한국음향학회지 제23권 제4호 pp. 248-252 (2004)

예외 단어 선별 작업을 이용한 자동 발음열 생성 시스템

Automatic Pronunciation Generator Using Selection Procedure for Exceptional Pronunciation Words

김 선 희*, 안 주 은**, 김 순 협**
(Sunhee Kim*, Ju-Eun Ahn**, Soon-Hyob Kim**)

*광운대학교 음성정보처리기술연구센터, **광운대학교 컴퓨터공학과
(접수일자: 2003년 11월 11일; 채택일자: 2004년 3월 17일)

실제 언어생활에 있어서 여러 다양한 경제적 문화적 사회적 환경에 따라 다른 어휘가 사용되고, 각각의 다양한 환경에서 새롭게 신조어가 추가되는 등 어휘의 양적인 변화가 일어난다. 이러한 역동적인 언어 현실을 자동 발음열 생성기에 반영하기 위하여, 본 논문은 추가된 텍스트로부터 예외발음사전을 구축하는 방법을 제안하고, 이러한 방법으로 구축된 예외발음사전을 이용한 자동 발음열 생성 시스템의 성능을 실험하였다. 본 시스템에 대하여 ETRI에서 출시된 음성인식용 텍스트 코퍼스 가운데 한 달 동안의 신문기사를 모은 53,750문장 (740,497 어절)을 이용하여 실험한 결과 100%의 성능을 얻었다.

핵심용어: 자동 발음열 생성, 예외 발음, 선별, 텍스트 코퍼스
투고분야: 음성처리 분야 (2.5, 2.7)

Cultural, social, economic and other various environmental factors affect our language and different words and terminology are used and coined for different contexts, resulting in quantitative change of vocabulary. This paper presents an automatic pronunciation generator using selection procedure for exceptional pronunciation words from added text corpus, which reflects this dynamic nature of language. For our experiment, we used the text corpus released by ETRI for speech recognition, consisting of 53,750 sentences (740,497 Eojols), and obtained a 100% performance level of the proposed automatic pronunciation generator.

Keywords: automatic pronunciation generation, exceptional pronunciation, selection, text corpus.
ASK subject classification: Speech Processing (2.5, 2.7)

I. 서 론

자동 발음열 생성이란 주어진 언어의 맞춤법 체계를 반영하고 있는 문자열을 음성 체계를 반영하는 발음열로 변환하는 것을 의미한다. 일반적으로 자동 발음열 생성 시스템은 규칙적인 형태 음운현상을 처리하기 위한 규칙부와 예외적인 발음을 나타내는 어휘들을 모은 예외발음사전부로 구성된다. 실제 언어생활에 있어서 여러 다양한 경제적, 문화적, 사회적 환경에 따라 다른 어휘가 사용될 뿐만 아니라, 각각의 다양한 환경에서 새롭게 신조어가 추가되는 등 어휘의 양적인 변화가 일어난다. 이러한 현상을 반영하는 새로운 텍스트가 추가되는 경우에, 자동 발음열 생성기에서 규칙적인 형태음운현상은 규칙부에 의해서 완벽하게 처리가 될 수 있으나, 예외발음의 경우는 그 예외발음사전의 보강이 없이는 전체 시스템의 성능 향상을 기대하기 어렵다. 다시 말하면, 자동 발음열 생성 시스템에 있어서 역동적인 언어 현실을 처리하기 위해서는 규칙부가 아닌 예외발음사전부에 대한 보완이 필수적이다. 그러나, 지금까지 자동 발음열 생성 시스템에 관한 연구는 주로 규칙부에 대한 것으로[1, 2, 3], 예외적인 발음을 나타내는 어휘들을 모으는 예외발음사전 구축 방법에 대한 체계적인 연구는 거의 없었다. 본 논문은 [4, 5]를 기반으로 추가된 텍스트로부터 예외발음에 해당하는 어휘를 선별하여 예외발음사전 구축하는 방법을 제시함으로써 자동 발음열 생성 시스템의 성능을 향상하는 것을 그 목표로 한다.

책임저자: 김 선 희 (sunhkim@daisy.kw.ac.kr)
서울 노원구 월계동 447-1 광운대학교 도서관 908-1
(전화: 02-940-5585; 팩스: 02-940-5296)

II. 자동 발음열 생성 시스템

본 논문에서 제안하는 자동 발음열 시스템은 규칙적인 음운현상인 일반음운현상과 형태음운현상을 각각 일반음운규칙과 형태음운규칙으로 규칙화 하고, 불규칙적인 음운현상을 보이는 어휘들을 추출하여 예외사전과 예외규칙을 만들어, 그림 1과 같이 형태음운규칙, 예외규칙 (예외사전 검색), 일반음운규칙의 순서로 적용한다.

전처리 된 입력 문장은 형태소 분석기를 통해 형태소로 분석되는데, 이 때 어간과 어미 정보를 가져는 동사어에만 형태음운규칙인 형태론적 경음화가 적용된다. 형태음운 규칙이 적용된 어절을 제외한 나머지 어휘들 중에, 예외규칙이 적용될 어휘를 선별하는데, 이때, 예외사전을 검색하여 예외사전에 있는 어휘들에게만 예외규칙을 적용한다. 마지막으로, 규칙적 음운 현상이 적용될 수 있는 어휘들에 일반음운규칙을 적용한다.

본 논문에서 제안하는 자동 발음열 생성 시스템은 표준발음을 생성하는 것을 목표로 하며, 본 시스템을 구성하는 각 음운현상과 그에 해당하는 세부 규칙 수는 다음 표 1과 같다.

표 1. 음운현상의 규칙화

음 운 현 상			세부규칙수
규칙적	형태	형태론적경음화(13)	20
	일반	종성 중화(1)	8
		자음군 단순화(2)	6
		유음의 비음화(3)	2
		음운론적 경음화(4)	59
		격음화(5)	23
		장애음의 비음화(6)	27
		유음화(7)	2
		융합(8)	69
		이중 비음화(9)	11
		ㅎ'탈락(10)	2
		ㅎ'비음화(14)	1
		연음(21)	437
		구개음화(22)	3
		이중모음의 단모음화(23)	3
불규칙적		어휘적 경음화(11)	116
		유음화 예외(12)	1
		ㄴ'첨가(24)	13
		중화/단순화+연음(25)	3

III. 예외발음사전 구축 방법

3.1 기본 예외발음사전 구축 방법

어휘의 중복을 최소화하면서 최대한 다양한 음소결합을 포함하는 목록으로부터 예외발음사전을 구축한 경우, 이를 여기에서는 기본 예외발음사전이라고 한다[5]. 여기에서 기본 예외발음사전은 일반사전의 표제어로부터 추출하였는데, 본 연구에서는 일반사전으로 고빈도 5만여 어휘를 표제어로 하는 [6]을 이용하였다.

[6]의 표제어 가운데 예외발음사전을 구성하는 예외 단어들을 선별하기 위하여, 위 표 1에서 제시한 불규칙한 음운현상, 즉, (1) 어휘적 경음화, (2) 비음의 유음화, (3) /ㄴ/ 첨가, (4) 종성중화나 자음군단음화에 따르는 연음, 등 예외발음이 관찰되는 음운환경 (예외발음환경)에 해당하는 어휘들을 분류하였다. 이 어휘들은 불규칙한 음운현상을 보이는 어휘들과 그렇지 않은 어휘들로 분류가 되는데, 불규칙한 음운현상이 관찰되는 어휘들만을 선별하여 기본 예외발음 사전을 구축한다. 여기에서 불규칙한 음운현상을 보이는 어휘들과 그렇지 않은 어휘들을 분류하는 기준은 표준발음[7, 8]이 된다.

이러한 기본 예외발음 사전을 구축하는 방법을 그림으로 나타내면 다음과 같다.

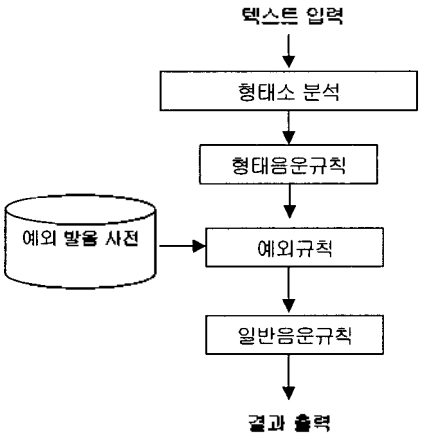

그림 1. 자동 발음열 생성 시스템

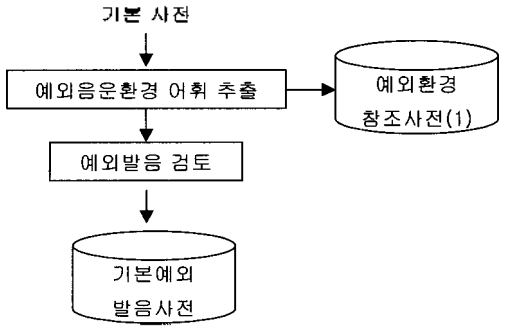

그림 2. 기본 예외발음 사전 구축 방법

위 그림 2에서 '예외환경 참조사전 (1)'이란 예외발음환경에 해당하는 어휘들을 모은 것으로, 예외 발음이 관찰되는 음운환경에 해당하는 모든 어휘가 수록되게 된다.

3.2 추가 예외단어 선별 방법

TTS나 ASR 시스템 구축을 위하여 텍스트가 추가되는 경우에는 이미 구축된 기본 예외발음사전을 기반으로 새로이 추가될 예외 단어들을 추출하여 새로운 예외발음사전을 구축할 필요가 있게 된다. 이와 같이 새로운 텍스트가 추가되는 경우에 그에 따른 예외 단어 추가 과정을 다음과 같은 그림 3으로 나타내었다[5].

먼저, 추가된 텍스트를 어절단위로 처리한 다음, 예외발음환경에 해당하는 어휘를 추출한다. 추출된 어휘를 예외환경어휘라 하고, 이것으로 예외환경어휘 사전 (1)을 구성한다. 이 예외환경어휘사전 (1) 가운데 기본 예외발음사전 구축 시에 만들어 놓은 예외발음 참조사전 (1)에 해당하는 어휘들을 모두 삭제하고, 그 결과로 남게 된 어휘들을 예외환경어휘사전 (2)로 모아 둔다. 즉, 예외환경어휘사전 (2)는 예외발음환경에 해당하는 어휘 가운데 기본 사전에서 예외발음환경에 해당하는 어휘들을 제거한 결과이다. 다음으로, 예외환경어휘사전 (2)에 수록된 어휘들을 수동으로 검토하여 예외발음 추가어휘를 추출해 내게 된다. 이렇게 얻어진 예외발음 추가어휘를 그림 2에서 얻어진 예외발음사전 (1), 즉 기본 예외발음사전에 추가하면 새로운 텍스트에 따른 예외발음사전 (2)가 만들어 지게 된다.

여기에서 예외발음사전을 생성해 내기 위하여, 기본예외발음사전이 아닌 예외환경 참조사전을 이용함으로써 수동으로 검토할 어휘들의 수를 줄일 수 있었다. 즉, 예외환경어휘사전 (2)는 예외환경참조사전 (1)과 결합하여 새로이 구성되고, 이 예외환경참조사전 (2)가 다시 새로운 텍스트로부터 예외발음사전을 생성할 경우에 참조사전으로 이용

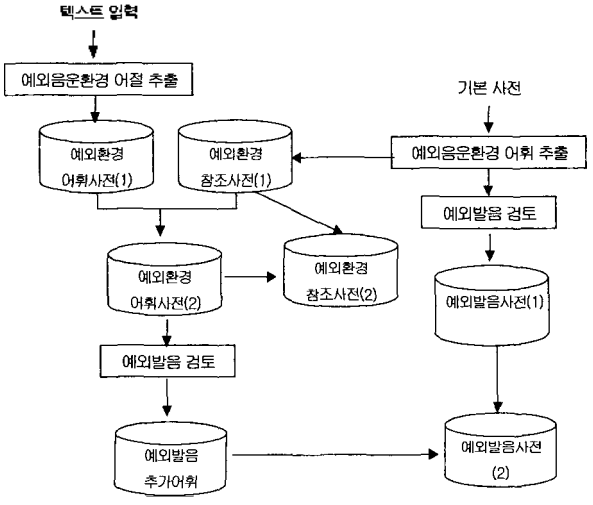

그림 3. 텍스트 적응 예외발음사전을 구축 과정(전체)

되게 된다. 새로운 텍스트가 추가되면 기존의 예외발음사전과 예외환경어휘사전을 참조로 같은 방법을 되풀이 하여 새로운 예외발음 추가어휘를 추출하여, 새로운 예외발음사전을 구축하는 것을 반복하게 된다.

VI. 실험 및 결과 분석

본 연구에서 제시한 자동 발음열 생성기의 성능을 실험하기 위하여, 먼저, 3.1.에서 제시한 방법에 따라 [6]을 이용하여 2,855개의 어휘로 구성된 기본 예외발음사전을 구축하였다. 입력한 49,561 표제어에서 중복된 단어를 제외하면 29,983개의 표제어가 검출되고, 이 가운데 예외발음환경에 해당하는 어휘는 22,932개였다. (이 22,932개의 어휘를 예외환경을 기준으로 분류하면, 한 단어에 예외환경이 2개 이상이 들어있는 경우가 있으므로, 원래 전체 예외환경을 포함한 어휘 수는 34,485개이다.) 이와 같은 방법으로 추출된 22,932개 어휘 가운데 수동으로 검토하여 2,855개의 예외발음을 보이는 어휘를 추출하여 기본 예외발음사전을 구축하였다.

이와 같이 구축된 기본 예외발음사전을 그 구성 요소로 한 자동 발음열 생성기의 성능을 실험하기 위하여 ETRI에서 출시한 음성인식용 텍스트 코퍼스 가운데 한 달 동안의 신문기사를 모은 53,750문장 (740,497어절)을 사용하였는데, 이 가운데 중복된 단어를 포함하는 어절을 제외한 총 189,481어절이 이용되었다.

위에서 구축된 기본 예외발음사전을 이용하여 제안한 자동 발음열 생성기를 실험한 결과 표 2와 같이 99.54% 성능을 얻었다. 이는 형태음운규칙 (형태)이 적용된 경우, 예외규칙 (예외)이 적용된 경우, 일반음운규칙 (일반)이 적용된 경우에 있어서 각각의 어절 수와 아무런 규칙이 적용되지 않은 경우의 어절을 분류하고, 각각의 경우를 전체 실험에 사용된 어절과의 비율로 나타내었다.

다음은 추가된 텍스트인 53,750문장 (740,497어절) 가운데 중복된 단어를 포함하는 어절을 제외한 189,481어절을 이용하여 추가 예외발음사전을 구축한 다음 다시 발음열 생성기의 성능을 실험하였다.

추가된 텍스트는 전체 740,497어절에서 중복된 단어를 제외하면 189,481어절이고, 이 가운데 예외발음환경에 해당하는 어휘만을 모은 예외환경어휘사전 (1)은 159,975어절이었다. 이 예외환경 어휘사전 (1) 가운데 예외사전 참조사전 (1)에 해당하는 어휘를 제외하여 예외환경 어휘사전 (2)로는 91,226어절을 얻을 수 있었다. 그런데, 이 91,226어절을 예외환경을 기준으로 분류하면, 한 단어에 예외환경이 2개 이상이 들어있는 경우가 있었으므로, 전체

표 2. 기본 예외발음사전을 이용한 실험 결과

	형태	예외	일반	불변	총계
어절수	19	11,765	112,471	64,360	189,481
백분율 (%)	0.01	6.21	59.36	33.96	99.54

표 3. 추가 예외단어 선별 결과

	예외환경어휘사전(2)	예외발음 추가어휘
어휘적경음화		68
유음화예외		-
'ㄴ'첨가 중화/ 단순화+ 연음		1
총계		69

표 4. 추가 어휘를 포함하는 예외발음사전을 이용한 실험 결과

	형태	예외	일반	불변	총
빈도	19	12,631	112,471	64,360	189,481
백분율 (%)	0.01	6.67	59.36	33.96	100

예외환경을 포함한 어휘 수는 149,219개가 되었다. 이 149,219어절 가운데 예외발음을 보이는 어휘로는 69개가 최종적으로 선별되었다. 이 69개의 어절은 이전의 기본예외발음 사전과 함께 새로운 예외발음사전을 구성한다.

이와 같은 방법으로 추출된 예외발음 참조사전의 어휘 수와 예외발음의 어휘 수를 위 표 1과 같이 음운현상을 기준으로 나타내면 표 3과 같다.

따라서 740,497어절의 텍스트가 추가된 경우에 새로이 추가되는 예외발음 어휘는 69개로 기본 예외발음사전 2,855개와 더하여 총 2,924개로 이루어진 예외발음사전을 생성하게 된다. 여기에서 추가된 어휘의 대부분이 어휘적 경음화 현상을 보이는 어휘들이라는 것을 알 수 있었다.

이와 같이 새로이 구축된 예외발음사전을 이용하여 제안한 자동 발음열 생성기를 실험한 결과 표 4과 같이 100%의 성능을 얻을 수 있었다.

V. 결 론

무제한 음성합성의 경우나 대용량 음성인식시스템의 경우 모두 최대한 많은 예외발음 어휘를 포함하는 예외 발음 사전을 생성해 내는 것이 필수적이다. 본 연구는 이러한 예외 발음 사전을 구성하는 예외 단어 선별 작업을 이용한 자동 발음열 생성기를 제안한 것이다.

예외발음 사전의 구축은 두 단계로 이루어지는데, 먼저, 기본 어휘가 수록된 일반 사전의 표제어의 목록을 바탕으로 2,855개의 예외발음을 보이는 어휘를 추출하여 기본 예외발음사전을 구축하였다. 다음은 적용 분야에 따라 추가되는 텍스트를 발음열로 전환하기 위하여 추가된 텍스트에서 예외단어를 선별하는 방법에 따라 69개의 예외발음 어휘를 추출하여 새로운 예외발음사전을 구축하였다. 이러한 방법으로 구축된 예외발음사전을 이용한 자동 발음열 생성 시스템의 성능을 평가하기 위하여 ETRI에서 출시된 음성인식용 텍스트 코퍼스 가운데 한 달 동안의 신문기사를 모은 53,750문장 (740,497어절)을 이용하여 실험한 결과 100%의 성능을 얻었다.

본 논문에서 제시하는 바와 같이 예외 단어를 선별하는 작업을 이용한 예외발음사전 구축방법은 예외발음사전을 그 구성요소로 하는 자동 발음열 생성기의 성능을 향상하는데 직접적으로 기여하고, 나아가 TTS나 ASR 시스템의 성능에도 결정적인 도움이 될 것으로 예상된다.

참 고 문 헌

1. 이경님, 전재훈, 정민화, "한국어 연속음성인식을 위한 발음열 자동 생성", 한국음향학회지, 20(2), 2001.
2. 이경님, 정민화, "발음열 자동 생성기를 이용한 한국어 음운 변화 현상의 통계적 분석", 한국음향학회지 21(7): 656-664, 2002.
3. Kim, B., G. G. Lee, & J.-H. Lee, "Morpheme-Based Grapheme to Phoneme Conversion Using Phonetic Patterns and Morphophonemic Connectivity Information, ACM Transactions on Asian Language Information Processing, 1(1), 65-82, 2002.
4. 김선희, "한국어 자동 발음열 생성 시스템을 위한 예외 발음 연구", 말소리 48, 57-67, 2003.
5. 김선희, "한국어 자동 발음열 생성을 위한 예외발음사전 생성", 음성과학 10(4), 167-177, 2003.
6. 연세대학교 언어정보연구원, 연세한국어사전, 두산동아, 1998.
7. 이현복, 한국어 표준발음사전, 서울대학교 출판부, 2002
8. 김석득, 이현복, 유재원, 표준 한국어 발음 대사전, 어문각, 1993.

저자 약력

• 김 선 희 (Sunhee Kim)

1963년 3월 9일생
1985년 2월: 연세대학교 불어불문학과문학사
1986년 10월: 파리7대학(프랑스) 언어학석사(음성학)
1990년 10월: 고등사회과학대학원(프랑스 파리) 언어학박사(음운론)
1991년~2000년: 연세대학교 외 시간강사
2000년~2001년: L&H Korea 책임연구원
2002년~현재: 광운대학교 음성정보처리기술연구센터 연구교수

• 안 주 은 (Ju-Eun Ahn)

1980년 12월 24일생
1999~2002년: 한국방송통신대학교 컴퓨터과학과
2002년~현재: 광운대학교 컴퓨터공학과(석사)

• 김 순 협 (Soon-Hyob Kim)

1947년 12월 28일생
1974년: 울산대학교전자공학과학사
1976년: 연세대학교 석사과정
1983년: 연세대학교 공학박사
1998년~1999년: 한국음향학회장
1979년~현재: 광운대학교 컴퓨터공학과 교수
2000~현재:한국음향학회명예회장