pISSN 2005-8063
eISSN 2586-5854
2018. 12. 31.
Vol.10 No.4
pp. 111-119

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2018.10.4.111

형태소 발음변이를 고려한 음성인식 단위의 성능*

Performance of speech recognition unit considering morphological pronunciation variation

방 정 욱 · 김 상 훈 · 권 오 욱**

Bang, Jeong-Uk · Kim, Sang-Hun · Kwon, Oh-Wook

Abstract

This paper proposes a method to improve speech recognition performance by extracting various pronunciations of the pseudo-morpheme unit from an eojeol unit corpus and generating a new recognition unit considering pronunciation variations. In the proposed method, we first align the pronunciation of the eojeol units and the pseudo-morpheme units, and then expand the pronunciation dictionary by extracting the new pronunciations of the pseudo-morpheme units at the pronunciation of the eojeol units. Then, we propose a new recognition unit that relies on pronunciation by tagging the obtained phoneme symbols according to the pseudo-morpheme units. The proposed units and their extended pronunciations are incorporated into the lexicon and language model of the speech recognizer. Experiments for performance evaluation are performed using the Korean speech recognizer with a trigram language model obtained by a 100 million pseudo-morpheme corpus and an acoustic model trained by a multi-genre broadcast speech data of 445 hours. The proposed method is shown to reduce the word error rate relatively by 13.8% in the news-genre evaluation data and by 4.5% in the total evaluation data.

Keywords: pronunciation variation, decoding unit, pseudo-morpheme, Korean LVCSR

1. 서론

한국어 대어휘 연속 음성인식(large vocabulary continuous speech recognition, LVCSR)을 위한 음성인식 단위로는 주로 의사형태소 단위(pseudo-morpheme, 방정욱·권오욱, 2014; Kwon *et al.*, 1999)를 사용한다. 의사형태소 단위는 어절 단위보다 적은 수의 인식 어휘로 다양한 단어를 표현할 수 있으며, 음절 단위보다 평균 지속시간이 길어서 넓은 문맥을 고려할 수 있다. 또한, 형태소 단위와는 다르게 발음이 유지되면서 길이가 짧은 단음소가 제거되고 높은 빈도의 형태소들이 병합되어 한국어 음성인식 단위(Kwon & Park, 2003)로 많이 사용된다.

의사형태소를 음성인식 단위로 사용하기 위해서는 형태소 내부와 형태소 경계에서 발생하는 발음변이 현상을 고려해야 한다. 의사형태소의 발음은 특히 인접한 형태소에 따라 단단한 형태학적 규칙에 지배를 받는다. 이러한 이유로 형태소 단위에서 추출된 발음은 어절 단위에서 얻어진 발음과 종종 다른 발음

* 본 논문은 한국전자통신연구원 연구운영비지원사업의 일환으로 수행되었음. (18ZS1140, Conversational AI 공통핵심기술 연구).
** 충북대학교, owkwon@cbnu.ac.kr, 교신저자
Received 29 August 2018; Revised 8 October 2018; Accepted 9 October 2018

(Kwon *et al*., 1999)을 가질 수 있다.

기존의 연구에서는 언어학적 지식을 토대로 구축된 발음열 자동 생성기(Jeon *et al*., 1998; Lee & Chung, 2004)를 사용하여 형태소에서 발생 가능한 다양한 발음변이를 발음사전의 다중발음(강병옥, 2003)으로 다루었다. 다중발음이 확장된 발음사전은 변이된 발음을 고려하여 음성인식 성능을 향상시키는데 기여한다. 하지만, 너무 많은 다중발음이 추가된다면 인식 과정에 혼란을 도래하여 성능을 하락시킨다(Lee & Chung, 2004).

본 논문에서는 발음사전에서 형태소 발음변이를 고려하기 위해서, 어절 단위 말뭉치를 토대로 의사형태소 단위의 새로운 발음열을 추출하고, 기존의 인식 단위를 발음에 따라 세분화함으로써 발음이 고려된 새로운 단위를 제안한다. 제안된 방법은 발음사전을 확장하여 형태소의 다양한 발음변이를 고려하면서, 인식 단위를 세분화하여 다중발음이 많이 추가되었을 때 발생하는 음성인식 성능 하락 문제를 감소시키는 효과를 가진다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 발음 생성 방법을 소개하고, 3장에서는 형태소 발음변이가 고려된 새로운 발음열과 인식단위 생성 방법을 설명한다. 4장에서는 새로운 발음이 추가된 발음사전과 제안된 단위로 학습된 언어모델의 성능을 확인한다. 마지막으로 5장에서 결론을 맺는다.

2. 기존 연구

한국어 대어휘 연속 음성인식에서는 의사형태소 단위를 음성인식 단위(Kwon & Park, 2003)로 많이 사용한다. 여기서 의사형태소 단위는 형태소 단위를 음성인식에서 사용하기 적합하도록 수정한 단위를 말한다. 형태소 단위는 형태소 분석기로 분할하는 과정에서 발음열의 변화가 발생할 수 있으며 길이가 짧은 단음소나 단음절이 빈번하게 나타나기 때문에, 음성인식에 그대로 사용하기 어렵다. 따라서 기존의 연구에서는 형태소 분석기를 수정하여 발음을 유지하면서 빈도가 높은 단어들을 병합시킨 의사형태소 단위를 주로 사용한다. 본 절 이후로 우리는 편의상 의사형태소를 “형태소”로 부른다.

형태소는 인접한 형태소에 따라 다양하게 발음이 변화하는 특성을 가진다(정민화·이경님, 2004). 이러한 발음변이 현상은 주로 복합명사나 조사, 접미사, 그리고 어미 등의 결합에 의해 생겨난다. 따라서 음성인식 단위로 형태소를 사용하기 위해서는 다양한 발음변화 현상에 대한 고려가 필요하다. 기존의 연구에서는 단어의 발음변화 현상을 음성인식기에 반영하기 위해서 크게 두 가지 접근법을 사용하였다. 첫 번째 방법으로 발음사전 측면에서 각 단어의 변이된 발음을 다중발음으로 추가하는 명시적(explicit) 접근 방법(정민화·이경님, 2004)이 존재하며, 두 번째 방법으로 음향모델 측면에서 의사결정 트리(decision tree)를 사용하여 비슷한 발음을 공유하거나 묶는 묵시적(implicit) 접근 방법(Young *et al*., 1994)이 존재한다. 본 논문에서는 발음사전에 다중발음을 추가하는 명시적 접근 방법에 중점을 두고자 한다.

발음사전은 언어모델에 존재하는 각 단어들이 음향모델에서 어떠한 발음으로 구성되는지를 보이는 중요한 모델이다. 발음사전은 각 단어에 대해 하나의 대표 발음을 가지며, 각 단어의 발음은 언어학적 지식을 기반으로 구축된 규칙 기반의 발음열 자동 생성기를 사용하여 다양한 다중발음과 함께 생성된다(Jeon *et al*., 1998). 최근에는 언어학적 지식을 사용하지 않고 음성 데이터를 기반으로 각 단어의 발음을 추정하고 발음사전을 구축하는 접근법이 제안되고 있다(Razavi & Magimai.-Doss, 2015).

명시적 접근 방법에 관한 다양한 연구에서 다중발음을 발음사전에 반영하는 경우에 시스템의 성능이 향상된다는 사실은 이미 잘 알려져 있다(강병옥, 2003). 음성은 매 순간 변화하는 특성이 있기 때문에 발음사전에서 각 단어가 가질 수 있는 다중발음은 매우 광범위하다. 앞서 언급한 형태소의 발음변이 이외에도 발화 환경이나, 화자의 연령, 성별, 사투리 등의 원인으로 쉽게 변화될 수 있다. 이러한 상황에서 변이된 발음들을 다중발음으로 발음사전에 반영시킨다면, 음성인식 과정에서 단어의 변이된 발음이 고려되어 해당 단어가 인식되는 긍정적인 효과를 기대할 수 있다. 하지만, 너무 많은 다중발음이 발음사전에 추가되는 경우에는 탐색 네트워크의 혼잡도가 증가되어 인식 성능을 하락시키는 부정적인 영향 또한 함께 야기할 수도 있다.

발음사전에서 적절한 개수의 다중발음을 정하는 것은 중요하다. 기존 연구에서는 말뭉치에서 각 단어들이 나타나는 빈도를 고려하여 높은 빈도로 나타나는 단어만을 인식어휘로 사용하거나(Lee & Chung, 2004), 형태소 내부나 형태소 경계에서 나타나는 음운 변화 현상을 선택적으로 적용하여 다중발음을 발음사전에 추가하는 방법을 사용하였다(정민화·이경님, 2004; Lee & Chung, 2003). 하지만, 발음사전에서 제거된 단어나 다중발음들은 탐색 네트워크에서도 완전히 제거되기 때문에 변이된 발음을 전혀 고려하지 못하게 된다. 따라서 예상 가능한 모든 발음변이를 고려하면서, 탐색 과정에서 발생하는 혼잡도를 감소시키는 방법이 제안될 필요가 있다.

3. 제안된 방법

본 논문에서는 다양한 발음변이를 고려하면서 탐색 혼잡도를 감소시키기 위해서 그림 1과 같이 새로운 단위를 제안한다.

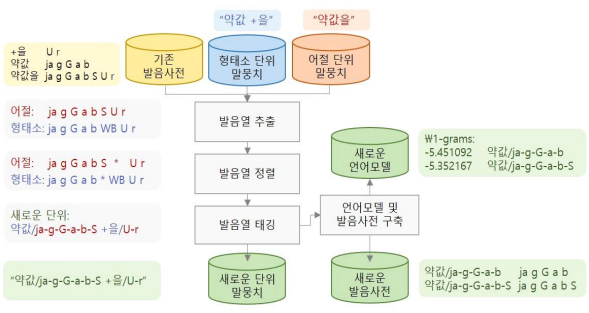


그림 1. 제안된 단위 생성을 위한 블록도
Figure 1. Block diagram for the proposed unit generation method

제안된 방법은 어절 단위와 형태소 단위의 두 말뭉치로부터 발음열을 추출하고, 어절 단위 발음열을 기준으로 두 발음열을 정렬한다. 이후, 형태소 발음변이가 고려된 새로운 발음을 어절 단위 발음으로부터 추출하고 이들을 형태소 단위 뒤에 붙임으로써 발음이 고려된 새로운 단위의 말뭉치를 구축한다. 제안된 단위는 발음사전과 언어모델을 구축하는데 사용된다.

3.1. 발음열 추출 단계

어절 단위와 형태소 단위로 구성된 두 말뭉치로부터 각 단어의 발음열을 추출한다. 그림 2는 어절 단위 말뭉치의 예시와 이들을 형태소 분석기를 이용하여 형태소 단위로 분할한 말뭉치이다. 이때, 생성된 형태소 단위는 발음열이 유지되면서 고빈도 단어가 결합된 의사형태소 단위를 의미하며, 이후 어절 단위로 복원 가능하도록 별도의 기호('+')가 삽입된다.

장 담그는 날이 다가온다
엄마의 손맛을 담은 맛 간장에 10분 뚝딱 양념 된장까지
잘 만든 장 하나로 입은 물론 마음까지 감동시킨다
약값을 우대해 신약 개발을 촉진시키고
신소재 산업과 에너지 산업도 집중 육성하기로 했습니다
당장은 이란과 중국 시장에 주력할 방침입니다

장 담그는 날 +이 다가온다
엄마 +의 손맛 +을 담은 맛 간장 +에 10분 뚝딱 양념 된장 +까지
잘 만든 장 하나로 입 +은 물론 마음 +까지 감동 +시킨다
약값 +을 우대 +해 신약 개발 +을 촉진 +시키고
신소재 산업 +과 에너지 산업 +도 집중 육성 +하기 +로 했습니다
당장 +은 이란 +과 중국 시장 +에 주력 +할 방침 +입니다

그림 2. 어절 단위 말뭉치(위)와 형태소 단위 말뭉치(아래)
Figure 2. Text corpora in the word unit (above) and morpheme unit corpus (below)

서로 다른 단위로 구성된 두 말뭉치는 미리 생성해둔 발음사전을 사용하여 발음열로 변환되며, 단어 경계(word boundary)를 구분하기 위해서 별도의 기호('WB')를 삽입한다. 어절과 형태소의 발음은 여러 다중발음 중에서 어떤 발음이 선택되는지 알기 어렵기에 각 단어의 대표발음을 사용하여 변환하였다.

형태소 단위에서 생성된 발음은 인접한 형태소에 따라 형태학적 규칙(정민화·이경님, 2004)에 영향을 받기 때문에 종종 어절 단위에서 생성된 발음과 다른 발음을 가질 수 있다. 그림 3은 발음사전을 이용하여 어절 단위와 형태소 단위로 구성된 말뭉치를 변환시킨 발음열을 나타낸다.

WB z a N WB d a m g U n U n WB n a r i WB ... WB
WB v m m a Wi WB s o n m a s U r WB ... WB
WB z a r WB m a n d U n WB z a N WB ... WB
WB j a g G a b S U r WB u d E h E WB s i n j a g WB ... WB
WB s i n s o z E WB s a n v b G wa WB ... WB
WB d a N z a N U n WB i r a n g wa WB ... WB

WB z a N WB d a m g U n U n WB n a r WB i WB ... WB
WB v m m a WB Wi WB s o n m a d WB U r WB ... WB
WB z a r WB m a n d U n WB z a N WB ... WB
WB j a g G a b WB U r WB u d E WB h E WB ... WB
WB s i n s o z E WB s a n v b WB g wa WB ... WB
WB d a N z a N WB U n WB i r a n WB g wa WB ... WB

그림 3. 어절 단위에서 얻어진 발음열(위)과 형태소 단위에서 얻어진 발음열(아래)
Figure 3. Pronunciation sequences obtained from the word unit (above) and the morpheme unit (below)

어절 단위의 단어 "약값을"에서는 발음열 "j a g G a b S U r"을 가지는 반면에, 이들의 형태소 분석 결과인 "약값 +을"에서는 발음열 "j a g G a b"과 "U r"을 가진다. 어절 단위 말뭉치는 발음기호 'S'가 존재하는 것과 다르게, 형태소 단위로 분할된 말뭉치에서는 해당 발음기호가 출력되지 않는다. 표 1은 각 음소 기호에 상응하는 발음과 예시를 나타낸다.

표 1. 음소 기호 별 발음 및 예시
Table 1. Pronunciation and examples by phoneme symbols

발음 기호	음소	예시	발음 기호	음소	예시
B	/ㅃ/	뿐 [B u n]	ju	/ㅠ/	타율 [t a j u r]
D	/ㄸ/	갖다 [g a d D a]	jv	/ㅕ/	열기 [jv r g i]
E	/ㅐ/	올래 [o r r E]	k	/ㅋ/	키위 [k i wi]
G	/ㄲ/	볼까 [b o r G a]	m	/ㅁ/	만두 [m a n d u]
N	/ㅇ/	사랑 [s a r a N]	n	/ㄴ/	나라 [n a r a]
S	/ㅆ/	했어 [h E S v]	o	/ㅗ/	소리 [s o r i]
U	/ㅡ/	스시 [S U s i]	p	/ㅍ/	파도 [p a d o]
Wi	/ㅢ/	의식 [Wi s i g]	r	/ㄹ/	라면 [r a m jv n]
Z	/ㅉ/	질주 [z i r Z u]	s	/ㅅ/	사람 [s a r a m]
a	/ㅏ/	사과 [s a g wa]	t	/ㅌ/	타임 [t a i m]
b	/ㅂ/	바른 [b a r U n]	u	/ㅜ/	우주 [u z u]
c	/ㅊ/	채소 [c E s o]	v	/ㅓ/	선수 [s v n s u]
d	/ㄷ/	다리 [d a r i]	wE	/ㅙ/	쇄도 [s wE]
e	/ㅔ/	할게 [h a r G e]	wa	/ㅘ/	화구 [h wa g u]
g	/ㄱ/	가자 [g a z a]	we	/ㅚ/	회식 [h we s i g]
h	/ㅎ/	하자 [h a z a]	wi	/ㅟ/	위기 [wi g i]
i	/ㅣ/	이다 [i d a]	wv	/ㅝ/	워드 [wv d U]
jE	/ㅒ/	얘기 [jE g i]	z	/ㅈ/	자라 [z a r a]
ja	/ㅑ/	약사 [ja g S a]	sil	묵음	
je	/ㅖ/	계산 [g je s a n]	#num	없음	#0 [], #1 []
jo	/ㅛ/	요가 [jo g a]			

다음 단계에서는 형태소 경계에 나타나는 발음변이 현상이 고려된 대표발음을 얻기 위해서 어절 단위의 발음열과 형태소 단위의 발음열을 서로 정렬한다.

3.2. 발음열 정렬 단계

발음열 정렬 단계에서는 어절 단위에서 생성된 발음열과 형태소 단위에서 생성된 발음열을 어절 단위를 기준으로 정렬한다. 정렬 알고리듬은 일반적으로 많이 사용되는 문자열 정렬 알고리듬인 Levenshtein alignment (Povey, 2016)을 개선하여 사용하였다.

일반적인 Levenshtein alignment 알고리듬은 두 발음열 사이의 거리 값이 최소가 되는 정렬 결과를 탐색하기 위해서 누적 거리 값이 기록된 탐색 테이블을 생성한다. 탐색 테이블은 먼저 아래의 식 (1)을 사용하여 발음 간의 거리를 계산하고, 식 (2)를 사용하여 재귀적으로 누적 거리 값을 기록한다.

$$score(p_{eoj}^{i}, p_{mor}^{j}) = \begin{cases} 0 & \text{if } p_{eoj}^{i} = p_{mor}^{j} \\ 1 & otherwise \end{cases} \quad (1)$$

$$lev(i,j) = \min \begin{cases} lev(i-1,j-1) + score(p^i_{eoj}, p^j_{mor}) \\ lev(i-1,j) + 1 \\ lev(i,j-1) + 1 \end{cases} \quad (2)$$

여기서, '$score(p^i_{eoj}, p^j_{mor})$'는 어절 단위의 발음열 '$p_{eoj}$'와 형태소 단위의 발음열 '$p_{mor}$'에서 각 '$i$'번째, '$j$'번째 발음 기호 '$p^i_{eoj}$', '$p^j_{mor}$' 사이의 거리 값을 의미한다. 또한, '$lev(i,j)$'는 두 발음열로 생성된 탐색 테이블에서 '$i$'번째와 '$j$'번째 발음 기호까지의 누적 거리 값을 의미하며, 테이블의 대각/위쪽/왼쪽의 누적 거리 값에 두 발음 사이의 거리 값을 더하거나 삽입/삭제 페널티를 더한 값 중에서 가장 작은 값을 선택하여 기록한다.

형태소 분석기로 분할된 형태소 단위는 어절 단위 문장보다 더 많은 단어 경계 기호를 가진다. 각 단위의 경계를 표현하기 위해서 발음열 추출 단계에서 단위 경계 기호 "WB"를 삽입하였다. 이러한 상황에서 기존의 문자열 정렬 알고리듬을 그대로 사용한다면, 형태소 단위의 발음열에 존재하는 단어 경계 기호가 어절 발음열에 존재하는 다른 발음 기호로 빈번하게 정렬되어 사라지는 현상이 발생할 수 있다.

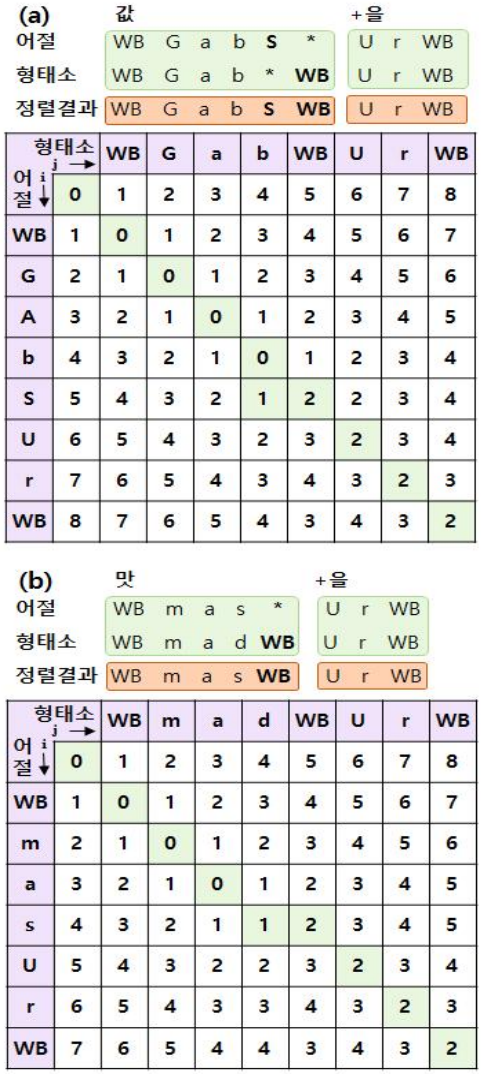
단어 경계 기호는 어절 단위 발음열에서 형태소 단위 경계를 찾는데 요구되는 중요한 기호이다. 어절 단위의 발음열에서 형태소 단위의 변이된 발음을 추출하는 과정에서 단어 경계 기호가 사라진다면, 이후 발음열 태깅 단계에서 형태소와 그들의 발음을 서로 연결하는데 어려움이 발생한다. 따라서 거리 계산 수식을 아래의 식 (3)과 같이 수정하여 형태소의 단어 경계 기호가 발견될 때 높은 거리 값을 가지도록 변경하였다.

$$score(p^i_{eoj}, p^j_{mor}) = \begin{cases} 0 & \text{if } p^i_{eoj} = p^j_{mor} \\ 3 & p^j_{mor} = WB \\ 1 & otherwise \end{cases} \quad (3)$$

단어 경계 기호의 거리 값으로 '1'을 사용하는 경우에는 기존의 알고리즘과 동일하게 동작하며, 거리 값으로 '2'를 사용하는 경우에는 어절의 발음 기호가 삽입되거나 치환될 때 발생하는 누적 거리 값인 '2'와 동일하여 여전히 형태소 단위의 단어 경계 기호가 사라지는 문제가 발생하였다. 따라서 우리는 형태소 단위의 단어 경계 기호 거리 값으로 '3'을 사용하였다. 그림 4는 단어 "값을"과 "맛을"을 제안된 방법으로 정렬한 예시이다.

먼저, 어절 단위의 "값을"은 발음열 변환 단계에서 "WB G a b S U r WB"로 변환되며, 이들의 형태소 단위인 "값 +을"은 "WB G a b WB U r WB"로 변환된다. 기존 방법으로 이들을 정렬할 경우에는 어절 단위 발음에 존재하는 발음 기호 'S'와 형태소 단위 발음열에 존재하는 단어경계기호 'WB'가 서로 치환되어 "WB G a b S U r WB"를 정렬 결과로 출력하며, 이후 형태소 경계를 파악하기가 어렵다.

반면에 제안된 방법으로 정렬할 경우에는 발음기호 'S'와 단어경계 기호 'WB'를 삭제/삽입된 단어로 인지하여 "WB G a b S WB U r WB"을 정렬 결과로 출력한다. 얻어진 정렬 결과는 형태소 단위의 단어경계 위치정보를 가지면서 형태소 단위 발음보다 더 정확한 어절 단위의 발음 정보를 가진다. 제안된 정렬 방법은 서로 다른 길이를 가지는 어절 "맛을"과 형태소 "맛 +을"의 정렬 예시에서도 단어경계기호가 유지되면서 발음변이가 고려된 발음 기호를 가진다.



어절 \ 형태소	WB	G	a	b	WB	U	r	WB
	0	1	2	3	4	5	6	7	8
WB	1	0	1	2	3	4	5	6	7
G	2	1	0	1	2	3	4	5	6
A	3	2	1	0	1	2	3	4	5
b	4	3	2	1	0	1	2	3	4
S	5	4	3	2	1	2	2	3	4
U	6	5	4	3	2	3	2	3	4
r	7	6	5	4	3	4	3	2	3
WB	8	7	6	5	4	3	4	3	2

어절 \ 형태소	WB	m	a	d	WB	U	r	WB
	0	1	2	3	4	5	6	7	8
WB	1	0	1	2	3	4	5	6	7
m	2	1	0	1	2	3	4	5	6
a	3	2	1	0	1	2	3	4	5
s	4	3	2	1	1	2	3	4	5
U	5	4	3	2	2	3	2	3	4
r	6	5	4	3	3	4	3	2	3
WB	7	6	5	4	4	3	4	3	2

그림 4. 발음열 정렬 예시 (a) "값을" (b) "맛을"
Figure 4. Examples of pronunciation sequence alignment

3.3. 발음열 태깅 단계

발음열 태깅 단계에서는 단어 발음이 고려된 새로운 단위의 말뭉치를 생성한다. 발음열 정렬 결과로부터 형태소 단위 말뭉치에서의 단어경계 위치를 찾을 수 있다. 발음열 정렬 과정에서 형태소 단위의 단어경계기호를 유지시켰다. 정렬 결과로 얻어진 단어 경계 기호 사이의 각 발음열은 형태소 단위 말뭉치의 형태소로 일대일로 연결 가능하다. 정렬 결과로 얻어진 새로운 발음열은 연결기호('–')를 사용하여 서로 이어 붙이고, 형태소 단위의 각 단어 뒤에 부착하였다. 그림 5는 제안된 단위로 재구성된 말뭉치의 예시이다.

장/z-a-N 담그는/d-a-m-g-U-n-U-n ... 다가온다/d-a-g-a-o-n-d-a
엄마/v-m-m-a +의/Wi 손맛/s-o-n-m-a-s ... +까지/G-a-z-I
잘/z-a-r 만든/m-a-n-d-U-n 장/z-a-N ... +시킨다/s-i-k-i-n-d-a
약값/ja-g-G-a-b-S +을/U-r ... +시키고/s-i-k-i-g-o
신소재/s-i-n-s-o-z-E ... 했습니다/h-E-d-S-U-m-n-i-d-a
당장/d-a-N-z-a-N +은/U-n 이란/i-r-a-n ... +입니다/i-m-n-i-d-a

그림 5. 제안된 단위의 말뭉치
Figure 5. Text corpus in the proposed unit

제안된 단위를 발음사전에 적용하는 경우에는 형태소 경계에서 다양하게 변이된 발음들이 반영되는 효과를 기대할 수 있다. 기존의 발음사전에서 형태소 단위의 단어 "약값"은 발음 [약깝]을 의미하는 발음열 "ja g G a b"만을 단일발음으로 가지는 반면에, 제안된 방법에서는 [약까], [약깝], [약깜], [약값]을 나타내는 "ja g G a", "ja g G a b", "ja g G a m", "ja g G a b S"의 다양한 다중발음들을 얻을 수 있었다. 각 발음들은 어절 "약값하고", "약값도", "약값만", "약값을" 등에서 얻어진 발음열로 기존의 발음 "ja g G a b" 이외에 3가지 다중발음이 추가된 것이다.

기존의 형태소 단위는 단어의 실제 발음을 예상할 수 없다. 따라서 변이된 발음을 모두 고려하기 위해서는 하나의 단어 "약값"에 4개의 다중발음을 모두 할당해야 한다. 이러한 방법은 탐색 과정에서 높은 혼잡도를 가지게 되며 인식성능 하락을 유발한다. 더욱이, 각 단어의 다중발음 개수를 축소시킨다면 제거된 다중발음이 탐색 과정에 반영되지 못하여 인식성능을 하락시킬 수 있다. 반면에, 제안된 단위는 각 단어가 어떤 변이된 발음을 가질지 미리 예측할 수 있다. 제안된 단위 "약값/ja-g-G-a"는 변이된 발음열 "ja g G a"를 가지며, "약값/ja-g-G-a-b"은 변이된 발음열 "ja g G a b"를 가지는 것을 예상할 수 있다. 따라서 제안된 단위로 구축된 발음사전은 다중발음에 의한 탐색 혼잡도를 줄일 수 있으며, 말뭉치로부터 얻어진 다양한 다중발음을 발음사전에 모두 반영시킬 수 있다.

제안된 단위를 언어모델에 적용하는 경우에는 발음에 따라 단어들이 세분화되어 탐색 과정에서 혼잡도가 감소되는 효과를 기대할 수 있다. 기존의 형태소 단위로 언어모델을 구축하는 경우에는 단어 "약값"의 빈도가 다음에 나타나는 단어 "+하고", "+도", "+만", "+을"의 발생 확률을 모델링하기 위해 동일하게 사용된다. 반면에 "약값/ja-g-G-a", "약값/ja-g-G-a-b", "약값/ja-g-G-a-m", "약값/ja-g-G-a-b-S" 등의 제안된 단위를 사용하여 언어모델을 구축한다면 발음열에 따라 세분화된 단어로부터 다음 단어가 나타날 확률을 모델링하게 되어 인식 성능 향상에 기여할 것으로 예상된다.

그림 6은 기존의 형태소 단위와 제안된 단위를 사용하여 구축된 단어 "약값"의 탐색 네트워크를 보인다. 실제로 본 실험에서 사용한 음성인식기의 탐색 네트워크는 최적화(optimization) 과정에서 그래프 형태로 변환되지만, 이해를 돕기 위해서 본 그림에서는 경로 최적화 과정이 생략된 네트워크를 나타내었다. 기존의 방법에서는 모든 다중발음이 발음사전에 반영되어 16가지의 복잡한 탐색 경로를 가졌다. 반면에, 제안된 단위에서는 단어의 변이된 발음을 미리 예상하여 4가지의 간소한 탐색 경로를 가졌다.

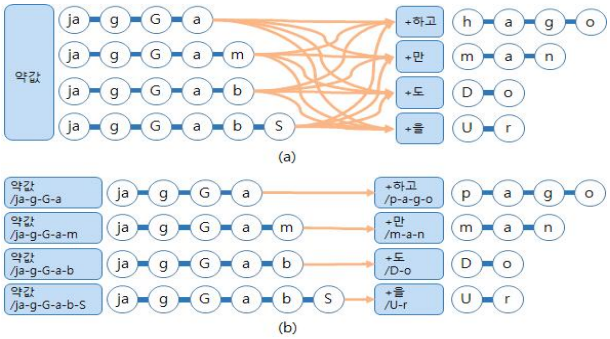


그림 6. 탐색 네트워크 (a) 형태소 단위 (b) 제안된 단위
Figure 6. Search networks constructed with (a) morpheme unit and (b) proposed unit

4. 실험 및 결과

4.1. 학습 데이터베이스

음향모델 학습에 사용된 음성 데이터베이스는 2016년 2월에서 3월까지 방송된 방송 오디오에서 자막 텍스트와 자막 타임스탬프(time stamp)를 이용하여 추출한 445시간의 음성 데이터를 사용하였다(Bang & Choi, 2017). 음성 데이터베이스는 16 kHz, 16 bits로 샘플링 되었으며, 다수의 화자들이 다양한 환경에서 녹음한 음성들로 구성된다.

제안된 단위 생성과 언어모델을 구축하는데 사용된 한국어 말뭉치는 2016년 2에서 2017년 3월까지 방송된 방송 자막을 사용하였다. 인식단위 생성에 사용된 자막 말뭉치는 약 86M개의 어절을 가지며, 이들을 형태소 분석기에 입력하여 얻어진 말뭉치에서는 약 117M개의 형태소를 가졌다. 발음열 추출 실험에 불필요한 문장 기호와 특수 문자는 제거하고 사용하였다.

4.2. 평가 데이터베이스

음성인식 실험에 사용된 평가 데이터는 2016년 4월에 방송된 뉴스, 어린이, 시사, 드라마, 예능의 5가지 장르를 가지는 방송 데이터를 사용하였다. 평가 데이터는 학습 데이터와 다른 날에 방송된 프로그램으로 구축하였지만, 방송 프로그램의 특성상 학습 데이터와 동일한 화자를 가질 수 있다.

장르별 특성으로, 뉴스 데이터는 평균 발화 길이가 길면서 잡음이 없는 환경에서 낭독체 스타일의 발화 특성을 가진다. 어린이 데이터는 선생님이 아이들에게 공룡에 대해 설명하는 내용의 방송으로 뉴스 데이터와 유사하게 정형화된 문장을 차분하게 발성하는 특징을 보인다. 하지만, "트리케라톱스" 등의 공룡 이름이나 "고생물학자"와 같이 언어모델에 낮은 빈도로 나타나는 어휘들이 많이 존재한다. 시사 데이터는 리포터가 사극 드라마를 소개하는 내용의 방송으로, 배우 이름이나 지명 등의 고유명사가 빈번하게 나타나고, 주인공과의 인터뷰 과정에서 웃음소리 등의 잡음이 나타나는 특징을 보였다. 드라마 데이터는 드라마 중간에 배우들의 감정표현이 강조된 발화들이 나타나며, 약간의 배경음악이 혼합되는 특징을 보였다. 마지막으로, 예능 데이터는 평균 발화 길이가 짧으면서 효과음이나 배경음악 등

의 다양한 잡음들이 혼합되었으며 자유로운 스타일의 발화 특성을 가졌다. 평가에 사용된 음성 데이터는 1,559개의 발화로 구성된 전체 2시간 23분의 길이를 가지며, 학습 음성 데이터와 독립적으로 구축하여 사용하였다.

4.3. 실험환경

모든 음성인식 실험은 Kaldi toolkit (Povey *et al*., 2011)을 사용하여 수행되었다. 음성인식기에 사용된 특징벡터로는 발화 단위로 평균과 분산을 정규화한 40차 로그 멜 필터뱅크를 사용하였으며, 문맥을 고려하기 위해서 좌우 7개 프레임을 연결하여 총 15개 프레임을 음향모델 입력으로 사용하였다.

음향모델은 깊은 신경망(deep neural network, DNN)과 은닉 마르코프 모델(hidden Markov model, HMM)을 사용한 DNN-HMM 하이브리드 방식을 사용하였다. HMM은 비묵음 기호와 묵음 기호에 대해 각 3개와 5개의 상태 열을 가지는 left-to-right HMM을 사용하였다. DNN은 15×40 차원의 입력층과 tanh 활성함수를 사용하는 1,024 차원의 은닉층 6개, softmax 활성함수를 사용한 8,033 차원의 출력층을 사용하였다. 전체 데이터는 학습에 15번 사용되며, 초기 10번까지는 0.001부터 0.01까지 학습률을 감소시키며 학습하였으며, 마지막 5번은 0.01의 고정된 학습률로 DNN 모델을 학습하였다(Povey, 2018). 언어모델은 SRILM 도구(Stolcke, 2002)를 사용하여 형태소 단위 말뭉치와 제안된 단위 말뭉치에 Kneser-Ney discounting 방법을 적용한 3-gram 모델을 사용하였다(Kneser & Ney, 1995). 발음사전 및 형태소 분석기는 한국전자통신연구원에서 제작된 도구를 사용하였다. 디코더는 음향모델 가중치를 0.077을 가지며, 탐색 빔 크기는 10으로 설정하였다. 인식 성능은 형태소 단위의 단어 오류율(word error rate, WER)을 계산하여 확인하였다.

4.4. 실험결과

4.4.1. 형태소 단위 인식실험

먼저 각 단어의 대표 발음으로 구성된 단일발음 사전과 형태소 내부에서 발생하는 음운변화 규칙을 반영한 다중발음 사전을 사용하여 음성인식 성능을 확인하였다. 실험은 약 1억 형태소 단위 말뭉치로 언어모델을 구축한 열린 평가(open test) 실험과, 평가용 데이터의 정답 텍스트로 언어모델을 학습시킨 닫힌 평가(closed test) 실험으로 나누어 수행하였다.

표 2는 단일 발음 및 다중 발음을 이용한 음성인식 성능 비교표이다. 열린 평가 실험에서 단일발음 인식실험은 60만 인식 어휘를 가지는 언어모델과 그들의 대표 발음으로 구성된 발음사전이 사용되며, 다중발음 인식실험은 형태소 내부에서 발생하는 발음변이가 반영되어 단일발음 사전보다 약 1.4배 확장된 82만개의 어휘 개수를 가졌다.

음성인식 결과에서는 잡음이 없으면서 낭독체 스타일로 발성된 뉴스 데이터에서 가장 낮은 단어 오류율을 보였고, 다양한 잡음 환경을 가지며 대화체 스타일로 발화된 예능 데이터에서 가장 높은 단어 오류율을 보였다. 다중발음을 사용한 실험에서는 단일발음을 사용한 실험과 미미한 성능 차이를 보였다. 다중발음이 고려됨으로써 나타나는 성능개선 효과는 전사 텍스트로 구축된 언어모델을 사용하는 닫힌 평가에서 두드러지게 나타났다. 우리는 제안된 방법의 성능을 비교하기 위해 형태소 내부에 발생하는 발음변이를 고려한 다중발음 실험을 베이스라인(‘Baseline’)으로 사용하였다.

표 2. 발음사전 유형 및 평가 방법에 따른 WER (%) 비교
Table 2. Comparison of WER (%) of the single pronunciation and the multiple pronunciation according to evaluation methods

장르	열린 평가		닫힌 평가	
	단일발음	다중발음	단일발음	다중발음
뉴스	6.7	5.8	1.4	1.2
어린이	23.7	23.7	6.5	5.8
시사	32.1	31.8	9.9	8.9
드라마	26.0	26.7	7.1	6.9
예능	60.6	60.6	32.2	32.1
전체	22.5	22.1	8.2	7.8

4.4.2. 제안된 단위 인식실험

제안된 방법의 성능을 확인하기 위해서 3가지 실험을 수행하였다. 첫 번째 실험(‘PronLM’)으로 기존의 다중 발음사전을 사용하되 제안된 단위로 언어모델을 구축함으로써 제안된 단위가 언어모델에 미치는 효과를 확인하였다. 다음으로 두 번째 실험(‘PronPM’)에서는 기존의 형태소 단위의 언어모델을 사용하되 발음사전에서 형태소 경계에서 발생하는 변이된 발음을 반영하여 확장된 다중발음의 효과를 확인하였다. 마지막으로, 제안된 단위로 언어모델을 구축하고, 발음사전에 적용한 실험(‘PronLMPM’)을 수행하여 음성인식 성능을 확인하였다. 표 3은 각 실험에 사용된 언어모델(‘LM’)의 크기와 발음사전(‘PM’)의 전체 어휘 개수를 보이며, 표 4와 표 5는 제안된 단위에 따른 열린 평가와 닫힌 평가 실험의 음성인식 성능을 보인다.

표 3. 실험별 언어모델 및 발음사전 크기 비교
Table 3. Comparison of step-by-step language model and dictionary size

변수		Baseline	PronLM	PronPM	PronLMPM
언어모델	1 gram	0.60M	0.62M	0.60M	0.62M
	2 gram	11.8M	12.1M	11.8M	12.1M
	3 gram	14.5M	14.5M	14.5M	14.5M
	단위	형태소	제안	형태소	제안
발음사전	개수	0.82M	0.84M	0.83M	0.62M
	규칙	내부	내부	내부+경계	내부+경계

첫 번째 실험에서 우리는 발음이 고려된 언어모델의 성능을 확인하기 위해서 제안된 단위로 생성된 언어모델과 기존의 발음사전을 사용하여 음성인식 성능을 비교하였다. 표 3에서 ‘Baseline’ 실험과 ‘PronLM’ 실험의 발음사전 단어 개수 차이는 제안된 단위로 세분화됨에 따라 인식 어휘 개수가 증가하여 발생한 차이이며, ‘Baseline’ 실험에 사용되었던 다중발음 이외에 새로운 발음이 추가된 것이 아니다.

제안된 단위로 언어모델을 구축한다면 단어의 발음 정보가

반영되어 음성인식 성능 향상에 기여할 것이라고 기대하였다. 하지만, 제안된 단위로 구축된 언어모델에서는 형태소 단위로 구축된 언어모델에 비해서 약 2만 개의 어휘만이 확장되었으며, 평가 데이터의 복잡도(perplexity)를 비교한 별도의 실험에서도 각 203.5와 204.0으로 미미한 차이를 보였다. 제안된 단위가 언어모델에 미치는 영향이 미미하다보니 각 장르별 음성인식 결과에서도 두드러진 개선 효과는 나타나지 않았다.

두 번째 실험에서는 언어모델의 기본 단위로 'Baseline' 실험과 동일한 형태소 단위를 사용하되, 발음사전에 형태소 내부와 경계에서 변이되는 발음을 모두 사용하였다. 확장된 발음사전은 형태소 내부에서 발생하는 발음변이만 고려하는 'Baseline' 실험의 발음사전보다 약 2만개의 다중 발음이 추가되었다.

다중발음을 확장한 실험에서는 긍정적인 영향과 부정적인 영향을 예상할 수 있다. 긍정적인 영향으로는 형태소 경계에서의 발음변이를 고려하지 못하여 잘못 인식되던 단어가 새로운 발음열을 추가함으로써 올바르게 인식시킬 수 있다. 예를 들어, 기존의 'Baseline' 실험에서는 단어 "것"의 발음열로 오로지 "G v d"만이 고려된다. 이들로 구축된 음성인식기는 "것 +이라고"의 올바른 발음열인 "g v s i r a g o"를 탐색 과정에서 고려하지 못하기 때문에, "거 시 +라고"라는 잘못된 인식 결과를 도출한다. 반면에, 다중발음을 확장한 실험에서는 단어 "것"의 다중 발음으로 "g v s"가 고려되어 올바른 음성인식 결과를 기대할 수 있다.

부정적인 영향도 예상할 수 있다. 발음사전의 확장은 발음변이가 고려되지 않던 단어들에게는 변이된 발음을 탐색 과정에 새롭게 고려시켜 긍정적인 영향을 미치지만, 이미 올바르게 인식되던 단어 측면에서는 탐색 네트워크만 크게 확장되어 음성인식 성능에 부정적인 영향을 미칠 수 있다. 확장된 탐색 네트워크는 다양한 발음의 단어를 인식할 수 있는 높은 자유도를 가지지만 탐색 과정에 너무 많은 인식 후보를 가지므로 음성인식 성능을 하락시킬 수 있다.

다중발음의 확장이 음성인식 성능에 기여하는 정도는 평가 데이터 내에서 다중발음 확장으로 인하여 긍정적인 영향을 받는 단어와 부정적인 영향을 받는 단어의 개수에 영향을 받는다. 표 4의 열린 평가 실험 결과로 미루어 볼 때 어린이 장르에서는 발음변이가 고려되지 않았던 단어들이 추가된 다중발음에 의해서 음성인식 성능이 개선된 것으로 나타나며, 예능 장르 평가 데이터의 경우에는 원했던 긍정적인 영향보다 부정적인 영향이 크게 작용하여 음성인식 성능이 하락하였다.

장르 별 평가 데이터에서 긍정적 영향을 받는 단어의 효과는 표 5의 닫힌 평가 실험에서 두드러지게 확인할 수 있다. 닫힌 평가 실험은 평가 데이터의 정답 텍스트로 언어모델을 구축하기 때문에 음향적인 성능 차이를 비교하는데 유용하다. 뉴스나 어린이 데이터의 경우에는 형태소 발음변이가 존재하는 정형화된 문장들이 많이 존재하여 높은 성능 향상을 보였다. 하지만 짧은 문장의 자유발화로 구성된 시사, 드라마, 예능 데이터의 경우에는 미미한 성능 변화를 보였다. 이 결과로부터 우리는 다중발음 확장에 따른 부정적 영향을 줄일 수 있다면 뉴스 및 어린이 데이터에서 성능이 크게 향상될 것을 예상할 수 있다.

세 번째 실험에서는 제안된 단위를 음성인식 단위로 사용하고 이들로 학습된 언어모델과 수정된 발음사전을 사용하였다. 수정된 발음사전은 'PronPM' 실험과는 다르게 그림 5의 (b)와 같이 각 단어에 태깅된 발음만을 사용하였다. 예를 들어, 단어 "약값/ja-g-G-a"의 경우에는 태깅된 발음열 "ja g G a" 만을 발음사전에 가지며, 단어 "약값/ja-g-G-a-b"의 경우에는 발음열 "ja g G a b"만 발음사전에 존재하도록 설정하였다.

'PronPM' 실험에서 단어 "약값"은 4개의 다중발음을 가지지만, 'PronLMPM' 실험에서 단어 "약값/ja-g-G-a"은 1개의 발음만을 가진다. 'PronPM' 실험에서는 형태소 단위로 구성된 각 단어에 다수의 다중발음을 할당함으로써 긍정적인 영향과 부정적인 영향을 모두 보이는 반면에 제안된 단위를 사용하는 경우에는 인식 단위로부터 변이된 발음을 미리 예상할 수 있기 때문에, 형태소 발음변이를 고려함으로써 얻어지는 긍정적인 효과는 유지하면서 다수의 다중발음을 가짐으로써 발생하는 부정적인 영향을 감소시킬 수 있다.

표 4. 인식 단위에 따른 열린 평가에서의 WER (%) 비교
Table 4. WER (%) in the proposed methods in the open test

장르	Baseline	PronLM	PronPM	PronLMPM
뉴스	5.8	5.9	6.0	**5.8 (0.0)**
어린이	23.7	23.7	23.1	**21.3 (10.1)**
시사	31.8	32.2	34.7	**31.6 (0.6)**
드라마	26.7	26.4	29.8	**25.4 (4.9)**
예능	60.6	60.4	62.9	**60.0 (1.1)**
전체	22.1	22.1	23.3	**21.6 (2.3)**

표 5. 인식 단위에 따른 닫힌 평가에서의 WER (%) 비교
Table 5. WER (%) in the proposed methods in the closed test

장르	Baseline	PronLM	PronPM	PronLMPM
뉴스	1.2	1.2	0.8	**0.7 (41.7)**
어린이	5.8	5.8	**5.0**	5.1 (12.1)
시사	8.9	8.9	9.1	**8.7 (2.2)**
드라마	6.9	**6.5**	6.6	6.7 (2.9)
예능	32.1	32.0	32.3	**31.9 (0.6)**
전체	7.8	7.8	7.6	**7.4 (5.1)**

4.4.3. 제안된 단위의 다중발음 확장실험

'PronLMPM' 실험 결과를 분석해보니 조사 "+의"를 "+에"로 잘못 인식하는 경우가 빈번하게 나타났다. 이는 우리가 형태소 발음변이를 고려한 단일발음만을 사용하였기 때문이다. 각 단어의 발음은 형태학적 규칙에 따라 변이되는 것 이외에도 주변 환경에 따라 다양하게 변이되어 읽힐 수 있다. 특히, 모음 "+의"의 경우에는 인접한 형태소에 상관없이 편의에 따라 [의]나 [에]로 읽힌다. 이러한 이유로 우리는 'PronLMPM' 실험의 발음사전에 'Baseline' 실험의 다중발음을 추가하여 음성인식 성능을 확인하였다. 표 6은 약 86만개의 발음으로 확장된 실험('PronLMPM+')의 음성인식 성능을 보인다.

표 6. 다중발음 확장에 따른 제안된 단위의 WER (%) 비교
Table 6. WER (%) in the proposed unit with multiple pronunciations

장르	Baseline	PronLMPM	PronLMPM+
뉴스	5.8	5.8	**5.0 (13.8)**
어린이	23.7	21.3	**20.8 (12.2)**
시사	31.8	31.6	**30.9 (2.8)**
드라마	26.7	**25.4**	25.8 (3.3)
예능	60.6	**60.0**	60.2 (0.6)
전체	22.1	21.6	**21.1 (4.5)**

제안된 방법은 모든 데이터에서 베이스라인 실험보다 높은 음성인식 성능을 보였으며, 'Baseline' 실험에 비해서 4.5% (=(22.1 – 21.6)/22.1×100)의 상대적 단어 오류율 감소 효과를 보였다. 더욱이 우리가 'PronPM' 실험의 결과에서 예상한대로, 뉴스 데이터와 어린이 데이터에서 각 13.8%, 12.2%의 높은 상대적 단어 오류율 감소 효과를 보였다.

5. 결론

본 논문에서는 형태소의 다양한 발음변이 현상을 음성인식에 반영시키기 위하여, 형태소 내부와 경계에서 다양하게 변이된 발음들을 데이터로부터 추출하고, 이를 의사형태소에 부착시켜 발음이 고려된 새로운 음성인식 단위를 제안하였다. 다양한 장르의 방송 데이터를 이용하여 음성인식 실험을 수행한 결과에서 제안된 방법은 4.5%의 상대적 단어 오류율이 감소되는 효과를 보이며, 뉴스와 어린이 데이터에서 각 13.8%와 12.2%의 높은 상대적 단어 오류율 감소 효과를 보였다.

본 논문에서 제안된 방법은 탐색 네트워크의 혼잡도를 크게 증가시키지 않으면서, 다양하게 변이되는 의사형태소의 발음들을 발음사전과 언어모델에 반영하여 전체적인 한국어 음성인식 성능을 향상시키는 효과를 가진다.

참고문헌

Bang, J. U., & Kwon, O. W. (2014). Performance of pseudomorpheme-based speech recognition units obtained by unsupervised segmentation and merging. *Phonetics and Speech Sciences*, 6(3), 155-164. (방정욱·권오욱 (2014). 비교사 분할 및 병합으로 구한 의사형태소 음성인식 단위의 성능. *말소리와 음성과학*, 6(3), 155-164.)

Bang, J. U., Choi, M. Y., Kim, S. H., & Kwon, O. W. (2017). Improving speech recognizers by refining broadcast data with inaccurate subtitle timestamps. *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH'17)*. Stockholm, Sweden. August 20-24, 2017.

Chung, M. H., & Lee, K. N. (2004). Modeling cross-morpheme pronunciation variations for Korean large vocabulary continuous speech recognition. *Malsori*, 49, 107-121. (정민화·이경님 (2004). 한국어 연속음성인식 시스템 구현을 위한 형태소 단위의 발음변화 모델링. *말소리*, 49, 107-121.)

Jeon, J., Cha, S., Chung, M., Park, J., & Hwang, K. (1998). Automatic generation of Korean pronunciation variants by multistage applications of phonological rules. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*. Sydney, Australia. November 30-December 4, 1998.

Kang, B. O. (2003). A study on the multiple pronunciation dictionary for spontaneous speech recognition. *Proceedings of the 2003 Conference of the Korean Society of Speech Sciences* (pp. 65-68). (강병옥 (2003). 대화체 연속음성인식을 위한 확장 다중발음 사전에 관한 연구. *한국음성학회 2003 학술대회논문집*, 65-68.)

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *Proceedings of the 1995 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP'95)*. Detroit, USA. May 9-12, 1995.

Kwon, O. W., & Park, J. (2003). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, 39(3-4), 287-300.

Kwon, O. W., Hwang, K. W., & Park, J. (1999). Korean large vocabulary continuous speech recognition using pseudomorpheme units. *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*. Budapest, Hungary. September 5-9, 1999.

Lee, K. N., & Chung, M. (2004). Pronunciation lexicon modeling and design for Korean large vocabulary continuous speech recognition. *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH'04)*. Jeju, Korea. October 4-8, 2004.

Lee, K. N., & Chung, M. H. (2003). Statistical analysis of Korean pronunciation variations. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*. Barcelona, Spain. August 3-9, 2003.

Povey, D. (2016). Align-text algorithm. Retrieved from https://github.com/kaldi-asr/kaldi/blob/master/src/bin/align-text.cc on July 1, 2018.

Povey, D. (2018). Neural-network training script. Retrieved from https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/nnet2/train_block.sh on July 1, 2018.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU'11)*. Hawaii, USA. December 11-15, 2011.

Razavi, M., & Magimai.-Doss, M. (2015). An HMM-based formalism for automatic subword unit derivation and pronunciation generation. *Proceedings of the 2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP'15)*. Brisbane, Australia. April 19-24, 2015.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH'02)*. Denver, USA. September 16-22, 2002.

Young, S. J., Odell, J. J., & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *Proceedings of the Workshop on Human Language Technology (HLT'94)*. Plainsboro, USA. March 8-11, 1994.

• **방정욱 (Bang, Jeong-Uk)**
충북대학교 제어로봇공학전공 박사과정 재학 중
충북 청주시 서원구 충대로 1(개신동)
Email: jubang@cbnu.ac.kr
관심분야: 음성인식, 음성정렬, 음성 데이터 정제

• **김상훈 (Kim, Sang-Hun)**
한국전자통신연구원 책임연구원
대전 유성구 가정로 218
Email: ksh@etri.re.kr
관심분야: 음성인식, 자동통역

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자공학부 교수
충북 청주시 서원구 충대로 1(개신동)
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 음성신호처리, 오디오신호처리