한국어 음성DB의 효율적 확보 방안
(다양한 환경의 음성DB 구축 방안)

오 영 환
한국과학기술원 전산학과

A Plan for Noisy Speech DB Construction

Oh, Yung Hwan
Dept. of Computer Science, Korea Advanced Institute of Science and Technology

1. 서 론

다양한 음향 환경에 강인한 음성인식기의 개발을 위해 잡음 처리 기술에 대한 많은 연구가 진행되어 왔으나, 그 동안의 연구 성과에도 불구하고 아직도 해결해야 할 많은 문제가 많이 남아 있다. 미국, 일본, 유럽의 경우 실험실 환경에서 높은 성능을 보이는 음성인식기를 개발하여 이의 성능 향상뿐만 아니라, 실제 현장에 적용하기 위해 실생활에 나타나는 여러 문제를 해결하는 연구를 수행하고 있다. 국내의 경우 음성처리 기술이 선진 각국에 비해 뒤떨어져, 음성인식기의 실용화를 위한 잡음처리 기술에 대한 연구도 미약한 형편이나, 점차 이에 대한 기술 개발의 필요성이 인식되고 있다

잡음 환경과 전송선의 특성에 대한 연구를 위해서는 먼저 다양한 환경의 음성 DB의 확보가 선결되어야 하며, 여러 환경과 잡음음성과 기술들을 평가하기 위한 표준적인 잡음음성과 잡음의 데이터베이스가 선정이 필요하다. 잡음음성의 인식에서 가장 적합한 알고리즘은 음향 환경과 인식 대상의 특성을 파악하고 이에 종속적인 방법을 사용하는 것이다. 따라서 잡음의 성질에 따라 최적의 잡음처리 기술이 다르므로, 잡음처리 기술의 개발과 객관적인 성능 평가 및 진단을 위해서는 표준적인 잡음 음성 데이터베이스가 필요하다

2. 전화음성 DB의 구축

음성인식 실용화 분야에서 가장 주목 받는 부분은 전화 음성 인식이다. 전화음성은 전송선의 채널 잡음, 대역폭 제한, 다양한 전화 송화기의 특성, 예측할 수 없는 많은 화자수, 주변잡음 등에 의해 광범위한 활용이 제한되고 있다

외국의 경우 대어휘 연속음성 DB인 TIMIT을 기반으로 전화음성 DB인 NTIMIT, 부전전화 음성 DB인 CTIMIT 등을 만들어 기초 연구 및 실제 시스템 개발에 효과적으로 활용하고 있는 실정이다

전화음성의 확보는 일반적인 음성 수집에 비해 비교적 간단한 장비로 수집이 가능하며 다수 화자의 자연스러운 음성을 수집할 수 있다. 그러나, 효과적인 인식기의 개발에 부합된 음성을 수집하기 위해서는 각 음향 환경과 전화회선의 특성에 따른 음향학적인 영향을 분석할 수 있는 다양한 환경의 자료가 요구된다.

3. 다양한 환경 잡음과 채널 잡음 DB의 구축

사무실 환경의 타자기, 프린터, 컴퓨터의 디스크와 팬, 전화 벨소리 등과 자동차 환경에서의 엔진과 냉각팬, 도로와의 마찰음, 바람소리 등의 잡음은 음성과 상관관계가 없이 가산적으로 첨가되는 가산잡음으로 모델링 된다. 가산잡음이외에도 음성의 특성은 음성인식기가 설치되어 있는 방의 특성에 따른 반향, 음성 입력 기기인 마이크의 위치와 종류에 따른 주파수 응답특성의 변화 등에 영향을 받게 된다. 따라서 환경잡음은 음성자료와 개별적으로 수집되어도 원 음성에 크기를 조절하여 가산함으로써 이용이 가능한 반면에 채널 잡음의 경우 전화음성과 마찬가지로 채널의 필터 특성을 분석하기 위해는 다양한 음운환경의 음성과 같이 수집되어야 한다.

4. Rombard 음성 DB의 구축

잡음환경에서 사용자는 rombard 효과에 의해 평소와 발성방식과 다르게 발성 되며, 따라서 잡음환경하의 음성은 단순히 음성에 잡음이 첨가되는 것이 아니라 음성의 여러 특성이 변하게 된다. 이러한 rombard 효과에 의해 발성된 음성은 조용한 환경에서 발성된 음성과 특성이 다르므로, 단순한 잡음 제거는 인식을 향상이 어렵다

일반적으로 rombard 효과는 잡음의 종류 및 크기에 따라 다르므로 체계적인 음성 수집이 용이하지 않고, 경우가 다양하여 DB의 구축에 어려움이 많다. 이를 고려한 음성 수집을 위해서는 잡음의 크기의 종류를 조절하여 발성자에게 헤드폰을 통해서 들려주고, 어와 같이 모의된 잡음 환경에서 발성 된 음성을 모으는 것도 한 방법이 될 것이다

5. 결 론

현재까지는 음성인식 시스템의 개발과 성능 평가에 있어서 표준적인 음성 DB의 부재로 인하여, 각자가 보유하고 있는 평가 자료를 사용하였다. 특히, 다양한 환경의 잡음이나 잡음음성 자료가 부족하여, 견고한 인식 기술 개발과 객관적인 평가 및 성능의 진단 정보를 얻을 수 없었다 따라서, 추후 다양한 환경에서의 표준적인 음성 DB의 구축은 음성인식 기술의 실용화를 위해 반드시 해결해야 할 과제이다.

음성 인식용 DB 구축 방안

°김 순협
광운대학교 컴퓨터 공학과

The Approach of Speech Database for Recognition

Soon-Hyob Kim
Kwang-Woon Univ. Computer Engineering.

사 용

컴퓨터 및 통신기기의 발달에 따라 인간과 기계와의 정보 교환이 빈번하게 발생하고 있어 기존의 방식보다도 편리한 새로운 인터페이스 수단의 필요성이 고조되고 있다. 따라서 인간의 정보 교환 수단 중에서 가장 자연스럽고 사용하기 편리한 음성을 매체로 한 인터페이스 수단으로 이용하기 위한 음성 인식,합성의 음성 정보처리 연구가 활발히 진행되고 있다. 이를 위하여 연구 초기에 반드시 확보되어야 하는 연구환경이 많은 사람의 다양한 음성 데이타이며, 또 개발된 시스템 및 알고리즘의 객관적인 성능 평가를 하기 위한 음성 데이타도 필요하다. 이외 다른 관점에서 음성 인식을 위한 연구 및 평가에 필요한 음성 데이타를 구축하는 일은 매우 중요하다고 보았다.

1. 음성 데이타 베이스 개요

음성 인식에 관한 음성정보처리 연구는 H/W 및 S/W 등 관련 기술의 진보에 따라 환경 이휘에 대한 인식 시스템이 실용화되고 있으며, 현재 임의 어휘를 대상으로 하는 인식 시스템의 개발에는 음소 단위에 의한 인식 기술 개발이 필수적이라 한다. 인접음성중의 음소는 발성시에 의한 영향(개인차)은 물론이고 전후에 발성되는 음소의 영향(조음결합)에 의해서도 크게 변화하기 때문에 기술 개발에 어려움이 많다. 따라서 이러한 개인차 및 조음결합의 현상을 파악하는 연구를 수행하기 위해서는 다수의 사람이 발성한 대량의 다종 음성 데이타가 필요하다. 또 음성 인식 시스템이 점차 실용화 되어감에 따라 시스템의 객관적인 성능 평가에 이용할 수 있는 표준 음성 데이타도 필요하다. 이와같은 음성 데이타를 수집하여 음성 연구시 필요로하는 각종의 음성(단어,음절,음소 등) 을 사용자의 이용할 수 있도록 구축한 데이타 베이스를 음성 데이타 베이스라 한다. 국내의 경우 지금까지 음성 인식 연구에 필요한 음성 데이타를 각 연구자가 각자 부문별으로 제작하여 사용하고 있으며, 이를 외부에 공개하고 있지는 않다. 따라서, 데이타의 양 및 이용 형태가 제한되고, 동시에 각 연구자가 발표한 인식 시스템의 성능 혹은 분석 방식의 평가도 각 연구자의 데이타에 의존하고 있기 때문에 객관적인 평가가 이루어지지 않고 있는 실정이다.

2. 음성 데이타 베이스 요구 사항

이상적인 음성 데이타 베이스는 어휘 수나 화자, 녹음환경, 발음 형태등에서 갖추어야 할 요건들이 있다. 특히 대용량 화자독립 시스템이나 자동 통역 시스템을 위한 음성 데이타 베이스는 더욱 엄격한 조건들을 요구한다. 먼저 이러한 데이타 베이스를 구축할 때 요구되는 것은 기술적인 면보다도 많은 사람에 다수 발생한 다양한 음성을 얼마나 많이 확보하느냐가 중요할 것이다. 음성 데이타 베이스를 구축시 고려해야 할 기본적인 요구사항은 다음과 같을 것이다.

(1) 발생 내용
(2) 발생자
(3) 데이타량
(4) 녹음 조건
(5) 기록매체
(6) 음성 데이타 보관 형태
(7) 관리방법

특히 객관성을 위해서는 (1) 항의 발생 내용이 중요한 것이며, 단음절, 단어, 연속단어, 문장등 다양한 형태의 데이타를 구축해야 인식 시스템에서 인식에 기본 단위로 설정하여 인식을 실행할 수 있을 것이다. (2) 항의 발생자 또한 다양한 연령별,직업별,출신 지역별, 하위별, 표준어 및 방언을 고려하여야 할 것이다. (3) 항의 데이타량은 많으면 많을수록 좋겠으나 이러 데이터 양으로 통계화와 및 불특정화자의 2회 이상의 발성을 고려해야 할 것이다. (4) 항의 녹음조건은 무향실이나 방음실 또는 자동차내 같음이 있다. 방송 이러환경에서 녹음을 하면 좋을 것이다. (6) 항의 음성 데이타 보관 형태는 음성 파형 혹은 분석처리된 한 특정 파라메터의 형으로 저장되는 방법이 있으며 주어진 경우 저장하기 위한 정보량의 입

축 및 이용시에 계산량의 절감 등의 장점이 있으나, 이용할 분석법에 대한 선택의 어려움과 특정 분석법에 의한 이용상의 제약이 있어 데모를 목적으로하는 경우에는 음성 파형의 형태로 사상하는 것이 바람직하다.

3. 국외 연구 동향

최근 미국, 일본및 유럽등의 음성 데이타 베이스 구축 사례를 보면 다음과 같은 측면으로 연구 방향이 변화하고 있다. 첫째로 어휘의 수와 태스크에서의 변화이다. 80년대에 구축된 음성 데이타 베이스들은 천 단위의 어휘의 크기와 특정 태스크에 집중되어 있었던 반면 최근에 구축된 것들은 어휘의 수가 수만에서 수십만에 이르는 텍스트를 이용하여, 특정 태스크에만 집중되지 않는다는 점이다. 둘째로 상용 시스템 개발을 위하여 많이 이용될 전화음성 데이타베이스의개발을 들 수 있다. 세째로, 실제 환경 조건에서의 음성 데이타 베이스 구축이다. 네째로 다국어 음성 데이타 베이스 구축이다. 여러 언어간의 음성 기술 개발과 성능 비교의 목적을 위해 개발하고 있다.

4. 맺는말

외국의 경우를 보아도 알 수 있듯이 음성인식 시스템을 위하여 필수적인 음성 데이타베이스의 구축이 여러 환경의 다양한 각도에서 전개되고 있으며 대용량 및 전화음성 뿐아니라 자연스러운 대화체 문장에 이르기까지 그 범위가 확대되고 있다. 국내에는 아직 음성인식을 수행하는 많은 연구기관이나 대학들이 있지만 표준화된 객관적인 음성 데이타 베이스가 없기 때문에 객관적인 평가를 할 수 없을 뿐더러 음성 인식 기술의 상호 교류도 힘들며 각기 제한된 연구 환경에서 나름대로의 음성 데이타를 만들어 사용함으로 중복되는 작업이 반복되고 있다. 이에 여러 환경에서도 음성을 인식할 수 있는 기본적인 음성 데이타 베이스가 구축된다면 많은 인식 시스템의 성능 향상을 기대할 수 있을 것이다.

한국어 음성 데이타베이스의 효율적 확보방안
- 음성 입출력 평가기술과 음성DB -

김 경 태
한남대학교 정보통신공학과
전화 : 042 629 7574 팩스: 042 629 7843

Speech Assessment Technologies & Speech Database

Kyung Tae KIM
(Dept. of Informations and telecommunications, Han Nam University)

● 목적

Interactive한 음성입출력 시스템의 성능평가를 위하여 데이타베이스의 개발이 요구되고 있다. 이러한 데이타베이스이 개발로 인하여, 상업용 및 연구중인 음성입출력 시스템에 적용하여 그 성능평가의 결과를 연구자 혹은 개발자에게 피드백시켜 객관적인 시스템의 성능을 알게함으로서 국내의 음성처리 기술의 향상에 기여한다.

● 중요성 및 필요성

성능평가 기술과 데이타베이스의 개발이 되면, 개발자의 면에서는 시스템의 각 요소별 성능평가를 통해 성능향상의 척도를 제공받을 수 있고, 시스템 사용자의 입장에서는 여러 제품들을 체계적으로 비교할 수 있어 효율적인 선택의 척도를 제공할 수 있게 된다.

또한, 기술평가에 관한 연구와 데이타베이스의 개발은 일반 기업체에서는 시간과 노력 등의 면에서 수행하기가 어렵고 연구가 되더라도 공공성의 문제와 객관성의 문제는 물론이고 연구 자체도 여러 기관이 협력해야 할 사항이 많이 있기 때문에 공공연구기관이 담당해야 할 필요가 있다.

마지막으로 시장개방의 가속화와 함께 우리말에 대한 연구를 외국에서 연구/개발하여 국내에 들어올 때를 대비해서 우리말시스템에 대한 우리의 평가기준과 음성데이타베이스를 마련하고 있어야 한다.

● 입력(인식) 기술의 성능평가

인식시스템을 가장 객관적으로 평가하는 방법은 모든 장소에서 모든 계층의 사람들이 모든 경우의 상황(스트레스, 물리적 상황, 정신적 상황 등)에 의한 공통의 평가용 데이타베이스로서 직접 테스트해서 인식결과를 서로 비교하면 될 것이다. 그러나 이러한 방법은 시간과 경제적인 여러 문제로 실현이 어렵다. 따라서 실제의 여러 상황에 직접 테스트를 하는 것이 아니라, 실험실에서 간편하게 테스트를 하면서도 객관적이고, 진단적이며, 예측적인 평가결과를 도출할 수 있는 평가법을 마련해야 한다. 이렇게 하려면 가능한한 실제의 상황을 그대로 시뮬레이션한 평가 절차와 측정된 인식결과로서 진단적이고 예측적인 평가를 할 수 있는 기술들이 개발되어야한다. *만약 공통의 평가용 데이타베이스가 구축되어있다면 적절한 평가항목을 선정해서 항목별 인식결과를 구하고, 인식결과와 평가항목간의 상관관계를 해석함으로서 전반적인 인식시스템의 성능을 진단하고 예측한다.*

● 음성합성의 성능평가

규칙에 의한 음성 합성기술의 평가는 합성기술의 완성도에 관한 진단적 평가와 합성시스템의 이용목적, 이용자, 이용환경을 포함한 종합적 평가(assessment)로 나누어 생각할 수 있다. 진단적 평가는 음성합성기술을 개발 연구하는 기술자나 연구자 자신이 개발과정의 피드백 요소로서 각 개발 단계별로 개별적으로 수행하는 경우가 많지만 최근에는 이를 계통적으로 규범화하려는 움직임을 보이고 있다. 또한, 규칙합성기술이 맨머신 인터페이스로서 폭넓게 이용되기 위해서는 이용자인 인간의 음성언어 표출 및 수용과정을 포함한 토탈 시스템의 구성요소로서 합성기술을 바라보면서 종합적으로 평가하는 종합적 평가방법의 확립이 시도되고 있다.

● 음성 데이타베이스 및 관리시스템

음성 정보처리를 연구하기 위하여 다양한 종류(성별, 연령, 방언, 인원, 발성횟수)의 음성 데이타베이스 구축의 필요성이 날로 더해가고 있다. 국내의 경우 지금까지는 음성 연구를 하고자 하는 각자가 필요에 따라 음성 데이타를 녹음, 이용, 보관하여 왔다. 그러나 음성 연구가 더욱 진전됨에 따라 처리하고자 하는 데이타량의 증가가 요구되고, 음성 정보처리 시스템의 연구 개발을 위하여 분석, 합성, 인식의 각종 방법을 비교, 평가할 수 있는 공통 음성 데이타가 요구되고 있다. 이렇게 여러 사람이 이용 가능한 각종의 음성 데이타를 수록, 보관, 공개 하는 일은 매우 중요한 일이며, 연구 개발 과정과 성능평가의 차원에서도 꼭 필요한 일이다. 음성 데이타베이스를 구축할 경우, 그 이용 분야로서 음성 인식을 비롯하여 음성 합성, 화자 인식 분야로 생각할 수 있다. *각 분야별로 독립적인 음성 데이타베이스가 있으면 좋겠지만, 경우에 따라서는 공통 음성 데이타베이스를 분야별로 이용할 수 있다.*

● 결론

음성 입출력 기술인 음성인식 시스템과 음성합성 시스템의 성능평가와 연구를 위한 음성 데이타베이스에 대하여 생각했다. 다양한 분석기법과 알고리즘, 그리고 서로 다른 음성데이타를 사용하여 구성된 시스템을 객관적으로 평가한다는 것은 매우 어려운 일이다. 평가의 최종 목적인 향상된 시스템의 개발이라는 측면에서 볼 때, 각 조건에 대해 얻어지는 정보로서 성능향상을 꾀할 수 있는 방법이 필요하게 된다. 이를 위하여 평가용과 연구용의 음성데이타베이스가 필요하게 된다. 음성언어에 관한 다른 연구도 그러하듯이 인간의 언어인지 및 생성과 관련된 문제들의 연구는 인접 학문 간의 학제적 공동 노력이 필요하며 평가법과 같은 표준화와 데이타베이스에 관련된 연구는 기관 간의 협조체제의 구축과 함께 나아가서는 국제적인 협력에도 관심이 모아져야 할 것이다.

음성 DB의 효율적 확보방안

이호영(부산수산대)

음성 합성기와 음성 인식기의 개발을 위해 국내 대학, 연구소, 기업 등에서는 PBW(Phonetically Balanced Word List), 숫자음 DB, 전화음성 DB, 문장음성 DB 등 낭독체 음성 DB들을 구축하고 있으나 대화체 음성 및 운율 DB는 구축하고 있지 않다. 자연스런 대화체 발화를 합성해 낼 수 있는 음성 합성기를 개발하고, 무한대 어휘의 대화체 발화를 인식할 수 있는 음성 인식기를 개발하기 위해서는 정교하게 제작된 방대한 양의 대화체 음성 및 운율 DB를 필수적으로 갖춰야 하므로 앞으로는 대화체 음성 및 운율 DB의 구축에 더욱 많은 관심을 기울여야 한다.

대화체 음성 및 운율 DB의 구축 작업에 사용될 음성 자료로는 사람들의 자연스런 대화나 방송 대담을 녹음한 것이 이상적일 것이다. 그러나 이 자료는 음성 합성기의 개발을 위한 음성 및 운율 데이타의 수집에 유용하게 사용될 수 있지만 음성 인식기의 개발을 위한 시스템 학습 자료로는 부적절할 수도 있다. 아직은 무한대 어휘 대화체 음성 인식기를 개발하는 기술이 확보되어 있지 않기 때문이다. 단시간에 시연할 수 있는 음성 인식기를 개발하려면 녹음할 대화의 주제를 크게 한정해야 한다. 음성 인식기의 태스크(task)와 관련된 대화 주제를 호텔이나 열차 예약, 국제 학회 참가 등록, 무역 상담, 관광 안내 등이 있다.

대화체 음성 자료의 수집과 더불어 준비해야 할 일은 대화체 음성 자료의 분절 방법과 표기법을 표준화하는 것이다. 이를 위해서는 DB 구축 참여자들이 워크샵 등을 통해 분절 방법과 표기법을 통일하는 작업을 해야 한다.

대화체 음성 및 운율 DB는 가능한 한 많이 구축해야 할 뿐만 아니라 최대한 정밀하게 제작해야 한다. 이를 위해서는 충분한 수의 전문 인력을 양성해야 하고, 충분한 예산을 확보해야 한다. 그리고 인력과 예산, 그리고 시간을 절약하기 위해서는 DB의 구축 초기 단계에서부터 치밀하게 기획해야 한다.

음성 DB용 PBW에 관한 검토

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

Some considerations for construction of PBW set

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

요약

음성연구에 있어서 음성데이타는 필수적이다. 본고에서는 음성 데이타베이스를 구축하는 데 있어서 다양한 음운현상을 포함하되 음소열간 중복이 가장 적고 고른 확률분포를 갖는 최소 단어들의 집합인 PBW(Phonetically Ballanced Words)를 선정하기 위한 몇몇 사항을 살펴보고, 필자들이 현재 구축하고 있는 PBW의 예를 소개한다.

1. 서론

음성 인식 및 합성 시스템의 개발등 음성연구에 있어서 다종다양한 음운현상을 포함한 음성 데이타베이스의 구축은 중요한 과제의 하나로서, 많은 시간과 노력이 요구된다. 개별적 음성 데이타베이스의 구축에 따른 중복 투자를 줄이고 분석, 합성, 인식의 각종 알고리즘을 적절히 비교 평가하기 위해서도 공통의 음성 데이타 베이스는 필수적이다.

이러한 공통 음성 데이타베이스는 발생가능한 모든 음운현상을 포함하며, 발성대상 단어나 문장도 특정 태스크에 집중되지 않는 것이 이상적이다.

본고에서는 음성의 분석, 인식 및 합성을 위해 음성DB를 구축함에 있어, 적은 단어(또는 어절)에 다양한 음운환경을 포함시킬 수 있는 단어세트인 PBW를 선정하기 위한 몇몇 고려사항, PBW의 선정 절차 및 추출된 PBW의 특성에 대하여 기술한다.

2. PBW (Phonetically Balanced Words)의 선정

2.1 PBW의 정의

음운밸런스가 취해진 상태란 음운의 출현빈도가 같은 상태를 말한다. 이러한 상태는 음운을 확률사상으로 했을때 엔트로피가 최대인 상태를 말한다. 음소열의 출현확률을 Pi 라고 할때 엔트로피 H는 다음식(1)에 의해 구할 수 있다.

$$H = \sum_{i=1}^{N} P_i \log P_i \qquad (1)$$

PBW는 발생가능한 모든 음운현상을 포함하며, 각 음운들이 고른 확률분포를 갖는 최소단어들의 집합이라고 할 수 있다.

2.2 PBW추출을 위한 준비 절차

본 연구에서는 PBW를 추출하기 위한 모집단으로서 고려대에서 구축한 1,288,000여 어절의 텍스트 코퍼스에서 3음절 이상의 고빈도어를 추출하여 발성상 부적절한 어절, 음성 데이타베이스의 대상으로서 부적절한 어절, 동일 위치에 50%이상의 음소가 중복되는 어절등을 삭제하고 최종적으로 고빈도 5,000어절을 선정하였다. 이러한 처리절차를 그림 1.에 나타 내었다.

가) 텍스트 코퍼스에서 3음절이상의 고빈도 어절의 추출

본 연구에서 3음절이상의 고빈도어를 PBW추출의 대상으로 선정한 이유는 텍스트 코퍼스에서 3음절 이상의 어절이 차지하는 비율이 87%에 이르며 가급적이면 다양한 음소를 포함하기 위해서이다.

본 연구에서 대상으로 삼은 텍스트 코퍼스는 다음과 같은 자료로 부터 추출된 것이며 총 어절수는 1,288,000여 어절이다.

1) 신문기사

1.1)조선일보(1993년) - 77500어절

1.2)동아일보(1993년) - 104000어절

1.3)한겨레신문(1993년) - 126000어절
1.4)이규태 코너 - 50778어절
2) 소설 (4종) - 227540어절
3) 소설 이외의 수필, 기행문등 - 477268어절
4) 구어 자료
4.1) 세미나 기록 자료 - 20000어절
4.2) 대담 및 인터뷰 - 20500어절

나) 발성상 부적절한 어절 삭제

1) "클린턴", "이제필이", "엘친대통령이"와 같이 인명이 포함된 어절은 삭제한다.
2) "루슬란"과 같이 외래어나, 의미가 난해한 어절은 삭제한다.
3) "지섭은", "제린은", "해숙은"등과 같이 소설등의 자료에서 포함되거나, 자료의 성격상 포함된 전문용어등의 어절은 삭제한다.
4) "A가", "1980년" 등과 같이 영문자, 숫자 또는 특수 기호가 삽입된 어절은 삭제한다.
5) "그라믄", "빌어묵을", "워할라꼬"등과 같은 방언은 삭제한다.
6) 기타 접미사, 그밖의 발성상 부적절한 어절은 제외한다.

다) 읽기 규칙을 적용하여 소리나는 대로 표기한다.

이상에서 추출된 3음절 이상의 고빈도 어절들을 글자 음운 변환거에 의해 소리나는 대로의 형태로 자동 변환하고 잘못 변환된 어절은 직접 수정을 통해 올바르게 표기한다.

라) 발음이 동일한 어절이나, 동일 위치에 음소가 50%이상 중복되는 어절의 삭제

"기운이 [기우니]", "기우니 [기우니]"등과 같이 서로 다르게 표기되지만 동일하게 발성되는 어절은 하나의 어절만 취하고 나머지는 삭제한다.

그리고 본 연구에서 PBW추출의 대상으로 삼은 것은 어절단위이므로 단지 조사나 접미사만 다르면서 고빈도어절중에 포함된 어절들이 많다. 이러한 어절들을 배제하기 위하여 음소변환을 통하여 동일한 위치에 50%이상의 음소가 중복되는 어절은 가장 빈도수가 높은 어절을 취하고 나머지는 삭제한다. 예를 들면 "언어가 [ㅓㄴㅓ가], 언어는 [ㅓㄴㅓ는], 언어를 [ㅓㄴㅓ를], 언어의 [ㅓㄴㅓㅢ]", "것이가 [거시기], 것이다 [거시다], 것이며 [거시며], 것이요 [거시ㅛ], 것인가 [거신가]" 와 같은 어절들이다.

마) 최종 모집단 선정

이상과 같은 처리를 거쳐 3음절 이상을 대상으로 하여 고빈도 순으로 최종 모집단 5000어절을 선정 하였으며 모집단에는 텍스트 코퍼스에서 3174의 빈도를 갖는 "그러나" 에서 부터 8의 빈도를 갖는 "과시했다"까지가 포함되었다.

사) 대상 음소및 2음소열의 규정

대상 음소로 자음은 19개, 모음은 21개로 구성하고 종성의 자음에 /ㅅ/, /ㅊ/, /ㅌ/등은 모두 대표자음으로 모아 다루었으며, 반모음은 후속모음과 합하여 하나의 음소로 취급하였다. (예, / ㅑ/, / ㅕ/, /ㅘ/, /ㅙ/등)

대상 2음소열은 위와 같이 분류된 음소의 쌍으로 구성하며, 이때 어두나 어말의 경우에는 공백소와 해당음소가 쌍을 이루는 형태를 2음소열의 종류에 포함하였다. (즉, Ba는 a가 어두에 오는 경우, aB는 a가 어말에 오는 경우를 말한다.)

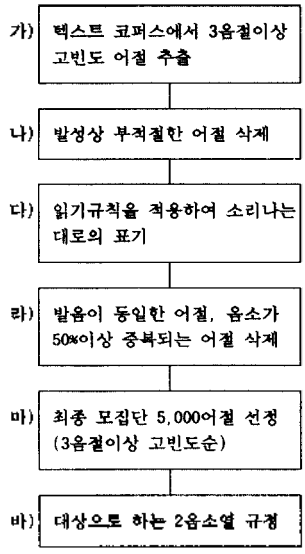

그림1. PBW추출을 위한 준비 절차

2.3 PBW의 선정

2.3.1 PBW의 선정절차

PBW를 선정하기 위한 과정을 그림2.에 나타내었다.

가) 모집단에서 1회밖에 나오지 않은 2음소열을 포함한 어절을 모두 PBW set에 포함시킨다.

본 연구에서는 고빈도 5000어절중에서 1회밖에 출현하지 않은 2음소열을 가지는 어절로 95개의 어절이 추출되었다.

나) 2음소열의 종류가 최대가 되도록 하는 어절을 고른다.

즉 현재 PBW set에 없는 2음소열을 가장 많이 가지고 있는 어절을 선택하여 PBW set에 추가한다.

다) 나)의 경우에, 그 대상이 복수 어절일 경우 엔트로피를 최대화하는 어절을 선택한다.

라) 나), 다)의 단계를 새로운 2음소열이 없을 때까지, 즉 모집단의 2음소열의 종류와 PBW set의 2음소열의 종류가 같아질때까지 반복한다. 나), 다) 및 라)의 과정을 통하여 본 연구에서 추출된 어절은 총 238어절이다.

마) 모집단의 나머지 어절들을 하나씩 PBW set에 추가해 보면서 추가되었을때 엔트로피를 최대화하는 어절을 선택한다. 마)의 과정을 통하여 총 167 어절이 추가 되었다.

바) 그 단어를 삭제해도 2음소열의 종류가 줄지않는 어절(즉, 2음소열이 중복되어 있는 어절)을 삭제하면서 아울러 삭제할 경우에 엔트로피를 최대화 하는 어절부터 삭제해 간다. 가) ~ 바)의 과정을 통하여 추출된 총 500어절을 바)의 과정을 통하여 48어절을 삭제하여 최종적으로 PBW 452어절을 추출하였으며, 최종 엔트로피는 6.157899이다.

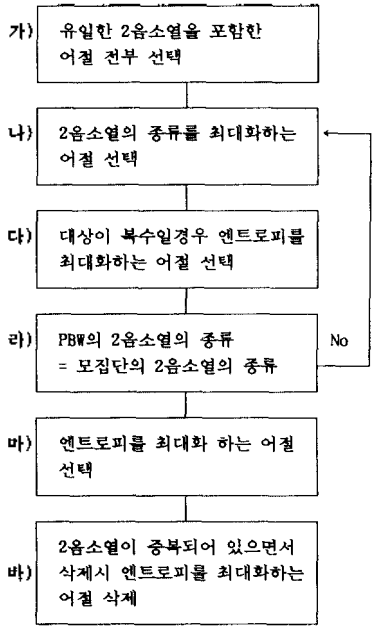

그림2. PBW 선정 과정

2.3.2 추출된 PBW

이상의 과정을 통하여 3음절 이상의 고빈도 5000어절의 모집단에서 PBW 452어절을 추출하였으며, 추출하는 과정의 엔트로피 변화를 그림3.에, 고빈도 5000어절과 PBW set에서의 2음소열의 형태별 빈도수를 표1.에 나타내었다.

표1. 고빈도 5000어절 및 PBW set에서의 2음소열의 형태별 빈도수

		고빈도 5000 어절 (%)	PBW set (%)
B	V	1004 (2.39)	149 (4.25)
V	B	2720 (6.47)	313 (8.93)
V	V	1000 (2.38)	248 (7.08)
V	C	13066 (31.08)	975 (27.83)
C	V	14782 (35.16)	1139 (32.50)
B	C	3996 (9.50)	303 (8.65)
C	B	2280 (5.42)	139 (3.96)
C	C	3195 (7.60)	238 (6.79)
총음소열수		42043	3504

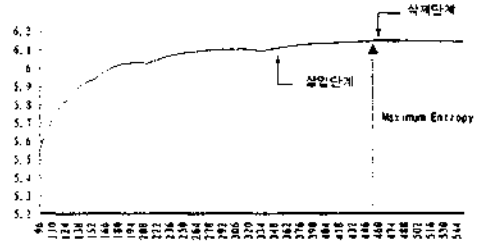

그림3. PBW 추출과정의 엔트로피 변화

3. 결과의 검토

모집단 5000어절과 PBW set에는 모두 서로다른 842종류의 2음소열을 포함하고 있다. 모집단에서는 VC, CV, VV, CC가 각각 31.08%, 35.16%, 2.38%, 7.60%의 분포를 보이고 있으나 PBW에서는 27.83%, 32.50%, 7.08%, 6.79%의 분포를 보이고 있다. 그림4. 과 그림5.는 모집단과 PBW set에서의 2음소열 출현빈도를 나타낸 것이다. 참고를 위하여 부록으로 추출된 PBW set를 실었다.

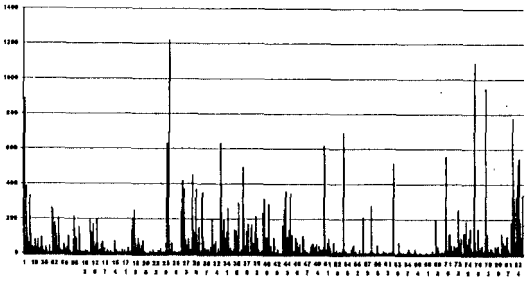

그림4. 고빈도 5000 어절의 2음소열 출현빈도

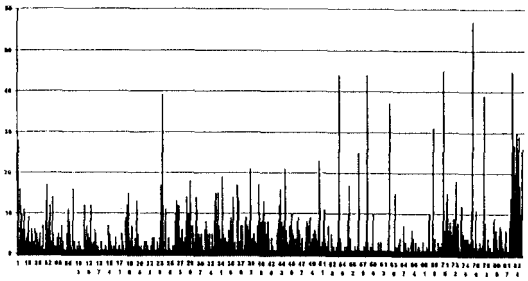

그림5. PBW의 2음소열 출현빈도

4. 결론

본 논문에서는 국내의 여러 관련기관에서 공통으로 사용할 수 있도록 하기 위해 국어공학 센터에서 추진하고 있는 한국어 음성DB구축의 일환으로 추출된 PBW set에 대하여 기술하였다. 구축중인 PBW 음성 데이타베이스는 국내의 음성 관련 기술의 개발과 성능 평가를 위해 대학이나 관련 연구기관에서 활용할 수 있도록 할 예정이며 앞으로 보완, 확장해 나갈 예정이다.

본 연구는 과기처의 연구비에 의해 수행된 것이다. 본 연구수행을 지원한 국어공학센타 박 동인 센타장, 국어 정보 베이스팀의 최 기선 박사께 감사드린다.

참 고 문 헌

[1] K. Shikano, "Phonetically balanced word list based on information entropy", Preprints Autumn Meeting Acous. Soc. Japan, Paper 3-3-10, Mar. 1984

[2] S. Hayamizu, et al, "Generation of VCV/CVC Balanced Word Sets for Speech Database", 일본 전자기술총합연구소, 제49권 제10호, 1985

[3] Yeonja Lim, Yonggjik Lee, "Implementation of the POW Algorithm for Speech Database", ICASSP-95, Vol. pp 89-92, Detroit 1995.

부록. 추출된 PBW set

청와대	컴퓨터	그에게
위대한	당노병	그야말로
에런대	분야에서	어두운
소프트웨어	했습니다	아니나는
야당의	니확암	요런대
자유와	주위의	최악이
의욕을	줘라고	과언이
되풀이	뇌물을	외외로
거에요	누워서	복해를
알코올	이집트	나와야
에이다	왔지만	경우와
왜곡된	뒤따라야	명예를
소외된	아쉬운	계열사
계획하고	로스엔젤레스	거액이
계약을	능치지	관세율보살
굉장히	스위스	의약품
최우선	쓰여진	임진왜란
집행유예	귀여운	노예가
두려워하지	핫반칠해	열쇠를
예민한	원만한	제약을
제도를	규제완화	좌우하는
폐적한	하였지만	관찰은
그쪽으로	뜨거워졌다	선행돼야
여의치	체장염이	취약한
합법화와	과부가	규법이
더위를	자외선	주체이트
공정거래위원회는		고역을
남품대금	분석했다	불완전한
세율을	쓸데없는	요소와
과일을	꿰뚫어	민방위
부유한	정계은퇴를	최고회의는
행하니	감수성이	고위급
변증법적	한국소비자보호원이	
서울기독교청년회		무역대표부
데이타베이스	오존층이	육체적
피해자의	어떻습니까	무엇보다도
부딪혀	못지않게	특특히
협동을	지할용	빼앗기고
에초에	케끗이	멀티미디어
법위내에서	살평균	에너지
죄책감을	효과적으로	위태롭게
세탁비누	이렇듯	바뀌어야
유의해야	응답했다	휴대용
끔에도	핵확산금지조약	발맞춰
영등포	발증에	동아일보
네다봤다	해외증권	헤매고
용크리고	에기입니다	국회의원
업격한	헌법을	개연성을
회겠지요	협싸여	왕조의
이워져야	수필집	연결되어
업체의	요즈음	카워야
외교와	핫마루에	알에까지
위치에	업계의	바르게
의미의	이유에	화폐가
통증이	의지와	체웠다
영어의	증여세	뉴욕에서
차례에	화해와	침체에
회화의	못살겠다	마음과
이외에는	금융실명제	지위에

바위가 교묘하게 우아한
중요성을 내뿜는 여학생은
인위적인 보궐선거 시야를
최도끼를 해택을 가요는
거처하는 뜻밖에 역스포
취임후 떠올랐다 해수욕장
두번째 올곡사업 주세요
예술적 끝끝내 더이상
인민대표대회 어깨에 의식과
외부의 빠아른 죽어야
마십시오 뒤었다 월요일
용의가 제조업 예부터
일제히 없어요 예전처럼
휘말려 이옥고 세포의
배노을 표적수사 터져나온
화증이 대통령 넘도록
알콧이 주의해야 예측할
소유의 필요에 눈시울이
교육과 참모습을 도입된
펼치고 배우자 이에앞서
코방귀를 않겠습니까 국세청
동양의 집계됐다 유심히
내려온 고마워 병원에
와서도 여태까지 회의의
처음부터 뚜렷이 싸우고
연평균 뒤쪽에 유치원
지역별로 부여와 조용히
소화성 외로운 자아의
배속에 여유가 터였다
메우고 규모의 예외가
더없이 일류의 재검토
못된다 쏘이고 도처에
끊어안고 웃도는 화면에
오트지 엣날에는 이유이다
불빛이 솜씨가 내었다
명확히 낮잘을 의의가
대법원 불잡고 마셨다
번화에 여의도 그처럼
놓았을 불량의 생애를
전야제 종소리가 명의로
기쁠을 수법을 회원이
의도가 교통사고 엉뚱한
못했지만 신회할 전원이
대응을 제외한 눈동자를
비호세력 개념과 약속했다
아카데미 오페라 흡어져
아파트 맹배해 염색체
외무부 매출액 앞세워
덧붙였다 위협적인 드다어
스포츠 구호품 있지요
에초부터 캘리포니아 오염된
쏘다져 에이즈 왕에게
생태계 특별히 어떻게
리더십 원래의 외세의
열다섯 태명앙 손쉽게
백화점 세번째 의자왕이
지켜야 때앗아 오후에
비추어 위원회 떨어져
남북관계 양조장 우주의
수수께끼 깨닫게 외교부

최초의 피해의 내세워
응답자의 어찌하여 취해야
악화될 우두커니 보험료
래이저 눈여겨 뜻대로
끝에서 재무부 윤곽이
배워야 왜냐하면 휴지통
법률적 어차피 위축되고
국민투표를 스웨덴 우리사회의
캄보디아 있어야 첫번째
앙쪽에 외에도 새벽녘에
뒤이어 염증이 외모에
총체적 분더러 편집국장
말겨야 요청에 결어되어
별기에 원활히 의하여
얄팍한 메체의 이것저것
창가에 헛물맘기 폭넓게
대이콤은 후원회 폴라스틱
갖춰야 뽑았다 와중에서
이때에 애기도 계속해서
두꺼운 엉거주춤 뒤어들러
요구에 깨닫고 아르바이트
유고적 노인성 영문과
제국주의 옳다고 섭취해야
팡계로 위기에 비소들히
어려웠면 띄었다 개입하지
시멘트 이용해 귀납적
유형의 두어야 표면에
류지컬 향하여 재조정
태초에 옆에서 의사도
원숭이의 떴었다 러사아
불법복제 완벽한 자율화
느껴졌다 위로부터의 양심과
때때로 할까요