한국음향학회지 제32권 제5호 pp. 438~445 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.5 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.5.438

pISSN : 1225-4428
eISSN : 2287-3775

한국어 음성데이터를 이용한 일본어 음향모델 성능 개선

An Enhancement of Japanese Acoustic Model using Korean Speech Database

이민규†, 김상훈

(Minkyu Lee† and Sanghun Kim)

과학기술연합대학원대학교 컴퓨터소프트웨어 및 공학, 한국전자통신연구원 자동통역인공지능연구센터 자동통역연구실

(접수일자: 2013년 5월 8일; 채택일자: 2013년 7월 5일)

초 록: 본 논문은 일본어 음성인식기 신규 개발을 위해 초기에 부족한 일본어 음성데이터를 보완하는 방법이다. 일본어 발음과 한국어 발음이 유사한 특성을 근거로 한국어 음성 데이터를 이용한 일본어 음향모델 성능개선 방법에 대하여 기술하였다. 이종언어 간 음성 데이터를 섞어서 훈련하는 방법인 Cross-Language Transfer, Cross-Language Adaptation, Data Pooling Approach등 방법을 설명하고, 각 방법들의 시뮬레이션을 통해 현재 보유하고 있는 일본어 음성데이터 양에 적절한 방법을 선정하였다. 기존의 방법들은 훈련용 음성데이터가 크게 부족한 환경에서의 효과는 검증되었으나, 목적 언어의 데이터가 어느 정도 확보된 상태에서는 성능 개선 효과가 미비하였다. 그러나 Data Pooling Approach의 훈련과정 중 Tyied-List를 목적 언어로만으로 구성 하였을 때, ERR(Error Reduction Rate)이 12.8 %로 성능이 향상됨을 확인하였다.

핵심용어: 음성인식, 일본어 인식, 자동통역, 언어적응, Data Pooling, Under-resourced Language

ABSTRACT: In this paper, we propose an enhancement of Japanese acoustic model which is trained with Korean speech database by using several combination strategies. We describe the strategies for training more than two language combination, which are Cross-Language Transfer, Cross-Language Adaptation, and Data Pooling Approach. We simulated those strategies and found a proper method for our current Japanese database. Existing combination strategies are generally verified for under-resourced Language environments, but when the speech database is not fully under-resourced, those strategies have been confirmed inappropriate. We made tyied-list with only object-language on Data Pooling Approach training process. As the result, we found the ERR of the acoustic model to be 12.8 %.

Keywords: Speech recognition, Japanese speech recognition, Automatic speech translation, Language adaptation, Data pooling, Under-resourced language

PACS numbers: 43.72. Ne

I. 서 론

최근 급격한 글로벌화로 인한 국제간 문화, 교육, 관광, 경제적 교류가 활성화 되면서 언어소통의 불편이 대두됨에 따라 자동통역의 수요가 점점 커지고 있다.

자동통역은 통역할 두 언어의 양방향 음성인식 기술, 자동 번역 기술, 그리고 음성 합성 기술이 필요한 고난이도 기술이다. 특히 음성인식 기술의 경우, 신뢰도 있는 음향모델 등 고품질 인식 성능을 확보하기 위해서는 해당 언어의 대용량 음성데이터 확보가 필요하다. 그러나 신규 언어의 음성인식기를 개발 하는 경우, 해당언어의 음성데이터를 새롭게 구축하는데는 많은 시간과 비용이 요구된다.

따라서, 신규 언어에 대한 신뢰도 있는 음향모델링을 하는 방법으로 이미 구축된 타 언어 음성데이터를 이용하는 연구들이 활발히 진행되고 있다.[1] 특히 언어적 특성이 유사한 언어들 간에 음성데이터를

†**Corresponding author:** Min-kyu, Lee (minkyu119@etri.re.kr)
Automatic speech Translation Research Section, Electronics and Telecommunications Research Institute, E218 Gajeongno, Yuseong-gu, Daejeon 305-700, Republic of Korea
(Tel: 82- 42-860-1805, Fax: 82-42-860-4889)

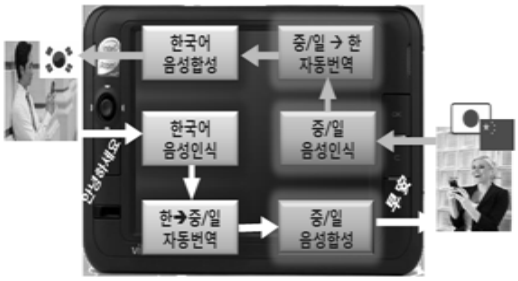


Fig. 1. Automatic speech translation system structure.

공유하여 목적언어의 음향모델 성능을 개선하는 방법에 대해서 의미 있는 결과를 내고 있다.

van Heerden은 남아프리카 소수 민족 언어들과 같이 음성데이터가 부족한 언어의 음향모델을 구성할 때, 유사한 다른 소수 민족 언어의 음성데이터를 이용하여 목적언어의 음향모델 성능을 개선 할 수 있음을 확인하였다.[2] Schultz는 'Global Phone Project'를 통해, 크로아티아어, 독일어, 일본어, 한국어, 포르투칼어, 러시아어, 스페인어, 스웨덴어, 타밀어, 및 터키어 등 10개 국어의 글로벌 음소 세트를 정의하였으며, 또한 각각 언어들을 조합하여 음향모델의 성능을 개선하였다.[3] Carkl는 터키어의 LVCSR(Large Vocabulary Continuous Speech Recognition) 개발에 부족한 터키어 음성데이터를 보충하기 위해 'Global Phone Project'의 다른 언어 음성데이터(15개 언어, 총 100명 화자가 각 20분씩 발화)를 사용하여 성능을 개선하였다.[4] 이 연구들을 통해 음향모델을 구성하려는 목적언어의 음성데이터가 작은 경우(약 100시간 이내) 다른 언어의 음성데이터를 이용하여 성능 개선이 가능함을 확인할 수 있다.

본 논문에서는 한국어와 일본어 모음의 포먼트 값을 대조 분석 하였을 때, 두 언어의 모음들은 국가별 차이점이 거의 나타나지 않다는 점에 두 언어가 유사한다는 근거를 세우고, 일본어 음향모델링 과정에 한국어 음성데이터를 이용하여 일본어 음향모델 성능 개선을 시도하였다.[5]

현재 목적언어인 일본어의 음성데이터를 어느 정도 확보한 상태이지만(약 367시간) 아직 신뢰도 있는 인식성능을 내기에는 부족한 실정이다 . 이에 이러한 조건을 감안하여 위의 논문들에서 제시한 방법들 중 현재 상황에 적절한 방법을 실험하여 그 성능을 검증해 보았다.

II. 일본어 음성인식 베이스라인시스템 구현

한국어 음성데이터를 이용하는 실험에 앞서, 일본어 음성데이터베이스만으로 일본어 음성인식시스템 베이스라인을 구현 하였다. 이를 위한 준비 작업으로 일본어 음소 정의, 기본 인식단위 설정, 발음사전 구축 그리고 형태소 분석을 하였다. 또한 안정적인 비교 실험을 위해 튜닝 값 조절 등을 통해 최적의 일본어 베이스라인 음향모델을 설정하였다.

2.1 일본어 음소 정의

일본어 음소정의는 ETRI(Electronics and Telecommunications Research Institue)에서 정의한 일본어 음소셋을 사용하였다. 하나의 히라가나로 일본어의 각 음소 표현이 불가능하다. 이에 다음과 같이 각 음소의 발음이 발생되는 히라가나를 조합하여 Table 1에 표기하였다. 일본 음성인식 오픈소스인 Julius에서 제공하는 일본어 모델에서 정의한 음소와 큰 차이가 없으며, 총 43개의 음소로 이루어져 있다.

Table 1. ETRI Japanese phone table.

a	あ	gy	ぎょ	py	ぴゃ
e	え	h	へ	r	ら
i	い	hf	ひゅ	ry	りゃ
o	お	hy	ひょ	s	さ
u	う	j	じ	sh	しゃ
a:	あー	jy	じょ	t	た
e:	えー	k	か	ts	つ
i:	いー	ky	きょ	w	わ
o:	おー	m	ま	xb	っぽ
u:	うー	my	みゃ	xg	っか
b	バ	n	な	xs	っさ
by	びゃ	N	んあ	y	や
ch	ちゃ	xm	んば	z	ざ
d	だ	ny	にゃ		
g	が	p	ぱ		

2.2 일본어 언어모델링

인식 실험을 위한 일본어 언어모델은 베이스라인 실험 및 다른 실험 모두에서 동일하다. 언어모델 구축하기 위해서, ETRI에서 제공한 100여만 문장을 정제하여 코퍼스로 사용하였다. 코퍼스는 온라인 메신저 채팅 등 다양한 분야의 대화체로 구성되어 있다.

일본어는 띄어쓰기를 사용하지 않기 때문에 단어별로 n-gram을 구성하기 위하여, 오픈 소스 형태소분석기인 Mecab를 사용하여 각 문장을 형태소 단위로 분리해 낸 후, 인식단위로 다시 후처리 과정을 거쳤다.[6] 그 결과를 다시 SRILIM(The SRI Language Modeling Toolkit)을 사용하여 3-gram 언어모델을 구축하였다.[7] 또한 Mecab 형태소분석기의 세그멘팅 결과를 통해 인식단위에 대한 발음열을 생성할 수 있는 발음변환기(Grapheme to Phoneme Conversion : G2P)를 구현하였다.

2.3 일본어 음향모델링

일본어 음향모델링을 위한 일본어 음성데이터는 ETRI의 스마트폰 환경 일본어 음성데이터 약367시간을 활용하였다. 해당 음성데이터의 수집된 전사문을 앞서 언급한 Mecab를 이용해 생성한 발음변환기를 통하여 모노폰 학습(Monophone Training) 단계를 위한 음소 MLF(Master Lable File)의 발음열로 변환하였다.

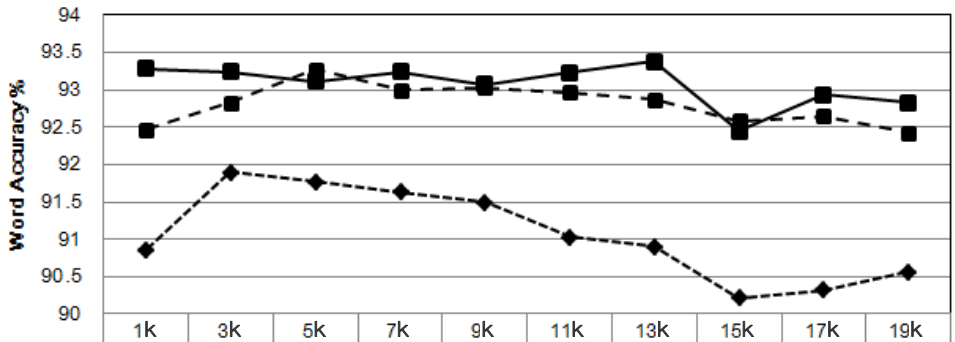
음향모델 훈련은 HTK(Hidden Markov Toolkit)로 이루어졌으며, 음성신호는 20 ms 단위의 Hamming window를 10 ms씩 이동하며 추출하였다. 특징계수로는 MFCC 39차(delta, delta-delta 포함)를 사용하였고, triphone -tying 이후 GMM(Gaussian Mixture Model)의 mixture의 수를 32개까지 늘렸다.

또한, HTK를 통한 음향모델 훈련과정의 tying단계에는 Question Tree Outlier Count 임계치인 RO 값이 존재한다. RO 임계값을 조절함으로써 최종적으로 생성되는 모델의 수를 조절할 수 있고, 생성되는 모델의 수에 따라 인식률의 차이가 발생한다.

이에 본 연구에서는 RO 임계치를 1,000부터 19,000까지 2,000의 등간격으로 조절하며, 전체 367시간 음성데이터 중 36.7시간(1/10), 183.5시간(1/2), 그리고 367시간(전체)에 대하여 인식실험을 통해 RO 임계값의 변화에 따른 음향모델의 성능 변화를 측정하였다.

인식률 측정을 위한 테스트 셋으로는 스마트폰 환경의 720발화(남자 10명, 여자 10명 각 36 발화)로 구성된 ETRI의 일본어 National 평가셋을 사용하였고, 디코더로는 ETRI의 음성인식기를 사용하였다. ETRI의 음성인식기의 구조는 WFST(Weighted Finite State Tranducer) 기반이며 음향모델과 언어모델을 각각 만든 후 하나로 묶어 이미지를 생성한 후 디코더에서 고속으로 인식하는 구조이다.[8]

베이스라인으로 단어 인식률이 가장 높은 모델을 설정하였다. Fig. 2의 인식결과 그래프를 보면, 일본



	1k	3k	5k	7k	9k	11k	13k	15k	17k	19k
36.7 hours	90.86	91.9	91.77	91.63	91.49	91.03	90.9	90.22	90.32	90.57
183.5 hours	92.47	92.83	93.27	93	93.03	92.96	92.87	92.58	92.65	92.42
367 hours	93.28	93.24	93.11	93.24	93.07	93.23	93.38	92.45	92.93	92.83

Fig. 2 . Adjust RO threshold of HTK between 1000 to 19000 for Japanese speech recognition system baseline (x-axis: RO threshold, y-axis: word accuracy, unit: %).

Table 2. Output of Japanese POS tool mecab.

Input	これはテスト文章です	
Output	Recognition Unit	POS information
	これ	名詞,代名詞,一般,*,*,*,これ,コレ,コレ
	は	助詞,係助詞,*,*,*,*,は,ハ,ワ
	テスト	名詞,サ変接続,*,*,*,*,テスト,テスト,テスト
	文章	名詞,一般,*,*,*,*,文章,ブンショウ,ブンショー
	です	助動詞,*,*,*,特殊・デス,基本形,です,デス

어 데이터만을 사용하였을 때 RO 임계값을 11,000으로 설정하였을 때 단어 인식률이 93.38 %로 가장 높게 나타났다. 해당 모델의 GMM과 HMM State 수는 각 2,960개, 4,802개이다.

III. Data Combination Strategies

보유하고 있는 음성데이터의 크기나 상황에 따라, 이종언어를 이용하여 목적언어의 음향모델을 생성하는 방법들을 몇 가지로 분류 할 수 있다. 선행 연구에서 다뤄진 내용 중 다음의 3가지 방법들을 소개한다.

3.1 Cross-language Transfer

기존에 있던 원시언어인 한국어 음향모델을 목적언어인 일본어 음성데이터의 추가 없이 그대로 사용하는 방법이다. 한국어 음향모델을 구축한 후 일본어 언어모델에 결합하는 방법이다.[9] 다만 일본어 발음사전의 발음표기를 한국어 음향모델의 음소 정의와 동일하게 사용해야 한다. 이 방법은 목적언어의 음성데이터 없이도 음향모델을 구성 할 수 있다는 장점이 있다. 그러나 목적언어의 음성데이터가 전혀 반영이 되어 있지 않기 때문에 일정 성능 이상으로 끌어올리기 힘들고, 한국어만으로 음향모델이 구성되어 있기 때문에 일본어의 음성데이터가 확보가 되어있는 상황에서는 적합하지 않다.

3.2 Cross-language Adaptation

목적언어인 일본어 음성데이터가 소량만 확보 되어 있는 경우 적용해 볼 수 있는 방법이다.[10-11] 원시언어인 한국어를 통해 음향모델을 생성한 후 일본어 음성데이터를 MLLR(Maximum Likelihood Linear Regression) 또는 MAP(Maximum A Posteriori)등의 적응 기법을 통해 기존 한국어 음향모델에 적응시키는 방법이다. 일본어의 데이터가 충분치 않을 때 효과적인 방법(10~20시간)이나, 현재와 같이 음성 데이터가 어느 정도 확보된 경우(약 367시간)에는 적응기법이 적절하지 않다.

3.3 Data Pooling

Cross-language Adaptation과 마찬가지로 목적언어인 일본어의 음성데이터가 충분치 않을 경우 적합한 방법이다.[12] 그러나 Data Pooling은 Cross-language Adaptation과 달리, 훈련의 첫 과정부터 한국어의 음성데이터와 일본어의 음성데이터를 함께 훈련 셋에 포함하여 음향모델을 생성하는 방법이다. 이를 위해, 한국어 전사문의 모든 인식 단위들을 한국어 발음변환기를 통해 발음열을 생성한다. 그 후 매핑 테이블이나 룰을 통해 한국어 음소 발음열들을 일본어의 음소들로 바꾼다. 이렇게 모든 훈련 데이터를 일본어 음소로 바꾸고 일반적인 음향모델 훈련과정을 거쳐 이종언어의 음성데이터를 모두 이용한다. 기존에 남아프리카 계통 언어를 통한 선행연구를 보았을 때 유사한 언어 쌍끼리 Data Pooling방법을 통해 음향모델을 구축하였을 때, 그 성능이 향상됨을 확인 할 수 있다.

IV. 한국어와 일본어간 음소매핑

4.1 한국어 음소정의

한국어와 일본어 간 발음이 서로 유사하도록 음소매핑을 수행하기 위해 일본어와 같이 한국어 음소 세트도 정의하였다.

한국어 음소 정의는 일본어 음소와 마찬가지로 ETRI에서 정의한 한국어 음소 셋을 사용하였다. 한국어 음소는 초성과 종성 자음을 구분한 형태의 음소 셋으로, 총 45개의 음소로 이루어져 있다.

Table 3. Korean phone table.

Phone	한글	Phone	한글	Phone	한글
B	ㅃ	i	ㅣ	je	ㅖ
D	ㄸ	k	ㅋ	jo	ㅛ
E	ㅐ	m	ㅁ	ju	ㅠ
G	ㄲ	n	ㄴ	jv	ㅕ
N	종성 ㅇ	o	ㅗ	wE	ㅙ
S	ㅆ	p	ㅍ	wa	ㅘ
U	ㅡ	r	ㄹ	we	ㅞ
Z	ㅉ	s	ㅅ	wi	ㅚ
a	ㅏ	t	ㅌ	wv	ㅝ
b	ㅂ	u	ㅜ	xb	종성 ㅂ
c	ㅊ	v	ㅓ	xd	종성 ㄷ
d	ㄷ	z	ㅈ	xg	종성 ㄱ
e	ㅔ	Wi	ㅟ	xl	종성 ㄹ
g	ㄱ	jE	ㅒ	xm	종성 ㅁ
h	ㅎ	ja	ㅑ	xn	종성 ㄴ

4.2 한국어와 일본어간 음소매핑

앞서 언급한 일본어와 한국어간 음소의 포먼트의 유사함을 일본어 음향모델링 과정에 한국어 음성데이터를 사용해도 된다는 근거로 삼았다. 이를 통해 일본어 전문가의 음성언어학적 자문을 받아, Table 3의 한국어 음소들을 Table 2의 일본어 음소로 변환하는 매핑 규칙을 만들었다. 변환 규칙은 아래 Table 4와 같다.

Table 4를 통해 한국어 음성데이터로부터 훈련 할 수 없는 일본어 음소는 5개의 장모음(a:,e:,i:o:,u:)이다. 그 외의 음소들은 모두 한국어에서 일본어로 음소 매핑이 이루어진다. 다만, 일본어 요음의 경우 한국어 이중모음 음소가 자음 쪽으로 매핑되어 Table 4에서의 정의와 같이 다른 음소들과 다르게 매핑되었다.

Table 4. Korean to Japanese mapping table.

#	Korean Phone ▶ Japanese Phone		#	Korean Phone ▶ Japanese Phone		#	Korean Phone ▶ Japanese Phone	
1	a	a	25	h + u	hf + u	49	p + ju	py + u
2	e	e	26	h + ja	hy + a	50	p + jo	py + o
3	E	e	27	h + ju	hy + u	51	r	r
4	i	i	28	h + jo	hy + o	52	r + ja	ry + a
5	o	o	29	z + i	j + i	53	r + ju	ry + u
6	u	u	30	z + ja	jy + a	54	r + jo	ry + o
7	없음	a:	31	z + ju	jy + u	55	s	s
8	없음	e:	32	z + jo	jy + o	56	s + ja	sh + a
9	없음	i:	33	k	k	57	s + ju	sh + u
10	없음	o:	34	k + ja	ky + a	58	s + jo	sh + o
11	없음	u:	35	k + ju	ky + u	59	t	t
12	b	b	36	k + jo	ky + o	60	Z	ts
13	b + ja	by + a	37	m	m	61	wa	w + a
14	b + ju	by + u	38	m + ja	my + a	62	xb	xb
15	b + jo	by + o	39	m + ju	my + u	63	xg	xg
16	c + ja	ch + a	40	m + jo	my + o	64	xd	xs
17	c + ju	ch + u	41	n	n	65	ja	y + a
18	c + jo	ch + o	42	xn	N	66	ju	y + u
19	d	d	43	xm	xm	67	jo	y + o
20	g	g	44	n + ja	ny + a	68	z + a	z + a
21	g + ja	gy + a	45	n + ju	ny + u	69	z + u	z + u
22	g + ju	gy + u	46	n + jo	ny + o	70	z + e	z + e
23	g + jo	gy + o	47	p	p	71	z + o	z + o
24	h	h	48	p + ja	py + o			

V. Data Combination 실험 및 결과

앞서 소개한 기법들 중 현재 확보된 일본 음성데이터 상황의 특성을 고려하였을 때 Data Pooling 방법이 가장 적합하다는 판단을 하였다. Data Pooling 방법을 응용하여 총 3가지의 실험을 수행하였으며, 각 실험들의 결과를 Fig. 2에서 설정한 일본어 베이스라인 음향모델과 비교 평가하였다.

3가지 실험에 사용된 일본어 음성데이터는 모두 동일하나, 아래 Fig. 3의 일반적인 음향모델 훈련절차에서 한국어 음성데이터가 반영되는 절차상의 위치와 그 방식이 다르다. 또한 훈련에 사용되는 한국어 음성데이터의 크기도 실험의 특성에 따라 서로 다르다.

5.1 Data Pooling Approach I

한국어 음성데이터에 대하여 한국어 음향모델 과정을 거쳐 한국어 모노폰 모델에 대하여 align된 한국어 음소로 구성된 음소 MLF를 생성하였다. 생성된 MLF파일을 Table 4를 이용하여 일본어 음소로 변환하였다. 한국어 음소 중 일본어 음소로 매핑 되지 않는 경우도 음향적으로 최대한 유사한 음소로 매핑하여[예 : v(한국어 '어') → o(일본어 '오')] 모든 한국어 트레이닝 목록을 일본어 트레이닝 셋에 포함하였다. 이후 Fig. 3의 Monophone Training After realignment부터 해당 한국어 리스트를 포함하여 일본어 음향모델 훈련을 진행하였다. 훈련에 사용된 일본어 음성데이터는 약 367시간, 한국어 음성데이터는 약 493시간이다. 전체적으로 음향모델 훈련에 사용된 데이터의 크기가 커지기 때문에 같은 Triphone의 발생 횟수가 늘어나, RO 임계값이 같을 경우 최종 생성되는 state수가 늘어난다. 마찬가지로 32 mixture까지 확장하였으며, 최종 GMM과 HMM state수는 5,204개, 10,129개 이다.

5.2 Data Pooling Approach II

위 실험에서 성능 하락의 원인을 한국어 음소 중 일본어 음소에 존재하지 않는 음소들을 강제로 일본어 음소에 매핑 한 것에 기인한다 판단하고[예 : v(한국어 '어') → o(일본어 '오')] , 강제 매핑에 해당되는 음성데이터는를 제외하고 실험을 하였다. 한국어 음성데이터를 통해 한국어 음향모델링 과정을 거쳐 한국어 모노폰 모델에 대하여 align된 한국어 음소 MLF를 생성하였다. Data Pooling Approach I과는 달리 음소 전사 내용 을 Table 4의 과정을 거쳐 매핑시킨 후, 매핑테이블을 통해 변환되어지지 않는 음소를 전사문에 포함하고 있는 음성데이터는 제외하고, 완벽히 매핑변환이 이루어진 전사문들을 선별하였다. 이후 Fig. 3의 Monophone Training After realignment 과정에서 선별된 한국어 음성 리스트를 포함하여 일본어 음향모델링을 진행하였다. 일본어 DB 약 367시간, 한국어 DB는 약 39시간이 훈련에 사용되었다. 32 mixture까지 확장하였으며, 최종 GMM과 HMM state 개수는 3,244개, 5,866개 이다.

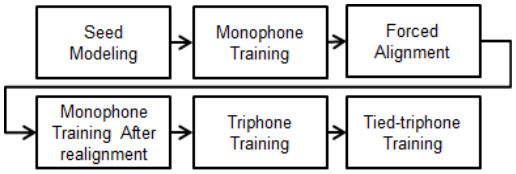


Fig. 3. Acoustic model training process.

5.3 Data Pooling Approach III

Data Pooling Approach II를 통해 변환 가능한 전사문만을 훈련에 포함시켰음에도 불구하고 기존 베이스라인 성능에 미치지 못하였다. 한국어 음성데이터의 음소분포가 일본어 데이터의 음소 분포와는 다른 것을 그 원인으로 찾을 수 있었다. 그에 따른 Tied-list의 변화를 억제하기 위한 실험을 진행하였다.

Data Pooling Approach II와 동일하게 한국어 음성데이터 리스트를 선별하였다. 그러나 이후 Data Pooling Approach II와는 달리, Fig. 3의 트라이폰 학습(Triphone Training)까지는 일본어 음성데이터만으로 진행을 하였다. 일본어 음성데이터로만 훈련된 트라이폰 음향모델로부터 Question-Tree를 Tying하여 triphone tying에 한국어 음성데이터의 영향을 차단하였다. 그리고 Tied-Triphone Training 과정에서만 선별된 한국어 음성 리스트를 추가하여 훈련하였다. 32 mixture까지 확장하였으며, 최종 GMM과 HMM state

수는 2,960, 4,802 개다. Tied-Triphone Training 과정에서만 한국어 음성데이터를 추가하였기 때문에 GMM과 HMM state 수는 일본어 음성데이터만으로 훈련을 진행한 베이스라인 음향모델과 동일하다.

5.4 성능 평가

각 실험을 통하여 생성한 모델들을 ETRI의 National, TgMobile 두 가지 평가셋으로 인식 성능을 측정하였다. 이미 언급한 National 평가셋은 스마트폰 채널의 남자 10명, 여자 10명의 각 36발화씩 총 720발화로 구성되어 있으며, 여행 상황에서의 발생 가능한 발화들로 이루어져 있다. TgMobile 평가셋은 스마트폰 채널의 남자 25명, 여자 25명 각 40발화씩 총 2,000발화로 구성되어 있다.

Table 5에서 알 수 있듯이, Data Pooling Approach I, II, III실험에 대해 베이스라인과 비교평가를 해보았을 때 Data Pooling Approach III의 ERR이 베이스라인 대비 National 평가셋에 대하여 12.8 % , TgMobile 평가셋에 대하여 4.95 %로 성능이 향상되었다.

VI. 결 론

본 논문에서는 일본어 인식 실험을 위한 음소 세트 정의, 일본어 발음사전 생성, 인식단위 선정, 언어모델링, 그리고 음향모델링을 수행하여 일본어 음성인식 시스템을 구성해 보았다. 또한 정교한 일본어 베이스라인 설계를 위해, 일본어 음향모델 훈련 과정에서 조절할 수 있는 RO 임계값 조정을 통한 인식률의 변화를 살펴보았다. 그 결과를 통해 일본어 베이스라인 음향모델을 설정하였다.

이후 일본어 음향모델링 과정에 있어 일본어 음성데이터의 부족한 부분에 대한 훈련을 보충하기 위해 발음상 유사한 한국어 음성데이터를 추가하여 성능을 개선하였다.

단순히 한국어 음성데이터를 일본어 음성데이터에 추가하여 훈련한 경우, 기존 367시간의 일본어 음성데이터로 훈련된 음향모델보다 성능이 하락됨을 확인하였다.

이는 한국어 음성데이터 추가가 Question-tree를 통해 생성되는 트라이폰 리스트에 부정적 효과로 작용된다고 할 수 있다. 한국어 음소 전사문에서는 자주 나타나는 트라이폰 구조지만 일본어의 언어적 특성으로 인하여 일본어 음성데이터에서 그다지 나타나지 않는 트라이폰 구조의 경우에 있어, 트라이폰 발생 횟수로 인해 본래 일본어 음성데이터만으로 훈련되었을 때와는 다르게 트라이폰 분기가 일어난다. 그로 인해 탐색공간에 부정확하게 모델링된 트라이폰 모델이 존재하게 되어 성능이 하락하는 것으로 해석된다.

이를 막기 위하여 일본어 음성데이터만으로 Seed Modeling , 모노폰 학습, 트라이폰 학습까지 진행한 뒤, tied-triphone training 과정에서만 한국어 음성데이터를 추가하였다. 그 결과, 해석에서 예측한 문제가 해결됨을 확인하였다. Question-tree를 통해 생성되는 트라이폰 리스트는 일본어 데이터를 훈련시켰을 때와 동일하지만 , 이 후 한국어 음성데이터의 추가로 인해 음향모델이 변화함을 보였으며, 그 변화로 인해 인식 성능이 향상됨을 확인하였다.

현재 한국어 음성데이터 가운데 음성 파일에 대한 음소 전사가 완벽하게 일본어 음소로 변환되는 경우만 훈련에 반영하고 있지만, 향후 모든 한국어 음성데이터를 반영할 예정이다. 선택적으로 반영된 한국어 음성데이터뿐만 아니라 모든 한국어 음성데

Table 5. Recognition experiment results of Japanese acoustic model baseline, Data Pooling Approach I, Data Pooling Approach II, Data Pooling Approach III.

	# of hmm	# of gmm	Word accuracy(%)		Error Reduction Rate(%)	
			National	TgMobile	National	TgMobile
Japanese Baseline	2,960	4,802	93.38	87.89	0	0
Data Pooling Approach I	5,204	10,129	91.33	86.21	-	-
Data Pooling Approach II	3,244	5,866	92.44	86.75	-	-
Data Pooling Approach III	2,960	4,802	94.23	88.43	12.8	4.95

이터의 일본어 음소로 변환되는 부분을 반영 할 수 있다면, 데이터 보충의 측면에서 장점이 있다. 또한 앞으로 중국어, 스페인어, 프랑스어등 신규 언어로의 음성인식기 확장에 있어, 이종언어의 음성데이터 활용은 초기 부족한 음성데이터베이스를 보충 및 여러 면에 응용분야가 있을 것으로 기대된다.

References

1. Ulla Uebler, "Multilingual speech recognition in seven languages," Speech. Commun. **35**, 53-69 (2001).
2. C. van Heerden, N. Kleynhans, E. Barnard, and M. Davel, "Pooling ASR data for closely reslated languages," in Proc. SLTU, 17-23 (2010).
3. Tanja Schultz and Alex Waibel, "Language Portability in Acoustic Modeling," Speech. Commun. **10**, 59-64 (2000).
4. Kenan Çarki, Petra Geutner, and Tanja Schultz, "Turkish LVCSR: toward better speech recognition for agglutinative languages," In Proc. ICASSP, 1563-1566 (2000).
5. J. K. Lee, "Comparative study on korean and japanese formant values by korean speakers and japanese speakers," Ono Yongu Studies in Linguistics. **15**, 61-75 (1997)
6. *Mecab: Yet Another Part-of-Speech and Morphological Analyzer*, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html, 2013.
7. *The SRI Language Modeling Toolkit,* http://www.speech.sri.com/projects/srilm, 2013.
8. S. H. Kim, I. Lee, and J. Park, "Developing fast/light korean recognizer through building FST_based search network," in Proc. KSCSP, **25**, 1, 21-24 (2008).
9. A. Constantinescu, and G. Chollet, "On cross-language experiments and data-driven units for ALISP," In Proc. ASRU, 606-613 (1997).

저자 약력

▸ 이 민 규(Minkyu Lee)

2011년 8월: 경북대 컴퓨터공학 학사
2011년 9월 ~ 현재: UST 컴퓨터 소프트웨어공학 재학

▸ 김 상 훈(Sanghun Kim)

1990년 2월: 연세대학교 전기공학 학사
1991년 2월: KAIST 전기전자공학 석사
2004년 3월: 동경대 전기전자공학 박사
1992년 3월 ~ 현재: 한국전자통신연구원 자동통역연구실 실장