

공동이용을 위한 음성DB의 설계 및 구축에 관한 연구

A Study on the Design and the Construction of a Korean Speech DB for Common Use

김 봉 완*, 김 종 진*, 김 선 태*, 이 용 주*

(Bong-Wan Kim*, Jong-Jin Kim*, Sun-Tae Kim*, Yong-Ju Lee*)

요 약

공동이용 가능한 각종 대량의 음성 데이터를 수록, 보관, 공개하는 것은 연구 개발 과정에서의 이용 및 음성 정보 처리 시스템의 성능평가 양면에서 필요하다. 이러한 공동 음성 데이터 베이스의 구축을 위해서는 발생 가능한 모든 음운환경을 포함하며, 특정 태스크에 집중되지 않는 발성 대상 단어나 문장의 설계가 필요하다. 본 논문에서는 이와같은 목적으로 신문, 소설, 기타 구어자료로부터 수집된 120만여 어절의 텍스트 코퍼스에서 PBW(Phonetically Balanced Word)를 추출하고 이를 발성목록으로 음성DB를 구축한 결과와 구축된 음성DB의 특성을 제시한다.

ABSTRACT

Speech database is an indispensable part of speech research. Speech database is necessary to use in speech research and development processes, and to evaluate performances of various speech-processing systems. To use speech database for common purpose, it is necessary to design utterance list that has all the possible phonetical events in minimal number of words, and is independent of tasks. To meet those restrictions this paper extracts PBW set from large text corpus. Speech database that was constructed using PBW set for utterance list and its properties are described in this paper.

1. 서 론

음성은 인간끼리의 정보 전달 매체로서 옛날부터 가장 많이 이용되어 왔다. 이러한 손쉬운 통신 수단인 음성을 인간과 기계와의 대화 수단으로 이용하고자 하는 시도가 오래전 부터 있어 왔다. 이를 위해서는 그 기반이 되는 요소기술로써, 음성 인식 및 합성으로 대표되는 음성처리 기술과 언어이해 및 기계번역으로 대표되는 언어처리 기술에 대한 연구가 필요하다. 이러한 음성 및 언어처리 기술의 연구를 위해 가장 먼저 확보 되어야 할 것이 음성, 언어 및 각종 사전 데이터 베이스이다.

음성연구가 진보되어감에 따라 처리 가능한 데이터수는 많아져 가고, 따라서 준비해야할 데이터량도 대폭적으로 증가되었다. 최근에는 음성인식의 경우 통계적 수법의 발달에 따라 대량의 음성 데이터가 시스템의 학습에 필요하게 되었으며, 합성의 경우에도 대형의 음성 데이터로부터 입의 길이의 음성부분을 골라내어 접속하므로써 좋은 합성품질을 얻고 있다. 또한 음성정보처리 시스템의

연구개발을 위해서는 분석, 합성, 인식의 각종 알고리즘을 적절하게 비교 평가할 필요가 있지만 이를 위한 방법으로는 현재까지는 공동음성 데이터를 이용하여 알고리즘을 수행하고 그 결과를 비교하는 방법 이외에는 알려져 있지 않다. 이러한 목적으로 이용하는 음성 데이터를 일반적으로 음성 데이터베이스, 음성 코퍼스 또는 음성 사전이라고 부른다(이하 음성DB로 표기).

선진 각국에서는 자국어 음성DB에 대한 체계적인 구축이 음성 및 언어처리기술 확보를 위한 가장 기본적인 연구환경임을 깊이 인식하고 이에대한 체계적인 확보가 공공연구기관을 중심으로 활발히 추진되고 있다[1]. 그러나 국내의 경우 지금까지 음성인식 및 합성연구에 필요한 음성 데이터를 각 연구자가 부분적으로 제작하여 자체 개발용으로 사용하고 있는 경우가 대부분이다[2].

다른 연구 목적간에 두루 쓰일 수 있도록 공동이용을 목적으로 음성DB를 구축하기 위해서는 가급적 다양한 종류로 대량의 것이 바람직하나 제한된 자원에서 가능한 양으로 목적을 달성하기 위해서는 대상이 되는 발성목록을 어떻게 작성할 것인가 하는 점이 중요한 문제가 된다[3, 4, 5]. 이러한 공동 음성 데이터 베이스의 구축을 위해서는 발생 가능한 모든 음운환경을 포함하며, 특정 태스크에 집중되지 않는 발성 대상 단어나 문장의 설

*원광대학교 컴퓨터공학과
접수일자: 1997년 1월 22일

게가 필요하다.

본 논문에서는 이와같은 목적으로 신문, 소설, 기타 구 어자료로부터 수집된 120만여 어절의 텍스트 코퍼스에서 PBW(Phonetically Balanced Word)를 추출하고 이를 발성 목록으로 음성DB를 구축한 결과와 구축된 음성DB의 특성을 제시한다.

II. 발성 목록의 설계

2.1 PBW(Phonetically Balanced Word)

2.1.1 PBW의 정의

음성인식 및 합성 시스템의 개발 등 음성 연구에 있어서 다종 다양한 음운환경을 포함한 음성DB의 구축은 중요한 과제의 하나로 많은 시간과 노력이 필요하다. 개별적 음성DB 구축에 따른 중복부자를 줄이고 분석, 합성, 인식의 각종 알고리즘을 적절히 비교 평가하기 위해서도 공통의 음성DB는 필수적이다. 이러한 목적으로 사용 될 음성DB는 적은 발성목록으로 실제 한국어의 발성에 나타나는 음운환경을 가능한 많이 포함하며, 특정 테스트에 집중되지 않는 것이 바람직하다. 이러한 조건을 충족시키는 발성목록으로 일본을 비롯한 선진국에서는 PBW라는 단어집합을 모아왔다[6, 7, 8, 9].

PBW란 발생가능한 모든 종류의 음운환경을 나타내는 음소열을 포함하되 음소열간 고른 확률분포를 갖는 최소 단어들의 집합을 말한다. 음소열간 고른 확률분포를 갖는 상태는 음소열을 확률사상으로 했을 때 엔트로피가 최대인 상태이다[10]. 낱말 세트에 나타난 총 음운환경의 종류를 N , 음운환경의 출현확률을 P_n , ($n=1, \dots, N$)라고 할 때, 낱말세트의 음운 엔트로피 S 는 다음식(1)에 의해 구할 수 있다.

$$S = - \sum_{n=1}^N P_n \log_2 P_n \quad (1)$$

이때 엔트로피 S 는 음운환경의 출현빈도가 모두 동일할 때 최대치 $\log_2 N$ 이 된다.

또한 음성DB를 위한 발성목록은 그 양이 많아지게 되면 발성자료를 수집하는 데 어려움이 따르게 된다. 따라서 PBW가 발성목록으로 사용되기 위해서는 엔트로피를 최대화 하면서도, 최소단어들의 집합에 발생가능한 모든 음운환경들을 포함할 수 있도록 알고리즘을 구성하여야 한다.

2.1.2 PBW추출을 위한 음운환경의 정의

실제 발화시 음성의 음향적 특징은 전후 음소의 영향을 받아서 변화하는 경우가 많다. 음운환경을 나타내는 음소열은 많을수록 좋다고 할 수 있으나 음소열의 증가에 따라 다루어야 하는 음운환경의 수와 계산량이 급격히 증가하게 된다. 따라서 본 논문에서는 한국어의 발성에 나타날 수 있는 모든 2음소열을 PBW추출의 기본단위인 음운환

경으로 정의하고, 추후 음성DB를 확장할 때 3음소열을 고려하기로 하였다.

본 논문에서 사용하는 대상음소로는 자음 19개, 모음 21개로 구성하고 종성자음은 모두 대표자음으로 모아 다루었다. 따라서 종성자음으로 올 수 있는 음소로는 /ㄱ/, /ㄴ/, /ㄷ/, /ㄹ/, /ㄴ/, /ㄷ/, /ㄹ/, /ㅁ/, /ㅂ/, /ㅅ/의 7가지 이다. 모음의 경우 반모음은 후속모음과 합하여 하나의 음소로 취급하였다.

대상 2음소열은 위와 같이 분류된 음소의 쌍으로 구성하였으며, 실제 발성할 때 특히 자음의 경우 같은 음소라도 어두, 어말의 위치에서 음향적 특성이 서로 다를 가능성이 있기 때문에 어두, 어말의 공백소도 하나의 음소로 취급하여 공백소와 해당음소가 쌍을 이루는 형태를 2음소열의 종류에 포함 하였다(즉, Bㄱ은 ㄱ이 어두에 초성으로 오는 경우, ㄱB는 ㄱ이 어미에 종성으로 오는 경우를 말한다.). 다음 표 1에 음운환경을 구성하는 음소들을 나타내었다.

표 1. 대상음소

자 음	ㄱ, ㄴ, ㄷ, ㄹ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㆁ
모 음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅚ, ㅜ, ㅡ, ㅝ, ㅟ, ㅡ, ㅞ, ㅟ, ㅡ, ㅠ, ㅢ, ㅡ, ㅣ, ㅤ, ㅥ, ㅧ, ㅨ, ㅩ, ㅪ, ㅫ, ㅬ, ㅭ, ㅮ, ㅯ, ㅰ, ㅱ, ㅲ, ㅳ, ㅴ, ㅵ, ㅶ, ㅷ, ㅸ, ㅹ, ㅺ, ㅻ, ㅼ, ㅽ, ㅾ, ㅿ, ㅿ, ㅿ, ㅿ
공백소	B

2.2 PBW추출을 위한 모집단

본 논문에서는 PBW를 추출하기 위한 모집단으로 신문, 소설, 수필등의 자료로부터 수집한 120만여 어절의 텍스트 코퍼스(Text Corpus)에서 가급적 많은 음운환경을 포함시키기 위하여 3음절 이상의 고빈도어절만을 선정하여 PBW추출의 대상으로 하였다.

고빈도어절을 추출한 텍스트 코퍼스는 다음과 같은 자료로부터 수집된 것이다.

- (가)신문기사
 - 조선일보(1993년) 78,000어절
 - 동아일보(1993년) 102,000어절
 - 한겨레신문(1993년) 126,000어절
 - 이규태코너 51,000어절

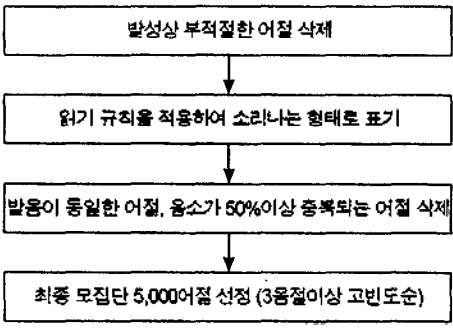


그림 1. 모집단의 선정 절차

(나) 소설(4종)	228,000어절
(다) 수필, 기행문 등	478,000어절
(라) 구어자료	41,000어절
(마) 기타	124,000어절

이후 PBW를 추출하기 위한 모집단을 선정하는 과정을 그림 1에 나타내었다.

2.2.1 발성상 부적절한 어절 삭제

선정된 고빈도어절들 중 음성DB를 위한 발성목록으로 부적절하다고 판단되는 다음과 같은 어절들은 모집단에서 제외하였다.

- (가) "콜린턴", "이제필" 등 인명이 포함된 어절
- (나) "무슬란"과 같이 외국어나 의미가 난해한 어절
- (다) "지섬은", "채린은" 등과 같이 소설등의 자료에서 포함되거나, 자료의 성격상 고빈도어절로 포함된 전문용어 등의 어절
- (라) "그라본", "빌어목을" 등의 방언
- (마) 그밖의 비어, 속어 등 발성상 부적절하다고 판단되는 어절

2.2.2 소리나는대로의 표기 변환

이상에서 추출된 3음절 이상의 고빈도 어절들을 글자 음운 변환기에 의해 소리나는 대로로 형태로 자동 변환하고 잘못 변환된 어절은 직접 수정을 통하여 올바르게 표기하였다. 글자 음운 변환기는 한글을 소리나는 대로 바꿔주는 읽기 규칙을 구현한 것이다[11]. 읽기 규칙의 기본 원칙은 교육부에서 고시한 표준어 규정의 표준 발음법을 따르고 있다. 읽기규칙은 크게 모음의 발음, 장음 처리, 음운 변동 규칙 및 음의 첨가의 4가지로 나눌 수 있으며, 본 논문에서는 장음 처리를 제외한 3가지의 규칙을 적용하고, 두가지 이상으로 적용 가능한 경우는 보다 일반적인 경우를 적용하였다. 그리고 규칙으로 구현할 수 없는 경우는 예외 발음 사전에 추가하여 처리하도록 하였다.

2.2.3 동일 위치에 음소가 50%이상 중복되는 어절의 삭제

"기운이[기+니]", "기우니[기+니]" 등과 같이 서로 다르게 표기되지만 동일하게 발성되는 이철동음어는 하나의 어절만을 취하고 나머지는 삭제하였다. 또한 본 논문에서 PBW추출의 대상으로 삼은 것은 단어단위가 아닌 어절단위 이므로 단지 어미의 변화에 의해 고빈도어절중에 포함된 어절이 많이 발생하게 된다. 이러한 어절들을 배제하기 위하여 음소변환을 통하여 동일한 위치에 50% 이상의 음소가 중복되는 어절은 가장 빈도수가 높은 어절을 취하고 나머지는 삭제하였다. 예를들어 "언어가[너너가], 언어는[너너는], 언어의[너너기]", "것이기[거시기], 것이며[거시며], 것이요[거시요]"와 같은 어절들이다.

2.2.4 최종 모집단의 선정

이상과 같은 처리를 거쳐 3음절 이상을 대상으로 하여 고빈도순으로 최종 모집단 5,000어절을 선정하였으며, 모집단에는 텍스트코퍼스에서 3,174번의 출현빈도를 갖는 "그러나"에서부터 8번의 출현빈도를 갖는 "과시했다"까지가 포함되었다.

2.3 PBW 추출 알고리즘

모집단에서 발생한 음운환경 n 의 출현빈도를 N_{Pn} , PBW 후보세트에서의 출현빈도를 N_n , 모집단에서 발생한 음운환경의 총 종류수를 T_p , PBW 후보세트에서의 음운환경의 총 종류수를 T 라고 한다면 본 논문에서 사용한 PBW 추출 알고리즘은 다음과 같다.

2.3.1 초기 PBW 후보세트의 구성

단계 1. 모집단의 모든 어절들을 조사하여 $N_{Pn}=1$ 인 음운환경을 갖는 모든 어절을 모두 PBW 후보세트에 포함 시키면서 모집단에서 제거한다.

2.3.2 추가 과정

단계 2. 모집단의 나머지 어절들을 모두 조사하여, 각 어절에 포함된 음운환경 중 $N_n=0$ 인 음운환경을 가장 많이 갖는 어절을 임시 선택한다.

단계 3. 선택된 어절이 복수일 경우, 한 어절씩 PBW 후보세트에 추가하여 PBW후보세트의 엔트로피를 계산하였을 때 PBW 후보세트의 엔트로피가 최대화되는 어절을 PBW후보세트에 포함시키면서 모집단에서 제거한다.

단계 4. 단계 2~단계 3의 과정을 T 가 T_p 와 같아질 때까지 반복한다.

단계 5. PBW 후보세트에 포함되지 않은 모집단의 나머지 어절들을 모두 한번에 한 어절씩 PBW 후보세트에 추가해서 PBW 후보세트의 엔트로피를 계산해 보고, 추가되었을 때 엔트로피를 최대화 시켰던 어절을 후보세트에 추가하고 이 어절을 모집단에서 삭제한다.

단계 6. 단계 5의 과정을 PBW 후보세트의 엔트로피를 증가시키는 어절이 더이상 없을 때까지 반복한다.

2.3.3 삭제과정

단계 7. PBW 후보세트를 구성하고 있는 어절들 중, $N_n \geq 2$ 인 음운환경으로만 이루어진 어절들을 모두 임시 선택한다.

단계 8. 선택한 어절들을 모두 한번에 한 어절씩 PBW 후보세트에서 삭제해서 후보세트의 엔트로피를 계산해 보고, 삭제 되었을 경우 PBW 후보세트의 엔트로피를 최대화 시켰던 어절을 PBW 후보세트에서 삭제한다.

단계 9. 단계 7~단계 8의 과정을 $N_n \geq 2$ 인 음운환경으로만 이루어진 어절이 더이상 존재하지 않거나, 존재하더라도 삭제했을 경우 PBW 후보세트의 엔트로피를 증가시키는 어절이 더이상 없을 때까지 반복한다.

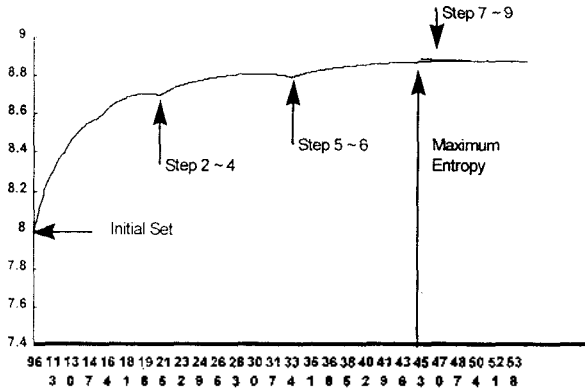


그림 2. PBW추출 과정의 엔트로피 변화

2.4 추출된 PBW

이상과 같은 알고리즘에 의하여 본 논문에서 PBW를 추출하는 과정에서 초기 후보세트로 95개의 어절이 추출되었으며 단계 2~단계 3의 과정을 통하여 238어절, 단계 5~단계 6의 과정을 통하여 167어절이 추출되었다. 이렇게 해서 추출된 총 500어절 중 단계 7~단계 9의 삭제과정을 통하여 총 48개의 어절이 삭제되어 최종적으로 452어절의 PBW를 추출하였다. 그림 2에 PBW를 추출하는 과정에서의 엔트로피 변화를 나타내었다.

모집단 고빈도 5,000어절과 PBW set에는 모두 842종류의 음운환경을 포함하고 있으며, 출현한 음운환경의 총 빈도수는 모집단에서 42,043, PBW set에서 3,504이다. 모집단에는 VC, CV, VV, CC가 각각 40.78%, 46.13%, 3.12%, 9.97%를 보이고 있으나 PBW에서는 37.50%, 43.81%, 9.54%, 9.15%의 분포를 이루고 있다. 모집단에서의 평균 음절수는 3.357, 평균 음소수는 7.408음소로 나타났고, PBW에서의 평균 음절수는 3.398음절, 평균 음소수는 6.752음

모집단에서의 2음소열 출현빈도

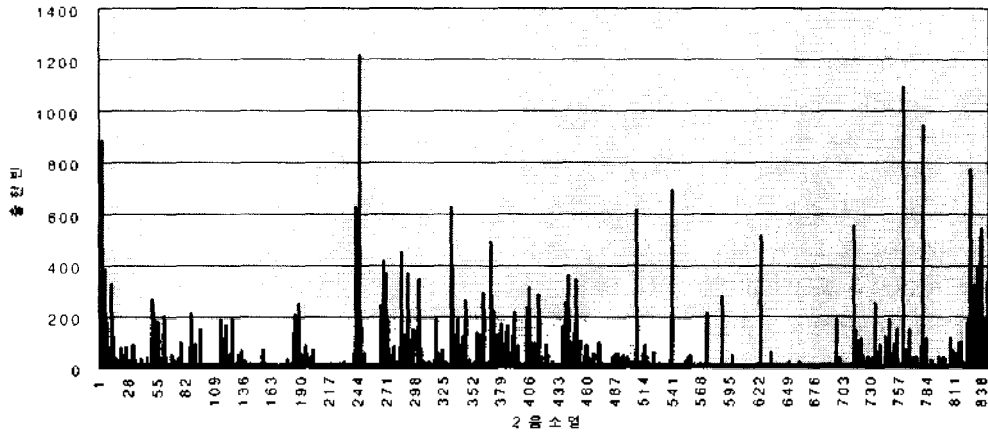


그림 3. 모집단에서의 2음소열 출현빈도

PBW에서의 2음소열 출현빈도

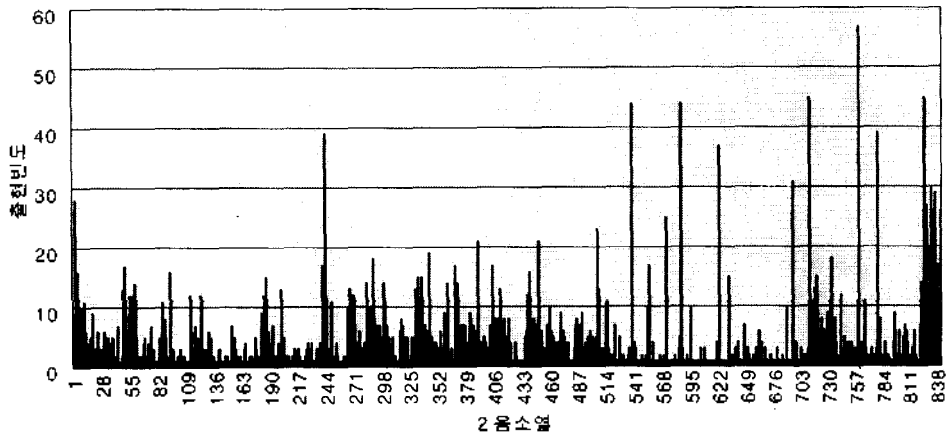


그림 4. PBW에서의 2음소열 출현빈도

소로 나타났다. 또한 추출을 위한 모집단에서의 엔트로피는 7.98993에서 추출된 PBW의 엔트로피는 8.88397로 음운의 균형이 모집단에 비해 좋아졌음을 알 수 있다. 그림 3과 그림 4에 모집단과 PBW에서의 2음소열 출현빈도를 나타 내었다. 그림에서도 PBW에서의 2음소열 출현빈도가 고른 분포를 보이고 있음을 알 수 있다.

2.5 기타 발성목록

무제한 어휘의 인식기 및 합성기를 위한 음성DB의 발성목록으로 태스크에 종속적이지 않으면서, 한국어의 음운환경을 충분히 고려한 PBW 외에도, 음성 연구자들이 꾸준히 요구하고 있는 것이 숫자음 음성이다. 관련연구에 의하면 필요로 하는 음성 데이터에 대한 설문 등에서 숫자음 음성은 우선순위가 높게 나타나고 있다[5]. 따라서 공동이용을 위한 음성DB의 기본적인 내용으로 포함되는 것이 바람직 하다.

2.5.1 숫자음

숫자음의 발성 목록으로는 우리말에서 흔히 쓰이는 단독숫자 37종과 지시어 4종, 그리고 숫자간 결합이 모두 포함되도록 구성된 4연숫자음 35종을 포함하여 구성하였다.

2.5.2 단 문

한국어의 자연스런 합성과 인식을 향상을 위해서는 한국어의 운율(Prosody)에 대한 연구가 이루어져야 한다. 이를 위해서는 한국어의 운율에 대한 정보를 추출하고 연구할 수 있는 운율음성DB가 구축 되어야 한다. 그러나 한국어의 특징을 잘 나타낼 수 있는 정형화된 운율 문장은 아직 국내에 설계된 바가 없다. 운율추출을 위한 발성목록의 설계는 문장수준의 음성DB 구축시 고려 해야할 사항이지만, 단어수준 음성DB에서도 운율 추출을 위한 기본적인 데이터를 제공하기 위하여 단문을 포함하였다. 포함된 단문은 국제음성학회 한국어 발음예문인 “바람과 햇님”을 발성목록으로 하였다.

2.5.3 고빈도 2,000어절

고빈도 2,000어절은 PBW를 추출하기 위한 모집단인 고빈도 5,000어절 중 음소중복을 최소화하고, PBW 452어절이 포함되도록 고빈도 순으로 2,000어절을 선정하였다.

III. 음성DB의 구축

음성DB를 구축하기 위해서는 음성시료를 수집하고 이를 편집하여 공통으로 사용할 수 있는 포맷으로 저장하여야 한다. 이러한 과정을 다음에 기술 하였다.

3.1 음성 데이터의 녹음

작성된 발성목록을 대상으로 표준어를 사용하는 서울 지역 일반인 및 어나운서 72명분의 음성을 수집하였다.

음성은 방음 부스에서 Senheizer HMD 224X Headset 마이크를 사용하여 녹음 하였으며, 발생된 데이터는 디지털 오디오 테이프에 저장하였다. 16kHz로 샘플링하고 16Bits로 양자화 하였다. 각 발성목록별 발생횟수는 표 2에 나타내었다.

표 2 발성목록별 발생횟수

발성목록	발성 횟수
단독숫자 및 지시어	일반인 70명 X 4회, 어나운서 2명 X 2회
4연숫자	일반인 70명 X 4회, 어나운서 2명 X 2회
단문	일반인 70명 X 2회, 어나운서 2명 X 2회
PBW	일반인 70명 X 2회
고빈도 2,000어절	어나운서 2명 X 2회

발성화자의 성별, 연령별 분포는 다음 표 3과 같으며 남성이 54.16%, 여성이 45.83%를 차지하고 있다.

표 3 발성화자의 성별, 연령별 분포

	남	여	계
10대	3	1	4
20대	27	25	52
30대	5	6	11
40대	4	1	5
계	39	33	72

3.2 음성 데이터의 편집

A/D하여 컴퓨터의 하드디스크에 저장된 음성 데이터는 그 양이 많으므로 각기 필요없는 부분은 제외하고 사용할 수 있는 부분만으로 다시 편집을 하여야 한다. 그 과정은 우선 양자화된 음성 데이터를 대상으로 자동 끝점 추출 알고리즘[12]을 사용한 음성 데이터 편집 툴을 구현하고 이를 이용하여 필요한 음성 데이터를 찾아내고, 불량 데이터의 제거, 수정 데이터의 삽입, 음성 데이터의 앞뒤에 일정한 길이의 무음구간의 확보 등 전반적인 사항을 수정한다. 이러한 편집과정을 그림 5에 나타내었다.

이와 같은 과정을 거친 각각의 데이터화일을 디렉토리 구조로 저장하여 CD-ROM형태로 제작되었을 때 효율적으로 검색하고 활용할 수 있도록 하였다.

3.3 음성 데이터 화일의 구조

음성 데이터를 보관하는 형태로는 음성파형으로 저장하는 방법과 분석처리를 한 후 파라미터의 형태로 저장하는 방법이 있으며, 후자의 경우에 저장하기 위한 정보량의 압축 및 이용시에 계산량 절감 등의 장점이 있으나 이용할 분석법에 대한 선택의 어려움과 특정 분석법에 의한 이용상의 제약이 있어 배포를 목적으로 하는 경우에는 음성파형의 형태로 저장하는 것이 바람직 하다[4].

음성 데이터를 저장하고 있는 각각의 화일들은 음성 데이터에 대한 정보를 유지하기 위해 헤더를 갖는다. 본문에서 구성한 헤더 구조는 미국의 DARPA-TIMIT

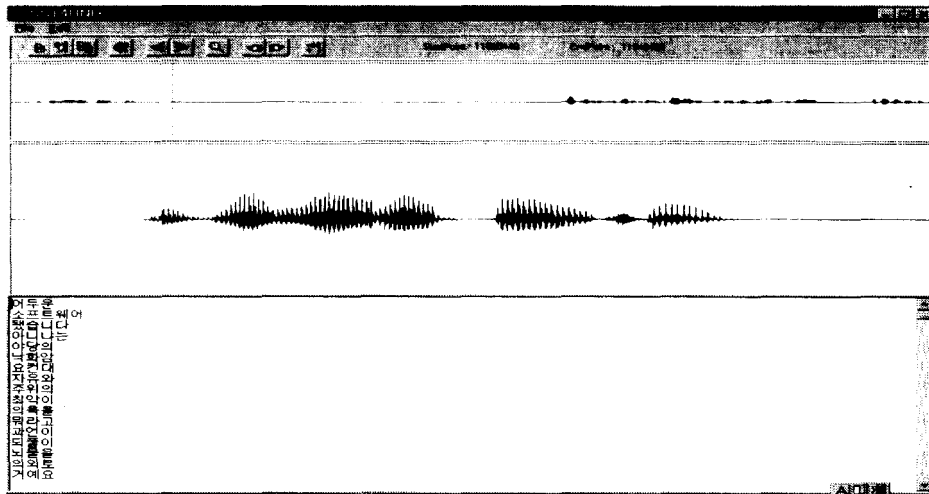


그림 5. 음성 데이터의 편집

Speech Database의 헤더 구조를 참조하여 상호 호환성을 유지하도록 구성 하였다[13]. 헤더의 크기는 1024바이트로 고정되어 있으며, 헤더 첫 부분의 고정영역을 제외하고는 필드의 추가, 삭제, 주석의 첨가 등이 자유롭게 이루어질 수 있도록 되어있다.

3.4 음성DB의 구조

음성 데이터의 수정과정을 거친 후 화일로 저장한다. 일반적으로 화일을 저장하는 방법은 편집된 데이터를 하나의 연속 데이터로 묶고, 필요한 화일만을 찾아낼 수 있는 Dictionary 화일을 만들어 사용하는 방법이 있다. 그러나 이와같은 방법은 데이터를 관리하기가 편하지만 개개의 화일을 복사할 때 번거로움이 따른다.

반대로 음성 데이터 화일 각각에 하나씩 이름을 붙여서 사용하면 화일을 관리하기는 어렵지만 필요한 화일을 복사하거나 이용하기가 용이하며, 각각의 데이터를 디렉토리 구조로 저장함으로써 효율적으로 데이터를 검색하도록 할 수 있다. 이는 CD-ROM화하여 배포하기에도 용이한 방법이다. 따라서 본 논문에서는 발성목록의 종류에 따른 각각의 음성 데이터를 디렉토리 구조로 저장하고 이를 CD-ROM으로 구축하였다[14]. 발성목록별 데이터의 양은 다음 표 4와 같다.

표 4. 음성DB CD-ROM의 데이터량

CD-ROM	예수	데이터량 (MBvte)
단독숫자 및 단문	1	단독숫자 : 277 단문 : 174
4연숫자	1	4연숫자 : 376
PBW1	1	PBW : 592(남성화자 20명분)
PBW2	1	PBW : 506(남성화자 18명분)
PBW3	1	PBW : 549(여성화자 18명분)
PBW4 및 고빈도 어절	1	PBW : 435(여성화자 14명분) 고빈도어절 : 120 (남녀 어나운서 각 1명분)

또한 각 CD-ROM의 DOC 디렉토리에는 음성DB CD-ROM을 사용하기 위한 표 5와 같은 On-line document가

표 5. CD-ROM의 On-line document

화 일 명	내 용
cdiglist.txt	해당 CD-ROM의 발성목록 및 시료번호에 대한 정보
diglist.txt	
narlist.txt	
pbwlist.txt	
frqlist.txt	
readme.doc	음성DB CD-ROM 전반에 관한 정보
spkrinf.txt	발성화자에 대한 정보

들어 있다.

그리고 구축된 음성DB의 실용성을 높이기 위해서는 음소단위 레이블링된 정보가 포함되는 것이 바람직하다. 그러나 국내에서는 아직 음성 데이터를 레이블링하는 작업이 대부분 수작업으로 이루어지고 있으며, 또한 그 기준마저 통일되지 않고 각 연구자의 필요에 따라 작성하여 사용하고 있으므로 기준 및 결과가 공동이용의 목적으로 사용하기에는 부적합하거나 객관적 검증이 이루어지지 못하고 있는 실정이다. 따라서 음성DB의 구축과 병행하여 여러분야에서 사용할 수 있는 한국어 레이블링 기준안을 마련하여 실제 데이터를 레이블링 하면서 검증 보완하는 연구와, 작성된 기준을 토대로 자동 레이블링 시스템을 구축하는 연구를 진행하고 있다[15].

IV. 결 론

공동이용을 위한 음성DB를 구축하기 위해서는 그 발성목록이 한국어의 다양한 음운환경을 포함하고 있으며, 일정한 도메인의 텍스트에 종속되지 않는 것이 바람직하다. 본 논문에서는 이를 위하여 120만 여 어절의 텍스트 코퍼스에서 3음절이상의 고빈도 5,000어절을 선정하고 여기에서 PBW를 추출하고 제시하였다. 추출된 PBW set은 다음과 같은 조건을 만족한다. 첫째, 모집단에 나타난

모든 음운환경을 포함한다. 둘째, 최소한의 어절로 구성 되어 있다. 셋째, PBW set을 구성하고 있는 음운환경들 간의 확률분포가 최대한 고르게 분포한다.

그의 발성목록으로 임의단위 합성 시스템을 위한 고반도 2,000어절을 작성하고 숫자음으로는 단독숫자음 및 4연숫자음 76종을 작성하고, 단문은 1종을 선정하였다.

이와같이 구성된 발성목록을 표준어를 사용하는 72명의 일반인 및 어나운서를 대상으로 음성시료를 수집하고 이를 양자화하여 편집하고 CD-ROM에 저장하여 공동으로 이용이 가능한 형태로 구축하였다.

앞으로의 연구방향은 공동이용을 위한 한국어 음성DB의 내용을 확장하여 문장수준의 음성DB를 구축하는 것이다. 이를 위해서는 한국어의 읽기규칙을 보다 확장시켜 어절과 어절간의 다양한 음운변화와 운율정보를 포함할 수 있도록 확장하고, 이를 통해 균형잡힌 텍스트 코퍼스로부터 음운균형 문장세트를 추출하고 이를 발성목록으로 하여 문장수준의 음성DB를 구축하여야 할 것이다.

감사의 글

본 연구는 과기처의 연구비에 의해 수행된 것이다. 본 연구수행을 지원한 시스템공학연구소의 국어공학센터 박동인 센터장, 한국과학기술원의 최가선 박사께 감사드립니다.

참 고 문 헌

1. 이용주, 김봉완 외, "해의 음성 DB 구축 동향," 한국음향학회 제12회 음성통신 및 신호처리 워크샵 논문집, 1995. 6.
2. 이용주, "음성데이터베이스의 현황 및 과제," 한국음향학회 제13회 음성통신 및 신호처리 워크샵 논문집, 1996. 8.
3. 이용주, "한국어 음성언어정보처리와 음성 데이터베이스," 한국어정보처리 소식 제2권 특별기고, 1994. 10.
4. 이용주, 정유현 외, 보급형 음성 데이터베이스 구축에 관한 연구, 1차년도 최종보고서, 과학기술처, 1992. 7.
5. 정유현, 최준혁 외, "공통 음성 데이터베이스 구축을 위한 사전 조사연구," 전자공학회 하계 학술대회, 1992. 6.
6. K. Shikano, "Phonetically balanced word list based on information entropy," proceedings of Acoustical society of japan, 1984.
7. Yeonja Lim, et al, "Implementation of the POW algorithm for Speech Database," ICASSP 95, Detroit, 1995.
8. Hayamizu, "Generation of VCV/CVC balanced word sets for speech data base," Bulletin of Electrotechnical Laboratory, 1986.
9. K. Iso, et al, "Design of a Japanese Sentence List for a Speech Database," Proc. ASJ., 2-2-19, March, 1988.
10. Jianhua Lin, "Divergence Measure Based on the Shannon Entropy," IEEE Transactions on information theory, Vol. 37, No. 1, 1991.
11. 최운천, 지민제, 이용주, "문장음성 변환시스템 글소리II를 위한 읽기 규칙," 1992년도 제4회 한글 및 한국어 정보처리 학술발표 논문집, 92. 10.

12. L. R. Rabiner, et al, Digital Processing of Speech Signals Prentice Hall, Inc., Englewood Cliffs, 1978.
13. NIST:Speech Corpora Produced on CD-ROM Media by The National Institute of Standards and Technology (NIST), April, 1991.
14. 김봉완, 이용주 외, "공동이용을 위한 단어음성DB의 구축 및 PBS설계에 관한 검토," 한국음향학회 제13회 음성통신 및 신호처리 워크샵 논문집, 1996. 8.
15. 김종진, 이용주 외, "한국어 음성DB 구축을 위한 한국어 레이블링 기준에 관한 연구," 한국음향학회 제13회 음성통신 및 신호처리 워크샵 논문집, 1996. 8.

▲김 봉 완(Bong-Wan Kim)

1995년 2월:원광대학교 컴퓨터공학과(공학사)
 1997년 2월:원광대학교 대학원 컴퓨터공학과(공학석사)
 1997년 3월~현재:원광대학교 대학원 컴퓨터공학과 박사과정
 ※주관심분야: 음성인식, 음성합성

▲김 종 진(Jong-Jin Kim)

1995년 2월:원광대학교 컴퓨터공학과(공학사)
 1997년 2월:원광대학교 대학원 컴퓨터공학과(공학석사)
 1997년 3월~현재:원광대학교 대학원 컴퓨터공학과 박사과정
 ※주관심분야: 음성인식, 음성합성

▲김 선 태(Sun-Tae Kim)

1996년 2월:원광대학교 컴퓨터공학과(공학사)
 1996년 3월~현재:원광대학교 대학원 컴퓨터공학과 석사과정
 ※주관심분야: 음성인식, 음성합성

▲이 응 주(Yong-Ju Lee)

1976년 2월:고려대학교 전자공학과(공학사)
 1985년 8월:고려대학교 대학원 전자공학과(공학석사)
 1992년 8월:고려대학교 대학원 전자공학과(공학박사)
 1976년 3월~1980년 7월:공군 통신전자장교
 1985년 8월~1986년 1월:일본 토호쿠대학 응용정보학연구센터 연구원
 1980년 8월~1994년 2월:한국전자통신연구소 실장, 책임연구원
 1994년 3월~현재:원광대학교 컴퓨터공학과 교수
 ※주관심분야: 음성언어정보처리전반, 멀티미디어, HCI