대화체 연속음성인식을 위한 확장 다중발음 사전에 관한 연구

강병옥
한국전자통신연구원 음성처리연구팀

A Study on the Multiple Pronunciation Dictionary for Spontaneous Speech Recognition

ByungOk Kang
Spoken Language Processing Team, ETRI
bokang@etri.re.kr

Abstract

본 논문에서는 대화체 연속음성인식 과정에서 사용되는 다중발음사전의 개념을 확장하여 대화체 발화에 빈번하게 나타나는 불규칙한 발음변이 현상을 포용하도록 한 확장된 발음사전의 방법을 적용하여 대화체 연속음성인식에서 인식성능의 향상을 가져오게 됨을 실험을 통해 보여준다.

대화체 음성에서 빈번하게 나타나는 음운축약 및 음운탈락, 전형적인 오발화, 양성음의 음성음화 등의 발음변이는 언어모델의 효율성을 떨어뜨리고 어휘 수를 증가시켜 음성인식의 성능을 저하시키고, 또한 음성인식 결과로 나타나는 출력형태가 정형화되지 못하는 단점을 가지고 있다. 이에 이러한 발음변이들을 발음사전에 수용할 때 각각의 대표어휘에 대한 변이발음으로 처리하고, 언어모델과 어휘사전은 대표어휘만을 이용해 구성하도록 한다. 그리고, 음성인식기의 탐색부에서는 각각의 변이발음의 발음열도 탐색하되 대표어휘로 언어모델을 참조하도록 하고, 인식결과를 출력하도록 하여 결과적으로 인식성능을 향상시키고, 정형화된 출력패턴을 얻도록 한다.

본 연구에서는 어절단위 뿐 아니라 의사형태소[2] 단위의 발음사전에도 발음변이를 포용하도록 하여 실험을 하였다. 실험을 통해 어절단위의 다중발음사전 구성을 통해 ERR 10.9%, 의사형태소 단위의 다중발음사전의 구성을 통해 ERR 4.3%의 성능향상을 보였다.

I. 머리말

음성인식 기술은 1960년대 연구가 본격적으로 시작된 이래 모음인식, 고립단어 인식의 단계를 넘어 현재는 대어휘의 연속음성인식의 단계로 접어들어 활발히 연구가 진행되고 있다. 낭독체 연속음성인식은 방송뉴스인식을 필두로 연구가 활발히 진행되고 있고 제한된 영역에서 가시적인 성과를 거두고 있다.

하지만, 대화체 음성은 화자가 발화할 때 발음이나 문법적 오류를 비교적 적게 포함하고 있는 낭독체 음성과는 달리 화자의 자연스런 발화로 인해 그 자유도가 높아 대화체 연속음성인식을 위해서는 처리해야 할 문제가 많다.[1] 즉, 실제 대화를 살펴보면 알 수 있듯이 대화체 음성에서는 사투리어휘, 거친 호흡이나 입술소리 등 화자가 만들어내는 잡음, '아', '음' 등의 간투어, 발화 반복, 발화중 머뭇거림, 발화 수정 등의 비문법적인 발화요소, 표준발음과 다른 발음변이 등을 포함하고 있어, 대화체 연속음성인식기의 성능을 떨어뜨리는 요인으로 작용한다.

이중 표준발음과 다른 발음변이는 '가르쳐'를 '갈쳐'로 발음하는 음운축약 및 탈락, '가고요'를 '가구여', '하셨고요'를 '하셨구여' 로 발음하는 양성음의 음성음화, '어떻게'를 '어트케'로 발음하는 전형적인 오발화 등의 형태로 나타난다. 대화체 음성에서 나타나는 어러한 발음변이는 언어모델의 효율성을 떨어뜨리고 어휘 수를 증가시켜 대화체 연속음성인식의 성능을 떨어뜨릴 뿐 아니라, 음성인식 결과가 정형화된 형태로 출력되지 못해 음성인식 결과를 이용해서 자동통역 등의 응

용으로 활용될 경우 문제를 야기하게 된다.

이에 본 논문에서는 대화체 연속음성에서 빈번하게 나타나는 이러한 불규칙한 발음변이를 현상을 포용하도록 한 확장된 다중 발음사전의 방법을 이용하여 한국어 대화체 연속음성인식기의 성능을 향상시킬 수 있음을 실험을 통해 보여준다. 특히 인식단위를 의사형태소 단위로 했을 경우의 발음사전에 앞서의 대화체 음성의 발음변이를 포용하도록 의사형태소 기반 다중 발음사전을 구성하였고, 그 실험 결과를 보여준다.

II. 확장된 다중발음사전

2.1 대화체 연속음성인식 시스템

<그림 1>은 일반적인 대화체 연속음성인식 시스템의 구성도이다. 입력된 음성은 특징추출부(101)에서 인식에 유용한 정보만을 추출한 특징벡터로 변환되고, 이러한 특징벡터로부터 탐색부(102)에서 훈련데이터를 통한 학습과정에서 미리 구축된 음향모델 데이터베이스(104)와 발음사전 데이터베이스(105), 언어모델 데이터베이스(106)를 이용하여 가장 확률이 높은 단어 열을 찾게된다. 마지막으로, 인식결과 출력부(103)는 탐색부(102)의 출력을 이용하여 인식결과를 제공한다.

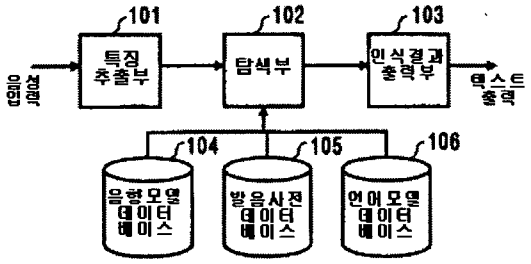
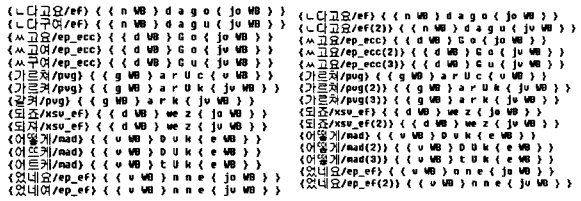


그림 1. 일반적인 대화체 연속음성인식 시스템

2.2 확장된 다중발음사전의 적용

확장된 다중발음사전을 적용했을 경우 위 <그림 1>에서 탐색부(102), 발음사전 데이터베이스(105), 언어모델 데이터 베이스(106)에 변경을 준다.

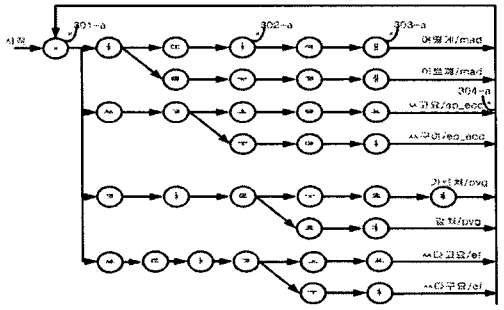
즉, 발음사전 데이터베이스(105)를 구성할 때 훈련데이터에서 나타난 음운 축약, 음운 탈락, 전형적인 오발화, 양성음의 음성음화 등을 각각의 대표어휘에 대한 변이발음으로 처리하고, 언어모델 데이터베이스(106)와 어휘사전은 해당어휘의 대표어휘만을 이용해 구성한다. 그리고, 탐색부(102)에서는 각각의 변이발음의 발음열도 탐색하되 대표어휘로 언어모델을 참조하도록 하고, 최종적으로 대표어휘를 이용하여 인식결과를 출력하도록 한다.



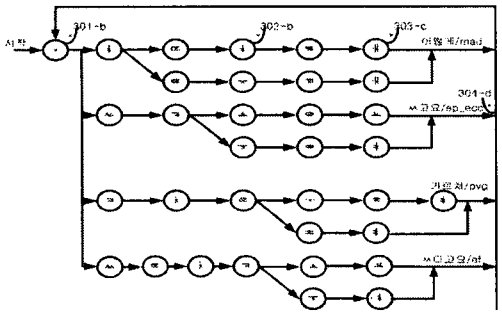
(a) 일반 발음사전 (b) 확장 다중발음사전

그림 2. 발음사전 비교(의사형태소 단위)

<그림 2-a> 는 일반적인 의사형태소 기반 발음사전 데이터베이스(105)의 예이다. 각 의사형태소 단위의 표준발음과 발음변이에 해당하는 발음이 모두 표제어에 나타남을 볼 수 있다. 즉, '어떻게/mad', '어뜨케/mad', '어트케/mad'가 모두 표제어로 나타난다. <그림 2-b>는 <그림 2-a>에 대응되는 확장 다중발음사전 데이터베이스(105)의 예이다. <그림 2-a>와 달리 해당 표제어의 발음변이가 있을 경우 표제어에 괄호가 붙은 형태로 대표어 뒤에 따라온다.



(a) 일반 트리탐색 예시도



(b) 다중발음사전 트리탐색 예시도

그림 3. 트리탐색 예시도 비교(의사형태소 단위)

<그림 3>은 <그림1>의 탐색부(102)의 예시도로서 종래의 발음사전과 언어모델을 이용할 경우(a)와 다중발음사전과 해당 언어모델을 이용할 경우(b)의 트리탐색의 예시도이다. <그림 3>을 보면 처음 시작 혹은 한 어휘가 결정된 후 탐색경로는 모두 하나의 가상적인 Root노드(301)에 연결된 형태를 갖는다. 이후에 음성입

력이 들어오면 매 프레임마다 트리의 모든 노드에서의 확률 값을 계산한 후에, 각 노드로 들어오는 천이들 중에 가장 확률이 높은 천이만을 남긴다. 탐색을 진행하여 Leaf 노드(303)에 도달해 어휘가 결정되면, Leaf 노드(304)에서 Root 노드(301)로의 천이는 단어가 변경되므로 어휘간의 연결에 통계적인 형태의 언어모델(105)이 적용된다. <그림 3-a>와 같은 종래의 발음사전은 대표어에 대한 각각의 발음변이를 모두 개별적인 어휘로 처리하여, 언어모델 역시 변이발음 각자에 대해 통계를 내야 하므로 언어모델의 효율성이 떨어지게 된다. 예를 들면, '어떻게/mad'와 '어트께/mad' 는 문장내에서 동일한 의미로서 언어모델 측면에서 동일한 통계적 특성을 갖는데도, 각각 따로 계산되게 된다. 반면에 <그림 3-b>와 같은 확장된 다중발음사전의 경우 각각의 Leaf노드(303-b)에서 다음 단어로 천이(304-b) 할 때 각 변이발음에 대한 대표어를 가지고 언어모델을 참조하게 된다.

2.3 의사형태소 단위 확장 다중발음사전 구성

<그림 4>는 본 논문에 따른 의사형태소 기반 확장 다중발음 사전을 만드는 블록다이어그램이다.

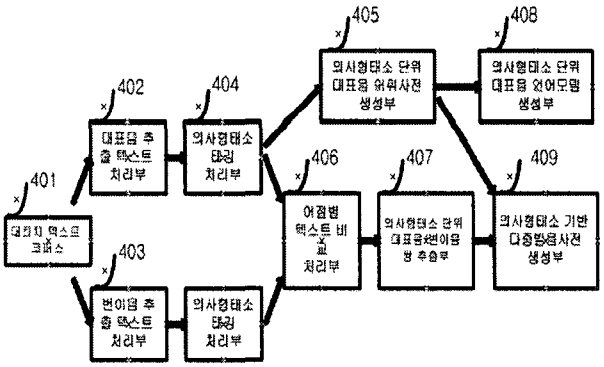


그림 4. 의사형태소 단위 확장다중발음사전 생성 블록다이어그램

(401)은 수집된 대화체 텍스트 코퍼스로서 발음사전과 언어모델 데이터베이스를 만드는데 사용된다. 미리 약속된 전사규칙에 의해 실제 화자가 발음한 발음과 표준발음의 어절 쌍이 전사되어 있다. 즉, '어떻게/어트께' , '했고요/했구여'와 같이 표준발음/화자발음이 전사되어 각각의 어절 쌍을 추출해 낼 수 있게 한다. 이로부터 대표음 텍스트 코퍼스(402)와 변이음 텍스트 코퍼스(403)를 뽑아낸다. 각각에 대해 의사형태소 태깅 및 후처리(404)를 한다. 변이음과 대표음 텍스트 코퍼스를 각각 처리한 결과 의사형태소 태깅된 대표음/변이음 쌍이 어절별로 차이가 나게되고, 어절별 텍스트 비교 처리부(406)에서 이들 쌍들을 뽑아낸다. 의사형태소 단위 대표음/변이음 쌍 추출부(407)는 앞단의 결과를 처리하여 의사형태소 단위 대표음/변이음 쌍을 추출해 낸다. (405)는 대표음만으로 추출되어 의사형태소 단위로 태깅된 코퍼스로부터 의사형태소 단위 어휘사전을 생성한다. (405)의 결과와 (407)의 결과를 이용하여 의사형태소 기반 다중발음사전 생성부(409)는 <그림 2-b>와 같은 의사형태소 기반 확장 다중 발음사전 데이터베이스(105)를 생성해 낸다. 한편 (408)의 의사형태소 단위 대표음 언어모델 생성부는 언어모델 데이터베이스(105)를 생성해 낸다.

Ⅲ. 실험 및 결과

3.1. 언어모델 효율성 실험을 통한 성능평가

본문에서 설명하는 방식과 같이 확장된 다중 발음사전을 이용하여 탐색에 적용할 경우 각 변이발음에 대한 대표음만을 가지고 언어모델을 참조하게 되므로, 언어모델이 좀 더 보강되는 효과를 가져온다. 이를 퍼플렉서티를 기준으로 실험적으로 증명하였다.

실험에 활동된 대화체 음성 및 텍스트 DB는 기존에 본 연구팀에서 보유한 유선 전화음성 DB인 여행계획 3차, 여행계획 4차 DB를 활용하였다. 여행계획 DB는 시나리오를 바탕으로 한 여행사 직원과 고객과의 자연스런 대화를 수집한 것으로 호텔예약, 여행문의 등을 비롯한 복합적인 내용으로 구성되어 있다. DB에 대한 간단한 설명은 다음과 같다.

여행계획 3차: 10시간, 100대화, 긴 대화(6분/대화), 50명 2인 1조, 4대화/조, 25개 시나리오 여행계획 4차: 8.7시간, 125대화, 50명 2인 1조, 5대화/조, 분명한 발음 요구, 15개 시나리오

여행계획 3,4차를 통합하여 총 11,496문장을 얻어내었고, 이중 9,642문장을 Train set으로 1,854문장을 Test set으로 활용하였다. 언어모델 구축 및 테스트를 위해 서는 CMU SLM Toolkit V2를 이용하였다.

표 1. 언어모델 효율성 성능평가 결과

	perplexity	어휘수	OOV
단일 발음사전	147.7	10.295	1,252
다중 발음사전	129.33	9,108	1,087
성능 변화	12% 감소	12% 감소	13% 감소

실험결과 다중 발음사전을 이용했을 경우 퍼플렉서티 측면에서 12% 정도 성능향상을 보여, 언어모델의 효율성이 증가됨을 알 수 있다.

3.2. 인식실험을 통한 성능평가

우선 3.1에서 설명된 여행계획 3차, 여행계획 4차 DB를 이용하여 어절을 인식단위로 하여 인식테스트를 하였다. 텍스트 처리 결과 총 11,496문장의 전체 어절 중 약 14.5%의 어절이 표준발음과 다르게 발성한 어절이었다. 사용된 음향모델과 언어모델은 본 연구팀에서 보유한 전화음성 DB인 여행계획 DB 총 30시간 분량을 이용하여 훈련하여 구축하였고, 인식테스트 결과 어절인식률에서 약 10.9%의 ERR을 보임을 알 수 있었다.

표 2. 어절단위 인식 성능평가 결과

	음절	어절
단일 발음사전	91.5%	86.3%
다중 발음사전	92.3%	87.8%
ERR	9.4%	10.9%

의사형태소를 인식단위로 할 경우 인식성능 평가를 위해 본 연구팀이 보유한 대화체 휴대폰전화 음성 DB를 활용하였다. 실험에 사용된 음성데이터는 총 37시간 2만문장 분량으로 크게 여행자 영역과 ARS 영역으로 나눠 실제 대화상황을 가정한 시나라오를 바탕으로 실제로 2인1조로 대화한 음성을 녹음한 DB이다. 실제 영역 환경에서의 잡음특성과 특히 대화체 음성 특성이 많이 반영된 음성 DB이다. DB에 대한 간단한 설명은 다음과 같다.

여행자 영역: 실제 대화상황 9개 영역, 250명 2인 1조, 4대화/조, 70개 시나리오, 21시간, 실제 잡음환경 ARS영역: 고객상담, ARS 상담 7개 영역, 250명 2인 1조, 50개 시나리오, 16시간

텍스트 처리결과 총 19,645 문장 중 13%의 어절이 표준발음과 다르게 발음된 어절이었다. 의사형태소 단위의 다중 발음사전 구성을 위해 대표음과 발음변이 텍스트 코퍼스의 태깅 후 의사형태소 단위 대표음/발음변이 쌍을 추출하기 위해 추가적인 작업이 필요하였다. <그림 4>의 블록다이어그램을 참조한다.

인식테스트 결과 약 4.3%의 ERR을 보임을 알 수 있다.

표 3. 의사형태소단위 인식 성능평가 결과

	단어(의사형태소)	테스트문장
단일 발음사전	62.5%	435
다중 발음사전	64.1%	435

일반 발음사전과 다중발음사전을 통해 출력된 인식결과를 비교해 보면 그림 5와 같다.

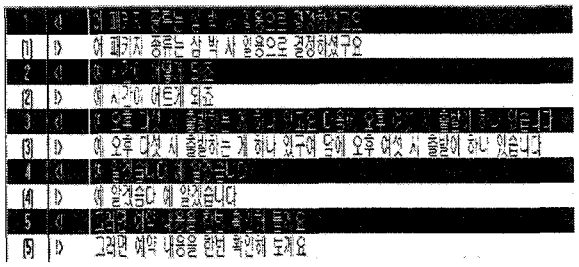


그림 5. 인식결과 출력패턴 비교

Ⅳ. 맺음말

본 논문에서는 대화체 연속음성에서 빈번하게 나타나는 불규칙한 발음변이를 현상을 처리하기 위해 확장된 다중 발음사전의 방법을 이용하였고, 이를 바탕으로 인식성능 실험을 하였다. 특히 인식단위를 의사형태소 단위로 했을 경우의 발음사전에 대화체 음성의 발음변이를 포용하도록 한 의사형태소 기반 다중발음사전을 구성하였고, 그 실험 결과를 보여준다. 확장된 다중발음사전을 통해 언어모델의 효율성을 높여 결과적으로 인식성능이 향상되었고, 정형화된 출력패턴을 얻을 수 있었다.

참고문헌

[1] 이항섭, 박준, 권오욱, "한국어 대화체 음성인식 시스템의 구현", 음성통신 및 신호처리 워크샵, pp.145-148, 1996.8.

[2] O.W. Kwon, K. Hwang, and J. Park, "Korean large vocabulary continuous speech recognition using pseudomorpheme units", EUROSPEECH 99, Budapest, Hungary, Sept. 1999