훈련용 단어 음성DB 검증

이수종, 김상훈, 이영직
한국전자통신연구원 음성/언어정보연구센터

A Validation of the Isolated Word Speech Database

Soo-jong Lee, Sanghun Kim, Youngjik Lee
Speech/Language Technology Research Center, ETRI
E-mail : {sjleetri, ksh, ylee}@etri.re.kr

Abstract

The purpose of this paper is to correct the errors in the isolated word speech database under the PC environment, and to analyze the various errors. The importance and procedures of the error detection are also described.

I. 서론

음성DB는 음성처리에 필요한 기반요소로서, 음성인식, 합성, 인증 등 그 용도에 따라 화자의 선택, 시차적용, 환경 및 규모를 고려하여 수집하게 되며, 마이크 또는 기타 장치를 통하여 입력받고 분류되어 저장 및 가공하여 활용하게 된다. 음성인식률의 제고를 위해서는 서비스 환경에 따라 관련 영역에 적합한 음성DB가 충분히 확보되어야 하므로 보다 고품질의 음성DB 확보에 많은 시간과 노력을 기울이고 있다.

본 논문에서는 음성정보처리기반기술의 일환으로 구축한 다양한 종류의 표준형 한국어 공통음성DB 중에서, PC환경 하에서 중가 마이크를 통하여 수집한 훈련용 단어 음성DB에 대한 오류를 검출하여 확인하고 다양한 오류의 유형들을 분류하고자 한다. 이하에서는 음성DB에 대한 오류검증의 필요성, 분석대상의 전체 음성DB 구조 및 검증환경, 오류검출 절차, 실제 오류확인 대상 자료의 추출 및 오류확인 결과, 오류의 유형 분류 등을 차례로 살펴보고자 한다.

II. 오류검증

음성DB에는 음성데이터, 음성데이터의 내용을 텍스트로 표기한 철자전사, 음성의 시작과 끝을 나타내는 음성구간 정보 등이 각각 하나의 단위를 이루는 묶음들로 구성된다. 철자전사의 각 어휘는 음소로 분리되어 어휘사전을 이루고, 음성데이터에 포함된 음향학적 정보로부터 해당 음운 및 언어적 정보와의 모델링 과정을 거쳐 통계적으로 서로 연결됨으로써, 음성과 어휘 간에 서로 가장 유사한 데이터를 찾아서 반응하게 된다. 따라서 음성데이터와 철자전사 간에 괴리가 있게 되면 음성처리에 심각한 문제가 발생하는 원인이 된다. 그러므로 음성DB를 신뢰성 있게 구축하기 위해서는 요구사항이 구체적이고 명확히 제시되어야 하고, 자연성 있는 시나리오에 의해 제작되어야 함은 물론 제작 이후에도 지속적으로 확인, 검증하여 오류를 추출 및 제거하여 그 적합성을 확보해 가야 한다.

음성DB에 대한 검증은 여러 단계에 걸쳐 이루어지고 있는데, 전체 데이터를 대상으로 하여 규격과의 비교검증 및 기초 데이터의 충실성 여부를 확인하는 기본확인단계, 인식오류의 유발가능성이 있는 데이터만을 추출하여 일정범위 내에서 확인하는 집중검증 단계 및 활용과정에서의 지속적인 오류보고에 의한 재확인 검증단계로 나누어 볼 수 있다. 본 논문에서는 집중검증 단계의 일환으로 가장 많이 활용될 것으로 보이는 분야를 대상으로 오류검증을 시도하였으며, 지속적인 오류검출 방법이 모색되어야 하겠다.

III. 오류검증 환경

오류검출에 앞서 분석의 대상이 되는 전체 음성DB의 내용과 구조, 분석에 활용된 도구에 대하여 살펴보고, 아울러 전체 음성DB 중에서 실제 오류확인 대상 자료를 추출하기 위해 수행한 인식실험의 형상을 소개하고자 한다.

3.1 음성DB 구조

분석에 사용될 단어음성DB는 10,000 단어를 대상으로 1,000명의 화자가 발성한 결과파일들로서 모두 100,000 발성어휘의 분량이며, 100개의 세트로 구분되어 있다. 각 세트는 100 단어를 대상으로 10명의 화자가 1회씩 발성한 어휘들의 묶음이다. 따라서 동일한 단어에 대해 10회씩 발성된 셈이다. 단어의 발성목록은 상장회사명, 지명, 인명, 상호명, 제품명, PC명령어, PDA명령어, 그 외에 일반명사로 구성되었다. 화자는 성별, 연령별, 지역에 따라 적정비율로 분포되었다.

표 1. 분석대상 음성DB 개요

종류	단어음성DB, 중가 마이크, PC환경
규모	100,000발성어휘(10,000 단어, 1,000 화자)
세트	100 set (1set : 100 단어 x 10명)

또한 음성데이터는 16kHz 샘플링, 16bit linear PCM으로 저장되어 있으며, 음성구간의 앞, 뒤에 200 msec의 묵음구간이 포함되었다. 음성DB의 각 파일명은 13개의 digit로 구성되어 코드화되어 있어서 이를 통하여 음성DB의 종류, 수집환경, 세트의 구분, 발성순서 등의 식별이 가능하다. 또한, 한 단위의 음성DB 묶음은 3종류의 확장자로 나누어져 있으므로 각각 필요한 정보를 확인할 수 있도록 되어 있다.

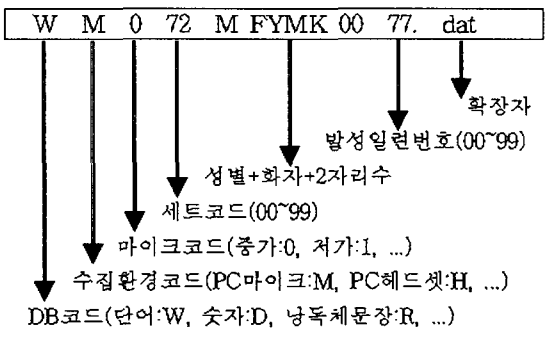

그림 1. 음성데이터 파일명 코드

다음의 <표 2>는 한 단위의 음성DB 묶음에 포함된 3종류의 파일을 보여주고 있다.

표 2. 음성DB 구성파일 (예)

단위별 파일 구성(3종)	파일 내용
WM072FYMK0077.dat	음성데이터(음성파형)
WM072FYMK0077.TEXT	철자전사(텍스트)
WM072FYMK0077.MARK	음성구간 정보(시간정보) 0.200063 -1 SBEG 1.302063 -1 SEND

따라서 분석대상 전체의 음성DB는 위와 같은 3종의 단위 묶음들이 100,000 종이 됨을 의미한다.

3.2 오류검증 환경

음성DB는 음성데이터가 포함되어 있어서, 대량의 음성DB를 효과적으로 처리하기 위해서는 대용량의 서버 활용이 불가피한데, 리눅스 서버를 포함하여 각종 상용도구, 스크립트 프로그램, 철자전사 텍스트의 각 음소를 음운으로 바꾸기 위한 GtP (Grapheme to Phoneme), 그리고 음성훈련기 및 가변어휘음성인식 엔진이 활용되었다.

스크립트 프로그램은 단계별로 작성되어 활용할 수 있었으며, 주요 활용내역은 다음과 같다. 먼저, 분석대상의 전체 음성DB를 대상으로 하여 철자전사 목록만을 수집하여 별도의 철자전사 목록 파일 작성 및 체계화, 트라이폰 생성, 어휘사전 생성, 특징추출 등의 자료를 생성하였다. 이들 스크립트 프로그램에 의하여, 전체 음성DB를 대상으로 하는 훈련 및 인식실험에 필요한 제반자료가 마련되었다. 음성데이터는 특징추출 단계를 거쳐 HMM 음소 훈련에 사용되며, 훈련에 사용된 음성데이터가 인식실험에 사용된다. 마킹정보는 음성구간 정보로 활용된다.

분석대상 음성DB는 모두 100세트이나 10세트씩을 단위로 하여 스크립트 프로그램 수행, 인식실험, 오류확인 및 결과 집계가 이루어졌다. 대규모 자료를 대상으로 하였으므로 오류검증에 활용된 장비의 이용률과 상태, 세트별 분석 수행의 순서와 시차에 따라 다소의 편차가 있을 수 있으나 전체 집계과정에서 해소된 것으로 사료된다.

IV. 인식실험

오류확인 대상 데이터의 범위를 좁혀 오류가능성이 높은 자료만을 대상으로 실제 확인하기 위하여, 전체 음성DB를 대상으로 close-test를 수행하여 오류 가능성이 높은 음성DB를 분리하였다. 인식실험에는 가변어

휘인식엔진을 활용하였다. 훈련 및 인식수행 형상으로서, 특징추출 파라미터는 MFCC 13차와 delta 13차를 합하여 26차로 하였고, 음소단위별 state 변화, 특징추출에 따른 HMM모델 mixture, 학습 횟수, 인식 score 비교에 의한 어휘인식은 3차까지 출력될 수 있도록 조건을 부여하였다.

표 3. 훈련 및 인식수행 형상

특징추출 파라미터	26차(MFCC 13차+ delta 13차)
State 수	3
Mixture 수	8
학습 횟수	30
N-best	N:3 (Top-I, Top-II, Top-III)
Test 형태	Closed Test

3-best 조건 하에 수행된 인식실험 결과는 음성데이터와 철자전사와의 일치여부를 3차까지 비교해 볼 수 있도록 출력된다. 즉, 1차에서 인식결과가 일치하여 인식성공이 이루어지는 경우, 1차에서는 인식오류이나 2차에서 일치하는 경우, 3차의 인식결과와 일치하는 경우, 그리고 3차례의 어느 경우와도 일치하지 않는 경우로 분류된다. 본 논문에서는 1차에서 인식 성공한 경우 외에는 모두 확인대상으로 하였다. 인식실험 결과에는 음성데이터, 철자전사, 3단계의 인식결과 외에도 인식 Score 값을 출력해 볼 수 있다.

표 4. 3-best 인식결과 (예)

일련번호		[철자전사]	[TOP-I]		[TOP-II]		[TOP-III]		
1	WM090FAMJ0017	+신한증권	최정헌	-1950.6	현대속셈학원	-1997.5	한스건설	-2032.3	error
2	WM090FAMJ0056	+진영축산	제일축산	-1743.5	김현순	-1802.5	두진소방	-1808.6	error
3	WM090FJDJ0082	+오라도화점	모라도화점	-1024.4	오라도화점++	-1035.1	동방박사제포	-1460.5	error
...	.....	.....	.....	.....	.....	.....	.....	.....	.....

오류가능범위로 분류된 음성파일에 대하여는 파일명 코드를 활용하여 원본파일의 저장경로를 추적한 후 녹음내용을 직접 청취하여 녹음상태를 확인하고 철자전사와 비교하였다. 녹음내용의 청취에는 녹음청취만을 위해 구현된 도구를 활용하여 신속한 작업이 이루어질 수 있었다. 음성데이터의 청취결과를 토대로 녹음상태의 적정성과 발음의 명료성 여부를 확인함과 동시에 철자전사와 비교하여 그 일치성 여부를 검증하였다.

V. 오류유형 분류

음성DB에 대한 오류분석은 음성자체의 품질확인과 발음의 표기에 대한 검토로 이루어지는데, 첫째는 음성의 품질과 관련하여, 발성의 내용은 명료한가, 잡음은 어느 정도인가, 그리고 음성파형의 최대 크기는 적절한가 등에 관한 것이다. 둘째는 음성의 지속시간 표시를 위한 마킹정보에 관한 것으로서, 묵음구간을 포함하여 음성구간이 적절히 표시되어 있는가이다. 셋째, 발음에 대응되는 철자전사가 발음의 내용과 서로 일치하고 철자전사가 맞춤법 및 표기법을 따랐는가. 넷째, 발음의 소리값을 그대로 나타낼 수 있도록 발음전사가 제대로 되어 있는가이다. 그러나 분석대상 음성DB의 경우에는 발음전사는 포함되어 있지 않았다.

5.1 오류구분

음성DB의 오류는 앞에서 언급된 바와 같이 음성오류와 음성표기 오류로 크게 나눌 수 있는데, 좀 더 세분하여 여섯 가지의 유형으로 분류하였다. 첫째의 유형은 녹음오류이다. 발성목록과 철자전사는 되어 있으나 음성데이터가 없는 경우이다. 아주 드문 경우에 속하나 치명적인 오류로 분류되며, 2건이 발견되었다. 둘째는 발음오류이다. 모호하게 발음함으로써 철자전사와 일치한다고 볼 수 없는 경우이다. 한꺼번에 발음하지 않고 더듬거린 경우에도 발음오류에 포함시켰다. 셋째로는 철자전사오류이다. 발성목록에 따라 발성하도록 되어있으나, 이와는 다른 단어로 발음한 경우라도 녹음상태가 양호한 경우에는 녹음내용을 기준으로 철자전사를 하게 되며, 음성처리과정에서는 철자전사 결과가 실제로 활용된다. 따라서 녹음품질을 우선 확인한 다음에 철자전사 결과와 비교하였다. 넷째는 띄어쓰기 오류이다. 스크립트 작성 및 인식테스트 과정에서 도구에 의해 자동으로 붙이도록 하기 때문에 경미한 오류로 분류될 수 있다. 띄어쓰기 오류에 대하여, 일정한 음절 다음에만 띄어져 있는 점을 중시하여 추적한 결과 텍스트 편집기의 오작동에 기인한 것으로 판명되었다. 다섯째, 맞춤법 오류이다. "청계휴게실"을 "청계휴계실"로 표기하거나, "물류본부"를 "물유본부"로 표기한 경우가 그 예이다. 마지막으로 여섯째는 파일명 오류이다. 여성의 발음을 남성의 파일명 코드로 표기한 경우이다.

다음의 <표 5>는 오류확인 대상 범위와 오류유형별로 집계한 결과를 보여준다. 분석대상 음성DB의 수는 모두 99,867이며, 이들 중에서 인식실험을 통하여 6,269종의 음성DB를 추출하여 오류확인을 수행하였다.

표 5. 오류유형 분류

오류유형						
녹음	발음	철자	띄움	맞춤법	파일명	합계
2	17	77	50	7	77	230
0.87%	7.39%	33.48%	21.74%	3.04%	33.48%	100%

이외에도 발음속도에 있어서 심한 편차가 있는 경우가 있었다. 화자의 발성음량이 기준이하인 경우는 발견되지 않았으나, 화자에 따라서는 자연성을 지나치게 의식한 나머지 화자 고유의 발성음이 훼손되었다고 판단되는 경우가 다수 확인되었다. 자연성의 기준은 실제 음성 서비스 활용 현장에서 이용자가 어떤 태도로 임할 것인가와 서비스 이용시점에서 음성시스템의 인식률 수준에 따라 이용자가 어떻게 반응할 것인가에 초점을 맞춰 마련되어야 할 것이다.

음성데이터의 오류는 실제 서비스 상황을 감안하지 않고 발성을 수행한 데에 주로 기인하는 것으로 판단된다. 따라서 바람직한 음성데이터의 수집을 위해서는 실제 서비스 상황을 고려한 테스트베드를 설치하여 화자로 하여금 스스로의 음성파형 및 음량, 발음속도, 띄어 읽기, 인식결과 등을 스스로 확인해 볼 수 있도록 하고, 다른 화자들과 비교할 수 있도록 하여 음성서비스 이용 상황에 적합한 자연성을 추구해야 하겠다.

5.2 음성DB보완

오류확인 결과에 따른 음성DB의 보완은 오류유형에 따라 달리 처리되는데, 오류데이터의 파일명 코드 정보에 따라 경로를 추적하여 해당 원본 데이터에 접근한 후 재확인 과정을 거쳐 삭제 및 정정 등 보완이 이루어진다. 녹음오류나 발음오류로 최종 판정되는 경우에는 해당 음성DB를 삭제하고, 음성데이터의 표기에 관련된 오류는 관련된 자료를 정정하게 된다. 음성표기에 대한 오류는 사후 보완이 가능하나, 일단 구축된 음성데이터의 오류는 삭제 외에는 보완방법이 없는 것으로 판단된다.

오류확인 과정에서 많은 도구를 활용함에도 불구하고 인적오류의 가능성이 항상 존재하는 점을 감안하여 원본 데이터에 대한 최종 보완이 이루어지기까지는 관련인원들에 의해 수차에 걸친 재확인 과정을 거치게 된다. 실제로 오류로 분류된 자료 중에는 정상범주로 재분류되는 사례로 많이 있었다. 따라서 분석대상 음성DB 데이터의 대규모화와 오류확인 과정에서의 휴먼에러의 가능성을 최소화하기 위해서는 실제 오류확인 대상 데이터를 효과적으로 추출할 수 있는 음성DB 검증전용 시스템의 구현 및 활용이 요구되고 있으며, 현재 구현이 진행되고 있다.

VI. 결론

본 논문에서는 음성정보처리의 기반요소로서 구축된 한국어 공통음성DB 중에서 훈련용 단어 음성DB를 대상으로 하여 오류를 검출하고 그 유형의 분류를 시도하였다. 녹음오류 및 발음오류 등의 음성데이터 오류와 음성표기 오류가 발견되었고, 재확인 과정을 거쳐 보완되었다. 이를 토대로 하여 여러 종류의 음성DB에 대한 오류확인이 계속되어야 하겠다. 또한, 오류확인 대상 자료의 정확한 확보와 검증과정에서의 휴먼에러를 최소화하기 위한 음성DB검증 전용 시스템을 활용함으로써 음성DB에 대한 신뢰성을 확보해 가야 하겠다.

참고문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Published by Prentice Hall PTR, 1993.

[2] Xuedong Huang, Alex Acero, Hsiao -Wuen Hon, "Spoken Language Processing", Prentice-Hall, inc. 2001.

[3] 이건상, 양성일, 권영헌, "음성인식", 한양대학교출판부, 2001.

[4] 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성DB 구축," *한국음향학회 하계학술대회논문집*, 제21권, 제1(s)호, pp.21-24, 2002

[5] ETRI 음성/언어정보연구센터: http://voice.etri.re.kr