을 녹음하는 것이 고려되고 있으며 segementation 및 음소구간의 labelling을 MIT에서 하는 것으로 되어 있다.

NATO음성처리 연구 그룹에서는 NATO 가맹국 중 6개국이 참가하여 음성인식 기술의 유용성에 대해서 검토를 시작했다. 현재 NATO 가맹국의 여러 언어에 의한 연속 숫자 음성 인식에 대해 공동 연구중이며, 19명의 발성한 음성 자국어 및 타국언어의 연속숫자 36,000어의 데이터 베이스를 만들어 각종 인식시스팀의 성능을 평가하고 있다.

프랑스에서는 GRECO project (Concerted Research Groupon Speech)가 입안되어 현재 많은 연구 기관이 참가하고 있으며 음성 데이터 베이스로서 10명의 프랑스인이 고립 또는 연속 발성한 숫자와 알파벳을 작성했다.[1] 또한 ENSERG group도 프랑스어의 대규모 데이터 베이스 작성을 시작했다.

일본에서는 1980년부터 일본전자공업진흥

협회에 일본어 정보처리 표준화 조사위원회가 구성되어 일본어 입력 방식에 관한 표준화 연구를 수행하였는데 그 결과로 숫자, 4연숫자, 기능어, 단음절, 외래어, 지명 등의 단어 리스트를 선정하고 이 음성 데이터를 비디오테이프에 의한 PCM 녹음형태로 수록하여 이중, 단일 및 4연숫자를 비롯한 74명분을 편집하여 배포중이다.[4]

CCITT에서도 1984년부터 SG Ⅻ의 Question 5로 음성인식 및 합성 장치의 성능평가 방법을 검토하고 있다.

이와같이 각국에서는 각기 자국어를 대상으로한 공통 데이터 작성의 움직임이 활발하다.

Ⅲ. 음소 Labelling Data Base

여기서는 이미 구성된 시스팀의 예와 발성용 텍스트생성 및 labelling방법 등을 논하고 한국어를 대상으로 데이터 베이스 구성시 연구 및 검토되어야 할 사항에 대하여 기술한다.

1. 시스팀 구성예

일본의 전자기술총합연구소(ETL)는 대용량의 음성 데이터를 저장, 검색할 수 있는음성연구용 파일 데이터 제어 시스팀.[5]을 개발한 바 있는데 여기서는 light pen과 graphic display를 이용하여 대화형식으로 애널로그 데이터의 수집, 음성 파형과 음소 기호의 대응(음소 labelling), 음소의 검색 및 청취실험이 가능하도록 구성되어 있어서 실험 ID, 단어, 음소 등의 지정에 의해 임의로 검색할 수 있다. 계산기의 갱신에 따라 새로운 수집 시스팀[1]을 구성하였는데 수집과 편집의 두부분으로 나누어, 수집 프로그램에서는

- 음성데이터를 20KHz, 12bit로 A/D 변환하고
- 음성 구간 자동 추출에 의해 불필요한 무음 구간을 삭제해서
- 음성데이터를 단음절, 문장단위의 세그먼트로 나누어 대응하는 구간 길이를 검출한다.
- 청취, 혹은 파형 표시에 의해 세그먼트 구분을 확인하고 labelling한 후
- 데이터 내용을 보여주는 제어정보를 음성데이터의 선두에 부가하여 master file을 만든다.

편집 프로그램에서는

- 다수의 master file중에서 지정된 화자군(성인남자, 성인여성, 어린이 등) 에 대하여 지정된 음소계열을 찾아내고
- 지정된 주파수로 표본화 주파수를 변환하며
- Control word 를 부가하여 user file 을 작성하여 사용하도록 되어 있다.

오오사까 대학은 전국의 음성 연구자의 공동이용을 목표로 SPEECH - DB[7]를 개발했는데 범용 DBMS인 INQ를 이용하였고 이용자의 다양한 요구에 부응하기 위해 음운파일, 음절파일, 단어파일, 환경파일 및 데이터파일로 되어 있으며 이는 각 대학의 대형 계산기 센터의 네트워크를 통해 TSS 단말로 액세스 할 수 있도록 구성한 것이다.

동북대학은 마쓰시타와 함께 directory file, label file, data file로 구성되는 음소단위 데이터 베이스를 구성하여 사용하고 있다.

2. 발성용 Text의 생성

음운특성 연구용의 text로서는 모든 음운이 가능한 한 다양한 음운 환경중에 있도록 샘플을 취하는 것이 바람직한데 이를 위해서 주로 사전의 표제어나 고빈도 단어를 대상으로 음운 밸런스가 취해지도록 단어set을 선택하는 방법을 쓰고 있다. 즉, 앞의 음운이 뒤의 음운에 영향을 준다는 관점에서CV 등의 2음소열을 대상으로 하거나, 전후의 음운이 가운데 위치한 음운에 영향을 준다는

관점에서 CVC또는VCV의 3음소열을 대상으로 하는 경우가 있다.

구체적인 선택방법으로 ETL 에서는

- 모든 음소열을 적어도 1회 포함하면서 가능한 한 적은 단어수로 많은 종류의 음소열을 포함하도록 종류수를 중시하는 것과
- 음소열의 entropy를 최대화하는 방법의 2가지 선택기준에 의해 일단 어떤 단어를 선택함으로써 증가는 음소열의 종류가 최대가 되는 단어를 우선으로 하고, 종류수가 최대가 되는 단어가 복수일 경우 음소열의 entropy가 최대화하는 단어를 채용하는 방법[8]을 쓰고 있다.

NTT ECL은 2음소열을 대상으로 부가에 의해 entropy가 가장 커지는 단어를 축차 부가하는 Add법과 모집단 전체를 초기 set 로 한후 삭제에 의해 entropy가 가장 커지는 단어를 축차 삭제해 가는 Delete법, 그리고, 부가와 삭제를 반복하면서 최적한 set를 구하는 A&D 법등을 제안[9]하고 이 A&D 법을 사용하여 744지명 및 국어사전의 중요어 5,244단어를 대상으로 100여 단어를, ETL 의 경우는 국어사전의 44,000단어중 492 단어를 선택하여 사용하고 있으며 동북대학은 마쓰시타[10]와 함께 각음소의 출현빈도가 일본어의 고빈도 4,600단어에서와 비슷하도록 선정한 212단어 set를, 오오사까 대학은 5개의 연속음성과 512개 VCV 음절을 선정하여 사용되고 있다.

3. Labelling 방법

ETL은 초기의 시스팀에서는 파형 데이터와 텍스트의 음소 기호를 사람이 참조하면서 서로 연결짓고 불확실한 부분은 잘라낸 음소의 음성출력에 의해 확인하였다. 그러나 새로운 시스팀에서는 몇개의 식별 파라미터를 이용한 자동 labelling 시스팀으로 일단 labelling한 후 사람이 확인 수정을 하도록 하였다.

오오사까 대학의 시스팀에서는시간파형과 포먼트주파수의 시간적인 변화를 화면상에 동시에 display한 후 수동으로 labelling 하였으며 음운간의 과도부분과 정상성에 대한 정보도 저장하고 있다. 동북대학에서는 29ch의 필터 bank로 주파수 분석한 후 10ms를 한 frame 단위로 하여 power 정보로서 대수 spectral의 합, 그리고 유성/무성, 비음성, 마찰성 등의 판별분석의 값 및 digital sonagraph의 출력을 CRT에 동시에 표시하여 눈으로 확인하면서 labelling 하였다.

마쓰시타 통신은 유사도에 의한 모음후보, 자음후보, U/V판정, 전 power, 고역 및 저역 power, spectral의 경사 등을 사용하여 labelling하고 있다.

4. 우리말 음성데이터 베이스

우리말을 대상으로 한 데이터 베이스 구축을 위해서는 다음과 같은 사항들이 연구 및 검토되어야 할 것이다.

가. 음소의 종류 및 분류

우리말의 음소를 몇 종류로 하며, 어떻게 분류할 것인가에 관해서는 음운특성의 변화를 연구한다는 관점에서 보면 가장 기본적이고 중요한 일이 된다. 그러나 이를 절대적으로 구분하는 통일된 기준은 없고 국어학자마다 주장하는 바가 다르므로 각 주장중 공통이 되는 면을 참고하여 기준을 마련한 후 우선 사용하고 표기법은 일반의 영문 터미널로 입력 가능하도록 미리 정해둘 필요가 있다. 특히 우리말의 경우, 초성과 종성의 음향특질이 같지 않으므로 구분해서 표기하는 등 세심한 검토가 필요하다.

이렇게 하여 일단 구성된 데이터베이스 를 이용하여 연구된 각 음소의 음향적 특성들은 앞으로의 통일된 음소 분류에도 기초자료가 될 것이다.

나. 발성용 text 선정을 위한 기초조사

다양한 음운환경의 단어를 text로 선정하기 위해서는 음소 표기된 대상 단어집단(고빈도 단어, 사전의 표제어, 중요어 등)의 컴퓨터 입력이 요구된다. 이를 위해 대상 단어집단 중 품사나 음절수를 기준으로 미리 제한하거나 고빈도 단어의 선정 등의 기초 조사한 예[11]도 많은 참고가 되겠으나 글자가 아닌 말을 대상으로 함에 따라 따로 고려해야 할 바가 많다.

다. 데이터 베이스

데이터 베이스의 구조에 따라 이용시의 편리성이나 효율에 크게 영향을 미치므로 이미 만들어서 사용중에 있는 음성 데이터 베이스의 예를 세밀히 검토하여 참고할 필요가 있다.

라. 자동 labelling

대상으로 하는 단어가 많아지면 labelling 작업을 수행하는 사람의 능력에 따라 그 결과가 제각기 다르게 되고 또한 많은 시간이 소요되므로 자동 labelling기법을 채용하지 않으면 안된다. 자동 labelling에서는 segmentation및 labelling에 유효한 파라미터를 찾는일이 중요하며 이를 위해서는 소량의 데이터를 대상으로 하여 수동으로 labelling 하면서 얻어지는 know how와 이미 개발된 자동 labelling기술을 능동적으로 수용해야 할 것이다.

마. 수집 및 편집시스팀

애널로그 데이터의 수집, 변환 및 labelling이 가능한 그래픽 기능이 완비된시스팀의 구성이 필요하다. H/W적인 사양도 까다롭지만 이러한 작업을 수행하기 위한 S/W의 제작에 많은 시간이 소요될 것이다.

Ⅳ. 결 론

지금까지, 공통된음성 데이터 베이스의 개발 동향을 소개하였고 음소 데이터베이스 개발예를 통하여 우리말 데이터 베이스 구축을 위한 고려사항을 논하였다.

가까운 장래에 우리나라도 국내 개발 또는 수입된 각종 음성인식 장치가 산업현장 또는 통신망 등에 활용할 수 있게 될 것이므로 우리 나름의 성능평가를 위한 공통음성 데이터의 제정을 위하여 정부, 학계, 산업계의 관계자로 구성된 조사 연구 그룹의 설치가 필요할 것이다. 또한 음소 데이터 베이스는 그 구성에 소요되는 장비, 인원, 예산 및 음성연구의 저변확대 등의 파급 효과를 고려하여 공공 연구 기관이 주관하여 제작 후 공동이용토록 제안한다.

한국전자통신연구소에서는 현재, 음소 데이터 베이스 구성에 관한 기초 연구를 추진중에 있다.

〈參 考 文 獻〉

1. Barker, at al., Proc. ICASSP 83, pp. 527～530, 1983.
2. 脇田壽, 일본음향학회지, Vol. 41－10, pp. 739～744, 1985.
3. Perennou, Proc. ICASSP 86, pp. 325～327, 1986.
4. 일본전자공업진흥협회, 일본어 정보처리 표준화에 관한 조사연구보고서, 1983, 84, 85.
5. 中島隆之외, 일본음향학회 음성연구회 자료, S73－07, 1973.
6. 速水悟외, 일본음향학회 학술발표회논문집, 3-6-15, 1982. 10.
7. 溝口理郎외, 일본정보처리학회 논문지, Vol. 24－3. 1983. 5.
8. 速水悟외, 일본음향학회 학술발표회논문집, 2-4-7, 1985. 10.
9. 鹿野清宏외, 일본음향학회 학술발표회논문집, 3-3-10, 1982. 3.
10. 秋場国夫외, 일본음향학회 학술발표회논문집, 1-4-22, 1982. 3.
11. 남궁 건, 서울대 석사학위 논문, 1982.