외국어 발화오류 검출 음성인식기의 성능 개선을 위한 스코어링 기법*

강효원(대전대), 권철홍(대전대)

<차 례>

<Abstract>

Scoring Methods for Improvement of Speech Recognizer Detecting Mispronunciation of Foreign Language

Hyo-Won Kang, Chul-Hong Kwon

An automatic pronunciation correction system provides learners with correction guidelines for each mispronunciation. For this purpose we develope a speech recognizer which automatically classifies pronunciation errors when Koreans speak a foreign language. In order to develope the methods for automatic assessment of pronunciation quality, we propose a language model based score as a machine score in the speech recognizer. Experimental results show that the language model based score had higher correlation with human scores than that obtained using the conventional log-likelihood based score.

* Keywords: Automatic detection of mispronunciation, Speech recognition, Scoring methods

* 본 연구는 한국과학재단 목적기초연구(R01-2002-000-00283-0) 지원으로 수행되었음.

1. 서 론

국제화 시대에 있어서 외국어 발화의 발음 교정에 대한 관심이 높아지고 있다. 이러한 발화 양상의 교정은 극소수의 언어·음성학 전문가들에 의해서만 수행되고 있는 형편이며, 아동의 언어발달과 성인의 외국어 습득을 위한 발음교정은 여러 차원에서 교육 및 재교육이 필요하나, 비용과 접근의 불편함 때문에 그 해결의 중요성에도 불구하고 기회 제공이 원활히 이루어져 오고 있지 않다. 이를 위해서 음성인식 기술을 이용하여 발음상의 미비함, 조음기관 동작의 부정확성을 정확히 인식하고 그 각각의 범주에 따른 교정 방안을 컴퓨터를 통하여 제시하는 시스템의 개발이 필요한 실정이다.

음성인식 엔진을 응용한 외국어 발음 학습기가 많이 출시되고 있으나, 출시된 제품은 외국에서 개발된 엔진이거나 엔진 활용 용도가 일반적인 상업용 엔진을 사용하기 때문에 한국인에게 적합한 수준의 학습 결과를 제시하지 못하고 있다. 본 논문은 HMM (Hidden Markov Model) 음성인식 엔진을 발음 교정 시스템에 최적화 시켜 외국어를 발음하는데 있어서 틀리기 쉬운 오류 발음을 정확히 찾아내 학습자에게 최적의 교정 지침을 제공할 수 있도록 하는데 그 목표가 있다.

발음 자동 교정 시스템은 한국인의 외국어 발화시 발음에 따른 오류 양상을 음성인식기로 자동으로 검출하여 오류 유형별로 발음 교정 지침을 제시하는 시스템을 말한다. 발음오류유형을 자동으로 분류하기 위한 음성인식기는 오류 유형별 발화 모델들을 미리 수집하여 HMM 모델을 통해 음소인식을 수행한다. 발음 교정 교수법은 언어·음성학 전문지식을 이용하여 발음 오류에 정확히 해당하는 교정 사항을 제공하여 학습자가 효과적으로 발음 교정을 수행하도록 교정지침을 제공한다.

본 논문에서는 한국인이 일본어 발화시 음소별로 발음오류를 자동으로 분류하는 HMM 음성인식기를 다룬다. 그리고 음성인식기의 성능 개선을 위하여 언어모델 스코어를 제안한다. 논문의 구성은 2장에서 발음오류 자동분류 HMM 음성인식기의 구현방법에 대하여 기술하고, 3장에서 제안된 언어 모델 스코어를 설명하고, 4장에서 실험 방법 및 결과를 논하고, 5장에서 결론 및 향후 연구과제에 대하여 기술한다.

2. 발음오류 검출 HMM 음성인식기

본 장에서는, 한국인이 일본어 발화시 일본어 음소와 혼동하여 발음하는 한국어 음소들을 분류하여 유사발음 음소 셋을 선정하고, 이 음소 셋을 이용하여 음소별로 발음오류를 검출하는 HMM 음성인식기를 설명한다.

2.1. 음소 셋의 선정

음소인식에 필요한 각 언어별 음소 셋 및 유사음소 셋을 선정한 방식은 다음과 같다. 한국어 음소 셋은 자음은 변이음을 고려하여 29개를, 모음은 음성학적 차이를 보이는 모음만을 고려하여 17개를 선정하여, 총 46개의 음소로 구성하였다[1]. 일본어 음소 셋은 일본음향학회에서 선정한 음소 셋을 참조하여[2], 38개의 음소를 선정하였다[1]. 유사발음 음소 셋은 한국인이 일본어 발화시 발음에 따른 오류 양상을 음성학 전문가의 전문 지식 및 수집한 음성 DB를 활용하여 일본어 음소별 한국인 화자의 발음오류 유형을 분류하였다[1]. 즉, 일본어 음성학 전문가가 한국인이 발화한 일본어 음성 데이터를 청취·분석하여, 한국인이 일본어 발화시 나타나는 음소별 발음오류 양상을 분석하였다. 유사발음 음소 셋의 한 예는 <표 1>과 같다.

<표 1> 유사발음 음소 셋의 예

일본어 음소	유사발음 일본어 음소	오류발음 한국어 음소
k	ky, g, gy	kkc /ㄲ/, khc /ㅋ/, kc, gc /ㄱ/

<표 1>에서 일본어 음소 [k]를 일본어 [ky, g, gy]로 발화하거나, 한국어 [kkc /ㄲ/, khc /ㅋ/, kc /무성 ㄱ/, gc /유성 ㄱ/]로 발화하는 오류가 발생함을 의미한다.

2.2. HMM 모델의 구성

일본어 음소 모델 구성을 위하여 정확한 발음의 일본어 음성 DB가 필요하므로, 고려대학교 부설 한국어교육센터에서 한국어 연수중인 일본인 70인을 대상으로 ATR(일본 자동통역 연구소)에서 작성한 일본어 PBW (Phonetically Balanced Words) 216개 단어를 녹음하였다. 또한 일본어 음소와 음가는 비슷하나 음성 자질이 틀린 한국어 음소 모델을 만들기 위하여, SITEC(음성정보기술 산업지원센터)에서 작성한 한국어 PBW 452개 단어를 대전대학교 대학생 70인을 대상으로 수집하였다. 녹음환경은 연구실 수준의 조용한 방에서 1명씩 녹음하였고, 사운드 카드는 Soundblaster Audigy를, 마이크는 SHURE 565SD를 사용하였다.

이와 같은 방식으로 수집한 음성 DB를 이용하여 다음과 같은 방법으로 음소별 HMM 모델을 구성하였다. 음성신호를 매 10msec 마다 25msec의 Hamming 창함수

를 사용하여 음성 특징(MFCC, Mel Frequency Cepstrum Coefficients)을 추출하였고, HMM의 구조는 3 state left-to-right continuous HMM을 사용하고, 일본어, 한국어 모두 Mixture 수는 15로 하여 음소별 HMM 음향모델을 생성하였다[1].

3. 언어모델의 가중치 조정 스코어

본 논문에서 연구의 초점은 발음오류유형 자동분류 음성인식기가 인식해 낸 결과와 해당 언어 음성학 전문가의 청취 판단 결과가 가능한 한 유사하게 나타내게 하는 데 있다. 이를 위해서 다음과 같은 방법으로 언어모델에 가중치를 적용함으로써 성능 개선을 시도하였다.

<그림 1>은 일본어 단어 (*i-k-i-o-i*)에서 [k] 음소에 대하여 발음오류를 분류하기 위한 발음 네트워크로, 일본어 [k] 음소에 대하여 일본어 [k, ky, g, gy]와 한국어 [kkc /ㄲ/, khc /ㅋ/, kc /무성 ㄱ/, gc /유성 ㄱ/]로 인식하기 위한 발음 네트워크이다. 이 그림에서 각 음소에 대한 언어모델의 확률은 다음과 같이 구한다.

$$X = \ln \frac{1}{N} \tag{1}$$

이 식에서 N은 오류발음유형 음소 개수이다. 일본어 음소 [k]의 오류음소의 개수는 8개이기 때문에 N은 8이고, 따라서 X는 -2.08로 동일함을 알 수 있다. 이것은 단순히 문법구조의 분류로만 볼 수 있을 뿐 언어모델의 필요성에 따른 역할을 수행한다고 볼 수 없다.

본 논문에서는 언어모델의 확률을 청취 판단에 근거하여 가중치를 적용하였다. 가중치를 적용한 수식은 식 (2)와 같이 구할 수 있다. 식 (2)는 각 음소의 청취 판단 확률에 자연로그 값을 취한 것이다.

$$X = \ln P \tag{2}$$

이 식에서 P는 오류음소별 청취 판단 비율을 의미한다. 즉, (청취 판단개수/총 개수) 이다. 그런데 청취 판단 개수가 0인 음소인 경우, P가 0이 되어 그 음소로 인식되는 경우가 없게 된다. 청취 판단 개수가 0인 경우에도 인식될 수 있는 최소한의 가능성을 열어 두기 위해 식 (2)를 다음과 같이 수정하였다.

$$X = \ln\left(\frac{1}{10N} + P*0.9\right) \tag{3}$$

이 식에서 $\frac{1}{10N}$ 을 유사음소 개수인 N개만큼 더하면 0.1이 되고, $P*0.9$는 유사음소들이 갖는 확률을 모두 더하면 0.9가 된다. 즉, 유사음소로 선정된 음소가 갖는 가능성을 1, 청취 판단한 결과의 가능성을 9로 하여 1:9의 비율로 값을 결정하였다. 이 비율은 실험적으로 결정한 값이다.

만약 총 개수가 20개이고, (ㄲ /kkc/)에 대한 청취 판단의 개수가 12개라면 오류음소 (ㄲ /kkc/)에 대한 언어모델의 가중치는 다음과 같다.

$$X = \ln(\frac{1}{10*8} + \frac{12}{20}*0.9) = -0.59 \quad (4)$$

<그림 2>는 청취 판단 결과가 (ㄲ /kkc/)는 12개($X = -0.59$), (ㅋ /khc/)는 8개($X = -0.99$), (무성 ㄱ /kc/)는 0개($X = -4.28$), (유성 ㄱ /gc/)는 0개($X = -4.28$)인 경우의 발음 네트워크이다. <그림 2>에서 적용된 언어모델은 인식과정에서 스코어링 하는 과정에서 음향모델과의 합으로 계산되어지며 기존의 방식으로 생성된 스코어와 비교된다.

```
 1 VERSION=1.0
 2 N=17   L=23
 3 I=0    W=!NULL
 4 I=1    W=!NULL
 5 I=2    W=SENT-START
 6 I=3    W=i
 7 I=4    W=kkc
 8 I=5    W=!NULL
 9 I=6    W=khc
10 I=7    W=kc
11 I=8    W=gc
12 I=9    W=k
13 I=10   W=ky
14 I=11   W=g
15 I=12   W=gy
16 I=13   W=i
17 I=14   W=o
18 I=15   W=i
19 I=16   W=SENT-END
20 J=0    S=16    E=1    l=0.00
21 J=1    S=0     E=2    l=0.00
22 J=2    S=2     E=3    l=0.00
23 J=3    S=3     E=4    l=-2.08
24 J=4    S=4     E=5    l=0.00
25 J=5    S=6     E=5    l=0.00
26 J=6    S=7     E=5    l=0.00
27 J=7    S=8     E=5    l=0.00
28 J=8    S=9     E=5    l=0.00
29 J=9    S=10    E=5    l=0.00
30 J=10   S=11    E=5    l=0.00
31 J=11   S=12    E=5    l=0.00
32 J=12   S=3     E=6    l=-2.08
33 J=13   S=3     E=7    l=-2.08
34 J=14   S=3     E=8    l=-2.08
35 J=15   S=3     E=9    l=-2.08
36 J=16   S=3     E=10   l=-2.08
37 J=17   S=3     E=11   l=-2.08
38 J=18   S=3     E=12   l=-2.08
39 J=19   S=5     E=13   l=0.00
40 J=20   S=13    E=14   l=0.00
41 J=21   S=14    E=15   l=0.00
42 J=22   S=15    E=16   l=0.00
43
```

<그림 1> 가중치 조정 전 발음 네트워크

```
 1 VERSION=1.0
 2 N=17   L=23
 3 I=0    W=!NULL
 4 I=1    W=!NULL
 5 I=2    W=SENT-START
 6 I=3    W=i
 7 I=4    W=k
 8 I=5    W=!NULL
 9 I=6    W=ky
10 I=7    W=g
11 I=8    W=gy
12 I=9    W=kkc
13 I=10   W=khc
14 I=11   W=kc
15 I=12   W=gc
16 I=13   W=i
17 I=14   W=o
18 I=15   W=i
19 I=16   W=SENT-END
20 J=0    S=16    E=1    l=0.00
21 J=1    S=0     E=2    l=0.00
22 J=2    S=2     E=3    l=0.00
23 J=3    S=3     E=4    l=-4.28
24 J=4    S=4     E=5    l=0.00
25 J=5    S=6     E=5    l=0.00
26 J=6    S=7     E=5    l=0.00
27 J=7    S=8     E=5    l=0.00
28 J=8    S=9     E=5    l=0.00
29 J=9    S=10    E=5    l=0.00
30 J=10   S=11    E=5    l=0.00
31 J=11   S=12    E=5    l=0.00
32 J=12   S=3     E=6    l=-4.28
33 J=13   S=3     E=7    l=-4.28
34 J=14   S=3     E=8    l=-4.28
35 J=15   S=3     E=9    l=-0.59
36 J=16   S=3     E=10   l=-0.99
37 J=17   S=3     E=11   l=-4.28
38 J=18   S=3     E=12   l=-4.28
39 J=19   S=5     E=13   l=0.00
40 J=20   S=13    E=14   l=0.00
41 J=21   S=14    E=15   l=0.00
42 J=22   S=15    E=16   l=0.00
43
```

<그림 2> 가중치 조정 후 발음 네트워크

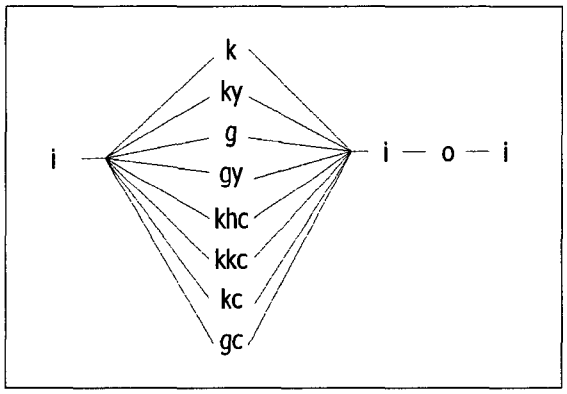
4. 실험 및 결과

4.1. 발음오류 자동분류 음성인식기의 인식률 및 정확도 분석

4.1.1. 실험 방법

발음오류 자동분류 음성인식기의 성능을 검증하기 위하여 비전공 한국인 대학생 30인이 일본어 PBW 단어 24개를 발화한 음성 데이터를 수집하였다. 이들을 대상으로 인식 실험을 수행하여 유사음소별 인식 결과를 구했다.

실험을 위한 발음 네트워크는 <그림 1>의 가중치를 조정하기 전 네트워크로, 이를 그림으로 도시한 <그림 3>과 같이 일본어 음소 [k]에 대하여 유사발음 일본어 음소(ky, g, gy)와 한국어 오류음소(khc, kkc, kc, gc)를 통합한 네트워크를 사용하였다.



<그림 3> 오류발음 검출 발음 네트워크
(일본어 단어 i-k-i-o-i 인 경우)

4.1.2. 자동분류 결과와 청취 판단과의 비교 분석

청취 판단과의 정확도 비교 실험은 각 화자별 각 단어별 인식 결과를 분류하여 실험 대상으로 선정하였다. 청취 판단 결과 일본어로 발화한 경우는 없으므로 일본어 [k, ky, g, gy]에 대한 실험 결과는 비교할 수 없다. 또한 청취 판단 결과 어두 음소와 어중음소가 각 음소별로 크게 차이를 보이므로[3], 실험 결과를 어두와 어중으로 나누어 비교하였다.

<표 2, 3>에서 보듯이, 자동분류 결과 중에서 어중 /ㄲ/, /ㅋ/ 음소에 대하여서는

어느 정도의 성능을 얻었다고 할 수 있으나, 나머지 경우에는 청취 판단과의 일치도가 좋지 않았다. 이와 같이 인식률의 정확도가 떨어지는 것을 개선시키기 위하여 3장에서 설명한 스코어링 방법을 적용하여 실험을 수행하였다.

<표 2> 어두 [k] 음소의 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	10	1	10.0
ㅋ(khc)	61	9	14.8
ㄱ(kc/gc)	139	37	26.6

<표 3> 어중 [k] 음소의 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	323	192	59.4
ㅋ(khc)	129	47	36.4
ㄱ(kc/gc)	58	6	10.3

<표 4> 음소별 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	333	193	58.0
ㅋ(khc)	190	56	29.5
ㄱ(kc/gc)	197	43	21.8

4.2. 언어모델 스코어링에 따른 인식률 및 정확도 분석

4.2.1. 실험 방법

언어모델의 가중치를 사용하면 언어적 지식을 기계 스코어(machine score)에 접목시켜 인식에 큰 영향을 미칠 수 있다.

실험을 위한 발음 네트워크는 기본적으로 <그림 3>과 같은 네트워크를 사용하였다. 그리고 3장에서 설명한 바와 같이 언어모델에 가중치를 적용하여 최종적으로 <그림 2>와 같은 발음 네트워크를 사용하였다.

4.2.2. 실험 결과

가중치 조정 없이 인식 실험을 한 결과 <표 5>에서 보듯이, 총 720개의 발화 데이터 중 일본어 [k] 음소로 인식된 결과는 45.8%인 330개, /ㄲ/는 31.7%인 228개, /ㅋ/는 12.8%인 92개, /ㄱ/는 9.7%인 70개로, 일본어 [k]와 한국어 /ㄲ/ 음소로 발화한 음성 데이터가 큰 비중을 차지하였다. <표 6>에 청취 판단을 근거로 각 음소별로 언어모델의 가중치를 조정하여 인식 실험한 결과가 보인다. 가중치를 조정한 실험 결과는, [k] 음소로 인식된 결과는 32.1%인 231개, /ㄲ/는 39.8%인 287개, /ㅋ/는 16.4%인 118개, /ㄱ/는 11.7%인 784개로, 한국어 음소의 인식 비중이 모두 증가하였고 일본어 음소 [k]로의 인식 비중이 줄었다.

<표 5> [k] 음소의 가중치 조정 전 인식 결과

음소		[k]	ㄲ(kkc)	ㅋ(khc)	ㄱ(kc/gc)
인식 비중	(%)	45.8	31.7	12.8	9.7
	개수	330	228	92	70

<표 6> [k] 음소의 가중치 조정 후 인식 결과

음소		[k]	ㄲ(kkc)	ㅋ(khc)	ㄱ(kc/gc)
인식 비중	(%)	32.1	39.8	16.4	11.7
	개수	231	287	118	84

4.2.3. 자동분류 결과와 청취 판단과의 비교 분석

<표 7, 8>을 보면 언어모델 적용 전(<표 2, 3>)과 비교해서 어중의 /ㄱ/ 음소를 제외한 모든 경우에서 청취 판단과의 일치도가 증가함을 알 수 있다. 음소별로 살

펴보면, /ㄲ/ 음소인 경우 58.0%에서 72.1%로 14.1%, /ㅋ/ 음소는 29.5%에서 40.0%로 10.5%, /ㄱ/ 음소는 21.8%에서 30.0%로 8.2%가 향상되었다(표 4와 9 비교). 청취 판단과의 비교대상 전체 발화 데이터 720개 중에서, 가중치를 조정하기 전 방식은 40.6%(292개)의 일치도를 보여 주었고, 가중치를 조정한 방식은 52.1%(375개)의 일치도를 보여 주어 11.5%의 성능 개선을 이루었다. 이와 같은 결과로부터 본 논문에서 제안한 언어모델의 가중치를 조정한 스코어링 방법이 상당한 성과를 얻었다는 것을 알 수 있다.

<표 7> 가중치가 조정된 어두 [k] 음소의 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	10	2	20.0
ㅋ(khc)	61	13	21.3
ㄱ(kc/gc)	139	53	38.1

<표 8> 가중치가 조정된 어중 [k] 음소의 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	323	238	73.7
ㅋ(khc)	129	63	48.8
ㄱ(kc/gc)	58	6	10.3

<표 9> 가중치가 조정된 음소별 청취 판단 결과와의 정확도 비교

	청취 판단 개수	청취 판단과 일치한 인식 개수	정확도(%)
ㄲ(kkc)	333	240	72.1
ㅋ(khc)	190	76	40.0
ㄱ(kc/gc)	197	59	30.0

5. 결 론

본 논문에서는 음성인식 기술의 응용분야인 외국어 학습기에 초점을 맞추어 발음 교정 시스템을 위한 발음오류유형 자동분류 음성 인식기를 구현하였다. 또한, 음성인식기가 표준 및 오류 발음을 자동으로 분류하기 위하여 학습자의 발음에 스코어를 주는 방법에 대한 연구를 수행하였다. 이 스코어는 전문음성학자의 판단과 상관관계가 높은 것이 필요하므로 음성학자의 청취 판단과의 결과를 비교하였다. 본 논문에서 제안한 언어 모델에 청취 판단의 결과를 기초로 가중치를 준 스코어를 통해 음성학자의 청취 판단과 높은 상관관계를 갖는 결과를 얻을 수 있었다.

향후 연구에서는 자동 분류의 정확도를 높여 자동 분류 결과와 음성학 전문가의 청취 판단을 좀 더 일치시키는 연구를 수행할 계획이다. 이를 위하여 언어별 음소인식에 존재하는 로그 유사도의 편차를 보상해 주는 스코어링 방법에 관한 연구를 진행할 계획이다.

참 고 문 헌

[1] 강효원, 이상필 et al., "음성인식기를 이용한 발음오류 자동분류 결과 분석", *대한음성학회 2003추계학술대회 논문집*, pp.29-32, 2003.

[2] T. Kawahara, T. Kobayashi et al., "Sharable software repository for Japanese large vocabulary continuous speech recognition", *Proc. ICSLP 98*, pp.3257-3260, Sydney, 1998.

[3] 이재강, 권철홍, "청각인상과 음성파형간의 관계 규명을 위한 일본어 /k/의 기초연구", *대한음성학회 봄 학술대회 논문집*, pp.52-55, 2003.

접수일자: 2004년 2월 5일
게재일자: 2004년 3월 15일

▶ 강효원(Hyo-Won Kang)
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2567
E-mail: kanghyowon@hotmail.com

▶ 권철홍(Chul-Hong Kwon) : 책임저자
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2555
E-mail: chkwon@dju.ac.kr