품사 부착 말뭉치를 이용한 임베디드용 연속음성인식의 어휘 적용률 개선*

임민규(서강대), 김광호(서강대), 김지환(서강대)

<차 례>

<Abstract>

Vocabulary Coverage Improvement for Embedded Continuous Speech Recognition Using Part-of-Speech Tagged Corpus

Minkyu Lim, Kwang-Ho Kim, Ji-Hwan Kim

In this paper, we propose a vocabulary coverage improvement method for embedded continuous speech recognition (CSR) using a part-of-speech (POS) tagged corpus. We investigate 152 POS tags defined in Lancaster-Oslo-Bergen (LOB) corpus and word-POS tag pairs. We derive a new vocabulary through word addition. Words paired with some POS tags have to be included in vocabularies with any size, but the vocabulary inclusion of words paired with other POS tags varies based on the target size of vocabulary. The 152 POS tags are categorized according to whether the word addition is dependent of the size of the vocabulary. Using expert knowledge, we classify POS tags first, and then apply different ways of word addition based on the POS tags paired with the words. The performance of the proposed method is measured in terms of coverage and is compared with those of vocabularies with the same size (5,000 words) derived from frequency lists. The coverage of the proposed method is measured as 95.18% for the test short message service (SMS) text corpus, while those of the conventional vocabularies cover only 93.19% and 91.82% of words appeared in the same SMS text corpus.

* Keywords: Vocabulary, Coverage, Embedded speech recognition, Part-of-speech (POS), Tagged corpus.

* 이 논문은 LG전자의 지원을 받아 수행된 연구임.

1. 서 론

휴대용 임베디드 기기 시장이 지속적으로 확대되고, 임베디드 기기들이 더욱 더 소형화되면서, 언제 어디서든 쉽고 편한 정보 접근이 가능해지고 있다. 임베디드 기기의 특성과 기능이 더욱 다양해지고 있지만, 제한된 기기의 크기로 인해서 디스플레이 또는 키패드 상에 다양한 메뉴버튼을 구성하는 것은 한계가 있다. 키보드, 마우스, 터치스크린 등, 인간이 기계와 커뮤니케이션할 수 있는 여러 다른 방법들이 있지만, 음성은 의사소통에서 가장 직관적이며, 핸즈프리(hands-free)와 아이즈프리(eyes-free) 조작을 가능하게 하기에, 효율적인 음성 인터페이스의 구현은 임베디드 환경에서 더욱 중요하다. 특히, 영어와 중국어권에서는 휴대폰 자판을 이용한 문자 입력이 힘들기 때문에, 이들 언어권에서 음성 인터페이스를 이용한 효율적 입력에 대한 관심은 더욱 크다. 현재 임베디드용 음성 인터페이스에 대한 연구는 휴대 전화, PDA, 네비게이터 등에서 음성인식과 음성 합성에 초점을 맞추고 있다.

음성 언어 처리 기술에서 괄목할 만한 진보가 있었지만, 아직 임베디드 환경에서 범용 연속음성인식을 위한 효율적인 솔루션은 제공되지 못하고 있는 실정이다. 메모리 및 계산 용량의 제약으로 인하여, 현재까지 상용화된 임베디드 환경에서의 음성인식기는 음성다이얼링 및 네비게이터에서의 목적지 설정과 같이 주로 가변어 단어 인식 수준에 머물러 있다.

그러나 응용 프로그램의 특성에 따라 어휘의 수와 문장의 형태가 상대적으로 제한된다면, 음향 모델(acoustic model), 어휘(vocabulary), 언어 모델(language model)의 최적화를 통해서 임베디드 환경에서 연속음성인식기의 구현이 가능하다. 예를 들어 short message service (SMS)의 경우 80 바이트 이내의 길이로 이루어진 생활 문장들로 구성이 되며, 이에 대한 연속문장음성인식기의 구현은 범용 연속음성인식기에 비해 상대적으로 적은수의 어휘와 메모리 용량이 작은 언어 모델로 구현 가능하다.

연속 음성인식은 인식가능 어휘를 정의하고, 어휘내의 단어들의 모든 조합에 대해서, 학습과정에서 생성된 음향모델과 언어모델을 이용하여, 입력된 음성에 대해서 확률이 가장 높은 단어열을 탐색(search)한다. 적용률(coverage)을 발성한 단어가 어휘내에 있을 확률로 정의한다면, 어휘내의 단어수가 많을수록 적용률은 높아지지만, 언어모델 및 탐색 과정에 의해 요구되는 메모리 용량이 커지게 되므로, 임베디드 환경에서의 연속음성인식기 구현에 있어서 어휘 최적화는 중요한 문제이다.

말뭉치가 주어진 경우, 일반적으로 해당 말뭉치에 대해서 최적화된 어휘를 정하는 방법은 각 단어별로 말뭉치 내에서 사용된 빈도를 구하고, 빈도순으로 어휘를 구성하는 방법이다. 따라서 응용 프로그램이 사용되는 상황에 대해 충분한 양

의 말뭉치 확보가 용이한 경우(예: 뉴스방송자료), 수집한 말뭉치를 이용해서 쉽게 최적화된 어휘를 정할 수 있다. 그러나 SMS와 같은 임베디드용 응용 프로그램의 경우, 거리적 시간적 제한 없이 다양한 사용자가 사용하고, 또한 그 내용이 개인 사생활에 대한 내용이 많은 경우 해당 응용 프로그램에 대한 충분한 양의 말뭉치를 수집하는 것은 매우 어려운 일이다.

빈도순으로 어휘를 구성하는 방법의 또 다른 단점은 동일 품사에 속하는 단어 중 일부는 어휘에 포함되지만, 일부는 말뭉치에서 측정한 빈도수가 작아서 어휘에서 제외되는 경우가 발생하는 것이다. 이 경우 어휘에서 제외된 단어는 어떤 경우에도 음성인식이 되지 않는다. 예를 들면, 'Monday'는 어휘에 있는데, 빈도 리스트에서 'Tuesday'에 대한 빈도가 낮아 어휘에서 'Tuesday'가 제외 된다면 음성인식기의 사용자는 항상 'Tuesday'의 발성시 음향모델과 언어모델의 정확도와는 관계없이 오인식 결과를 얻게 된다. 실제로 American National 말뭉치를 이용하여 5,000 단어급 어휘를 빈도 리스트로부터 생성하는 경우, 'twelve'와 'twenty'는 미등록어로 처리된다.

본 논문에서는 품사 부착 말뭉치를 이용하여 어휘를 구성하는 방법을 제안한다. 영어권에서 임베디드용 음성 인터페이스에 대한 소비자의 필요성이 상대적으로 높은 점을 감안하여, 제안한 방법을 영어 SMS에 대해서 적용하고 평가한다. 본 논문의 구성은 다음과 같다. 2장에서는 음성인식에서의 어휘 최적화와 관련된 기존의 연구에 대해 논한다. 3장에서는 제안하는 품사 부착 말뭉치를 이용한 어휘 구성 방법을 설명한다. 4장에서는 영어권에서 가장 대표적인 품사 부착 말뭉치인 Lancaster-Oslo-Bergen (LOB) 말뭉치를 이용하여, 제안한 방법으로 5,000단어 어휘를 구성하고, 미국에서 휴대폰을 이용하여 사용자가 직접 작성한 SMS를 대상으로 제안한 방법을 평가한다. 끝으로, 5장에서 결론을 맺는다.

2. 관련 연구

영어, 이태리어, 프랑스어, 독일어에 대한 신문 자료 텍스트 말뭉치를 이용해서 총 단어수, 총 어휘수, 어휘수에 따른 적용률이 측정되었다[1]. <표 1>은 이 측정 결과를 보여주고 있다. 영어의 경우 5,000 단어의 어휘로 얻을 수 있는 최대 적용률은 90% 수준으로 측정되었고, 20,000 단어의 어휘 사용시에는 97% 수준으로 측정되었으며, 미등록어(out-of-vocabulary: OOV) 비율을 1% 미만으로 하기 위해서는 최소 60,000 단어 수준의 어휘가 필요함이 측정되었다. 언어에 따라, 어휘수에 따른 최대 적용률은 차이를 보였다. 4개 언어 중 특히 독일어의 적용률이 낮게 측정되었는데, 독일어는 단어와 단어를 조합한 복합단어(compounding word)의 구성이 상대적으로 자유롭기 때문에, 같은 적용률을 얻기 위해서는 더 많은 단어를 포함

<표 1> 각 언어별 신문 자료 텍스트 말뭉치에 대한 총 단어수, 개별 단어수, 어휘수에 따른 적용률[1]

언어	총 단어수	총 어휘수	어휘수에 따른 최대 적용률		
			5K	20K	65K
영어	37.2M	165K	90.6%	97.5%	99.6%
이태리어	25.7M	200K	88.3%	96.3%	99.0%
프랑스어	37.7M	280K	85.2%	94.7%	98.3%
독일어	36.0M	650K	82.9%	90.0%	95.1%

하는 어휘가 필요하다.

어휘수가 고정되어 있고, 말뭉치를 이용하여 말뭉치 내에서 각 단어별 사용된 빈도에 따라 어휘를 구성하는 경우, 말뭉치의 크기에 따른 적용률의 변화에 대한 실험이 수행되었다[2]. 20,000 단어, 40,000 단어, 60,000 단어의 세 가지의 고정된 크기로 어휘를 설정하고, 신문 자료 텍스트 말뭉치의 단어수를 5M개씩 증가시켜 가면서 말뭉치내의 각 단어별 사용 빈도에 따라 어휘를 생성하여, 신문 자료 테스트 자료에 대해서 적용률을 구한 결과, 말뭉치의 크기가 커짐에 따라 적용률은 높아졌지만, 말뭉치의 크기가 30M 단어를 넘어가면서 부터는 적용률의 향상이 미미해짐을 보였다. 어휘의 크기에 따라 수렴되는 적용률은 20,000 단어, 40,000 단어, 60,000 단어에 대해 각각 약 96%, 98%, 98.5% 수준이다.

도메인에 따라 같은 크기의 어휘에 대한 적용률은 큰 차이를 보인다. <표 1>의 [1]의 실험 결과에 따르면 신문 자료 텍스트 말뭉치에 대한 5,000 단어 어휘의 적용률은 90% 정도로 측정되지만, British National 말뭉치의 구어체(spoken) 부분에 대한 5,000 단어 어휘의 적용률은 96.9%로 나타났고, 5M개의 단어로 영국과 아일랜드에서 녹음한 대화들을 전사(transcribe)한 CANCODE 말뭉치의 경우에도 5,000 단어 어휘의 적용률은 96.1%로 나타났다[3]. 또한 문어체에 내에서도 도메인에 따라 동일 어휘의 적용률이 많은 차이가 발생한다. [4]에 따르면 문어체 말뭉치로부터 많이 사용되는 2,000개의 단어로 구성된 General Service List에 대해서, 학술자료, 신문, 잡지, 소설 각각의 적용률은 78.1%, 80.3%, 82.9% 87.4%로 정리되어 있다.

소용량 어휘의 적용률을 높이기 위한 방법으로 자동 결합(automated compounding)을 통한 실시간 어휘 확장이 제안되었다[5]. 이 방법에서는 우선 기존의 대용량 어휘로부터 품사별로 하나의 엔트리만을 가지도록 기존의 어휘를 수정하고, 복합어를 제거하며, 단어별 사용 빈도를 고려한 크기를 축소한 어휘를 만든다. 그 후, 규칙들과 통계적인 지식을 이용하여 새로운 단어들이 실시간 확장 모듈에 의해 어휘에 추가되고, 최적화된다.

어휘의 크기를 줄이기 위한 방법으로는 perplexity에 기반한 어휘 최적화 방법

이 제안되었다[6]. 본 방법에서는 단어 기반의 기존의 어휘를 이용한 방법의 perplexity와 같은 수준의 perplexity를 유지하면서 어휘의 크기를 줄이거나, 같은 크기의 어휘를 유지하면서 perplexity를 줄이는 것을 목표로 한다. 이의 구현을 위해서 인식 단위를 단어가 아닌 음소, 음절, 형태소 단위로 각각 구현했고, 형태소를 사용한 경우, 학습자료가 테스트 자료를 잘 포함하고, 사용 문법이 정해진 경우에 대해서 어휘의 크기를 기존의 어휘의 절반으로 축소하였다.

중국어에 대한 어휘 최적화 방법이 [7]에서 제시되었다. 이 방법에서 어휘 최적화는 통계적(statistical) 방법에 기반하여 말뭉치로부터 어휘 추가에 적절한 단어들을 선택하고, 그 후 perplexity 최소화 척도를 사용하여 어휘를 간결하게 다듬는(prune) 과정을 계속 반복해 나감으로써 기존의 어휘로부터 어휘가 계속 확장되는 방식이다. 유사한 방법론은 [8]에서도 적용되었다.

고유명사는 음성 자료 검색에 있어서 중요한 정보를 가지고 있기 때문에, 어휘 구성에 있어서 고유명사 처리에 대해 더 많은 주의가 요구된다[9]. 그러나 이러한 고유 명사들은 종종 OOV가 되고, 따라서 음성인식의 주요한 에러 원인이 된다. [9]에 따르면 65,000 단어의 어휘에서 약 28%의 단어들은 고유명사 또는 약어들이었다.

3. 품사 부착 말뭉치를 이용한 어휘 적용률 개선

본 장에서는 본 논문에서 제안하는 품사 부착 말뭉치를 이용하여 어휘를 구성하는 방법을 기술한다. 본 논문에서 목표로 하는 SMS용 연속음성인식기의 어휘는 현재의 휴대폰의 계산용량 및 메모리 공간을 고려하여 5,000 단어를 목표로 한다.

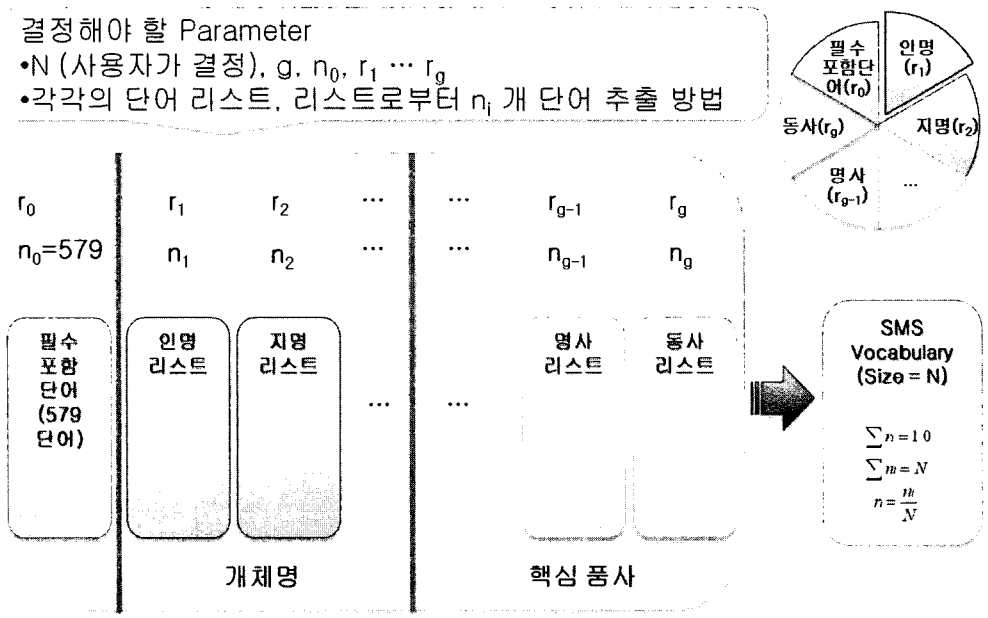
기존의 어휘는 텍스트 말뭉치에 대해서 각 단어별로 말뭉치 내에 사용된 빈도를 구한 후, 빈도순으로 어휘가 구성되어 있다. 이 방법은 정해진 말뭉치에 대해서 제한된 단어수의 어휘에 대해서 적용률을 가장 높게 해주는 어휘를 구성하는 방법이다. 그러나 서론에서 기술한 바와 같이 충분한 양의 SMS 말뭉치를 확보하는 것이 매우 어려워, 정확한 빈도 순서를 정하기가 어려운 단점을 가지고 있다. 또한, 동일 품사에서 속하는 단어 중 일부는 어휘에 포함되지만, 일부는 말뭉치에서 측정한 빈도수가 작아서 어휘에서 제외되는 경우가 발생하는 단점을 가지고 있다.

본 연구는 각 품사별로 품사에 해당하는 단어들의 어휘 반영 방법을 다르게 적용함으로써, 기존 방법의 두 가지 단점을 개선하고자 한다. 예를 들어 3인칭 단수 대명사, 전치사에 해당되는 단어들은 말뭉치 내에서 사용되는 빈도에 관계없이 모든 단어들이 어휘에 포함되어야 한다. 또한 이러한 품사의 경우 단어수가 적기 때문에 해당하는 모든 단어에 대한 어휘 포함 여부가 전문가에 의해 쉽게 결정될

수 있다. 그 반면, 인명, 지명 등과 같은 개체명(named entity)의 경우, 해당되는 단어수에 제한이 없기에 모든 단어에 대한 어휘 포함 여부를 전문가의 판별로 수행하는 것은 불가능하며, 해당 단어 전체 중 일부만이 어휘에 포함된다. 명사, 동사, 형용사, 부사와 같은 핵심 품사에 대해서는 의미상 유사 단어군(예: 과일 사과에 대한 의미상 유사 단어는 배와 같은 다른 과일이 됨)을 형성하여 단어군별 우선순위에 따라 정렬된 리스트를 만들어서 단어군별로 어휘 포함 여부가 결정되어야 한다.

<그림 1>은 본 연구를 이용하여 어휘가 구성되는 과정을 보여주고 있다. 사용자가 어휘의 크기 N을 결정하면, 기 분류되어 있는 n_0개의 어휘 필수 포함 단어가 어휘에 포함된다. 그 후 개체명 및 핵심 품사로 분류되어 있는 g개의 품사들로부터 각각 n_i개의 단어를 선택해서 어휘에 포함하여 어휘를 구성한다.

본 논문에서 기술하는 현재까지의 연구는 필수 포함 단어 선정(n_0) 및 g개의 개체명 및 핵심 품사에 속하는 품사 분류까지 진행이 되어 있다. 대용량 개체명 리스트 확보 및 해당 단어들의 우선 순위 결정, 그리고 핵심 품사에 해당하는 단어들에 대한 유사 단어군 생성에 대한 연구는 추후 이루어질 예정이기 때문에, 개체명 리스트와 핵심 품사별 단어 리스트, 품사별 어휘 구성 비율(r_i)은 말뭉치에서의 빈도 분석을 통해서 결정한 값을 사용한다.



<그림 1> 제안한 어휘 구성 과정

3.1. LOB 말뭉치

본 연구에서는 대표적인 품사부착 말뭉치인 LOB[10] 말뭉치를 사용한다. 이 말뭉치는 영국식 영어로 구성된 텍스트 말뭉치로서 약 1.16M개의 단어를 포함하고 있다. LOB 말뭉치의 총 단어수, 총 어휘수, 품사 태그수는 <표 2>와 같다.

<표 2> LOB 말뭉치의 총 단어수, 총 어휘수, 품사 태그수

총 단어수	총 어휘수	품사 태그수
1,157,278	56,174	152

LOB 말뭉치에서는 152개의 품사 태그를 정의한 후, 말뭉치내의 모든 단어에 대해서 이들 152개 품사 태그중 하나의 태그를 부착했다. LOB 말뭉치는 '단어_품사태그'의 형식으로 기록되어 있다. 예를 들면, LOB 말뭉치내의 문장들은 아래와 같은 형태로 기록되어 있다.

she_PP3A had_HVD been_BEN in_IN mental_JJ hospitals_NNS since_IN 1944_CD ._.

위의 예에서 단어 'she', 'in', 'hospitals'에 대응되는 품사태그는 각각 'PP3A', 'IN', 'NNS'인데, 이는 '3인칭 단수 대명사', '전치사', '복수형 일반명사'를 뜻한다. <표 3>은 위의 문장에서 사용된 태그를 중심으로 LOB 말뭉치의 태그의 의미를 설명한다.

<표 3> LOB 말뭉치의 품사 태그의 예 및 의미와 단어 예시

품사태그	의미	단어 예시
PP3A	3인칭 단수 대명사	she, he
HVD	Have 동사의 과거형	had, 'd
BEN	Be 동사의 과거분사	been
IN	전치사	about, in, to
JJ	형용사	brave, greate, terrible
NNS	복수형 일반명사	abuses, boys, ears
CD	기수	five, hundred, 1900
.	마침표	.

3.2. 품사 부착 말뭉치의 품사별 분류

LOB 말뭉치의 152개의 품사에 대해서 품사에 대응되는 단어들의 어휘 포함

여부의 결정 방법에 따라 품사들을 크게 4개의 그룹들로 분류 했다. <표 4>는 이들 4개 그룹에 대한 설명이다. 그룹 1은 품사에 해당하는 모든 단어들이 어휘에 포함되어야 하는 필수 단어로 구성된 경우이다. 따라서 그룹 1에 해당하는 품사에 대응되는 단어들은 모두 어휘에 포함된다. 그룹 2는 문법상으로 중요한 품사들이지만, 해당 단어들의 어휘 포함 여부는 어휘 크기와 관련이 없는 경우이다. 그룹 1, 2로 분류되는 품사들에 해당되는 단어들은 품사당 평균 29.5단어로, 품사별로 해당되는 단어들이 많지 않기 때문에 각 단어별로 음성인식 전문가가 어휘 포함 여부를 판별하는 것이 가능하다. 그룹 1과 2로 분류한 품사에 해당하는 단어들 중 어휘에 포함되는 단어들은 어휘의 크기에 관계없이 항상 어휘에 포함된다. 그룹 3은 크게 개체명과 핵심 품사인 명사, 형용사, 동사, 부사로 구분되어지는 품사들이다. 해당되는 단어들이 많은 품사들로서, 각 단어에 대한 어휘 포함 여부는 어휘의 크기에 따라 변경 되며, 응용 도메인에 따라서 달라진다. 그룹 4는 어휘에 포함될 가능성이 전혀 없는 단어들로 구성된 품사 태그이다.

<표 4> 어휘 포함 여부의 결정방법에 따른 품사 분류 및 예시 (예시의 형태: 품사태그-해당단어)

품사 분류	설명
그룹 1	품사에 해당하는 모든 단어들이 어휘의 크기에 관계없이 어휘에 포함됨 (예: BE-be, BEM-am, EX-there, HV-have, HVD-had,'d)
그룹 2	문법상 중요한 품사들이고 품사별로 해당되는 단어들이 많지 않아서, 해당되는 모든 단어에 대해 어휘의 크기에 관계없이 어휘에 포함되는 여부를 전문가가 결정함 (예: CD-two,three,hundred,10,45,1000,..., CD-10th,15th,tenth,twenty-first,...)
그룹 3	개체명 및 핵심 품사(명사, 형용사, 동사, 부사)로서 해당되는 단어들이 많은 품사.
그룹 4	품사에 해당하는 단어들이 어휘에 포함될 가능성이 전혀 없는 경우 (예: 특수 기호)

3.3. 품사 분류별 어휘 생성

LOB 말뭉치의 152개 품사들에 대한 품사 분류 결과는 <표 5>에 정리되어 있다. 그룹 1에는 be동사, 조동사, 접속사 등이 포함되었다. 그룹 2에 해당하는 품사로는 기수 (태그명: CD. 예: one, two, hundred), 서수 (태그명: OD. 예: first, second) 등이 있다.

'One'과 '1'이 모두 기수로 분류되어 있는데, 음성인식 결과는 대소문자 구분을 하지 않고, 숫자 대신 그에 해당하는 단어로 표시하는 것을 고려해서 '1'은 어휘에

포함시키지 않고 'one'만을 어휘에 포함 시킨다. 같은 이유로 서수의 경우 어휘에 '1st'와 '1950s' 등은 포함하지 않았다. 같은 방법으로 그룹 1과 그룹 2로 분류되는 품사들에 해당 하는 모든 단어들에 대해서 음성인식 전문가가 어휘 포함 여부를 판별하고, 어휘수에 관계없이 항상 어휘에 포함되는 필수 단어를 분류했다. 분류 결과 579단어가 어휘 크기에 관계없이 모든 어휘에 필수적으로 포함되어야 하는 것으로 조사되었다.

<표 5> LOB 말뭉치에서 정의된 품사들의 분류 결과

품사 분류 그룹	품사 태그
그룹 1 (tag수: 76개)	ABL, ABN, ABX, AP, AP″, AP$, APS, APS$, AT, ATI, BE, BED, BEDZ, BEG, BEM, BEN, BER, BEZ, CC, CC″, CS, CS″, DO, DOD, DOZ, DT, DT$, DTI, DTS, DTX, EX, HV, HVD, HVG, HVN, HVZ, IN″, MD, PN, PN″, PN$, PP$, PP$$, PP1A, PP1AS, PP1O, PP1OS, PP2, PP3, PP3A, PP3AS, PP3O, PP3OS, PPL, PPLS, PPLS″, QL, QLP, RB$, RI, RN, RP, TO, TO″, WDT, WDT″, WDTR, WP, WP$, WP$R, WPA, WPO, WPOR, WPR, WRB, IN
그룹 2 (tag수: 11개)	CD, CD$, CD-CD, CD1, CD1$,CD1S, CDS, OD, OD$, XNOT, ZZ
그룹 3 (tag수: 51개)	RB″, RBR, RBT, UH, JJ, JJ″, JJB, JJB″, JJR, JJR″, JJT, JJT″, JNP, NC, NN, NN″, NN$, NNP, NNP$, NNPS, NNPS$, NNS, NNS″, NNS$, NNU, NNU″, NNUS, NP, NP$, NPL ,NPL$, NPLS, NPLS$, NPS, NPS$, NPT, NPT″, NPT$, NPTS, NPTS$, NR, NR$, NRS, NRS$, RB, VB, VB″, VBD, VBG, VBN, VBZ
그룹 4 (tag수: 14개)	&FO, &FW, !, (,), *', **', *_, „, ., ..., :, ;, ?

그룹 3에 해당되는 품사들은 크게 개체명과 명사군, 동사군, 형용사군, 부사군이다. 그룹 3으로 분류되는 품사에 해당되는 단어의 개수는 상대적으로 많다. 이들 단어들은 SMS에 사용빈도가 높은 단어들도 있고, 사용될 빈도가 거의 없는 단어들도 있기 때문에 어휘의 크기 및 응용 도메인에 따라서 이들 단어들의 어휘 포함 여부가 달라지게 된다. 그룹 3으로부터 어휘에 추가하는 단어들의 개수는 전체 어휘의 크기에서 그룹 1과 그룹 2로부터 선정한 필수 포함단어수(579개)를 뺀 나머지로 결정된다. 그 후, 개체명 리스트와 핵심 품사별 단어 리스트, 품사별 어휘 구성 비율(r_i)은 LOB 말뭉치에서의 빈도 분석을 통해서 결정한 값을 사용한다.

4. 실 험

테스트 자료는 연구 지원기관으로부터 제공 받았다. 이 테스트 자료는 미국현지에서 휴대폰 사용자들이 휴대폰을 이용하여 수신 또는 발신한 SMS를 전사(transcribe)하여 수집했다. <표 6>은 본 논문에서 사용한 테스트 자료(이하 SMSText_US)의 구성을 보여주고 있다. 총 50명의 휴대폰 사용자가 참여했고, 일인당 20~30개의 SMS를 수집했다. 총 문장수, 총 단어수, 총 어휘수는 각각 2,704개, 13,699개, 2,395개로 집계되었다.

<표 6> 테스트 자료의 구성

자료명	총 문장수	총 단어수	총 어휘수	수집 방법
SMSText_US	2,704	13,699	2,395	50명의 휴대폰에 있는 SMS (20~30개/인)를 전사해서 수집

제안한 방법의 검증을 위해 <표 7>과 같이 3개의 어휘를 구성했다. Voc_ECSR은 임베디드 음성인식(embedded continuous speech recognition)을 위해 본 논문에서 제안한 방법으로 구성한 어휘이다. 총 어휘수는 5,000이다. SMS는 텍스트이지만 구어체 스타일의 문장으로 구성되므로 기존의 어휘는 대표적인 구어체 말뭉치인 American National Spoken 말뭉치[11]에서 가장 많이 사용된 5,000단어로 구성한 구어체용 어휘(Voc_ANC_Spoken)와 대표적인 구어체 말뭉치인 American National Written 말뭉치에서 가장 많이 사용된 5,000단어로 구성한 문어체용 어휘(Voc_ANC_Written)의 두가지 어휘를 구성했다.

<표 7> 제안한 방법으로 구성한 어휘 및 기존 방법으로 구성한 어휘에 대한 설명

어휘명	어휘수	어휘 구성 방법
Voc_ECSR	5,000	제안한 방법으로 구성
Voc_ANC_Spoken	5,000	American National Spoken 말뭉치에서 가장 많이 사용된 5,000 단어로 구성
Voc_ANC_Written	5,000	American National Written 말뭉치에서 가장 많이 사용된 5,000 단어로 구성

4.1. 어휘에 따른 적용률 측정 결과

각각의 어휘를 이용하여 테스트 자료인 SMSText_US에 대해서 적용률을 측정

했다. <표 8>은 어휘에 대한 적용률을 보여준다. Voc_ANC_Spoken의 적용률은 93.19%로써 Voc_ANC_Written의 적용률 91.82%보다 높게 나왔다. 이는 SMS 텍스트가 문어체보다 구어체에 가까운 특성을 가지고 있음을 보여준다.

기존 연구를 통해 영어 구어체 문장에 대해 동일 구어체 문장으로부터 선정한 5,000개의 어휘로 최대로 얻을 수 있는 적용률이 약 96%이고, 영어 문어체 문장에 동일 문어체 문장으로부터 같은 크기의 어휘로 최대로 얻을 수 있는 적용률이 약 90%임을 감안하고, Voc_ANC_Spoken 및 Voc_ANC_Written이 SMS 텍스트가 아닌 별도의 말뭉치로부터 구성된 점을 감안하면 이들 두 어휘로부터 얻은 적용률 93.19%와 91.82%는 SMS 텍스트에 대해 5,000단어 규모의 어휘로 얻을 수 있는 높은 수준의 적용률임을 알 수 있다. 본 논문에서 제안한 방법으로 구성한 Voc_ECSR의 적용률은 95.18%로 나타났다. 이는 기존의 어휘를 적용했을 때 얻은 적용률 보다 높은 수치로써, 제안한 방법이 기존의 어휘 구성 방법보다 더 유용한 방법이라는 것을 보여주고 있다.

<표 8> 어휘별 SMS 텍스트 자료(SMSText_US)에 대한 적용률

어휘명	적용률(%)
Voc_ECSR	95.18
Voc_ANC_Spoken	93.19
Voc_ANC_Written	91.82

5. 결론 및 향후 연구

본 논문에서는 임베디드용 연속음성인식을 위한 품사 부착 말뭉치를 이용한 어휘 구성 방법을 제시했다. 대표적인 품사부착 말뭉치인 LOB 말뭉치에서 정의한 152개의 품사에 대해서 품사에 대응되는 단어들의 어휘 포함 여부의 결정방법에 따라 품사들을 크게 4개의 그룹들로 분류했다. 각 그룹별로 음성인식 전문가의 지식을 활용하여 어휘 크기에 관계없이 어휘에 포함되어야 하는 필수 단어를 선정하고, 개체명 및 핵심 품사에 대해 LOB 말뭉치를 이용하여 각 품사별 리스트를 만들고 품사별 어휘 반영 단어수를 결정했다.

본 논문에서 제안한 방법으로 구성한 어휘의 적용률은 95.18%로 나타났다. 이는 기존의 어휘를 적용했을 때 얻은 적용률(93.19%와 91.82%)보다 높은 수치로써, 제안한 방법이 기존의 어휘 구성 방법보다 더 유용한 방법이라는 것을 알 수 있다.

본 논문에서 제안한 방법은 명사, 동사, 고유명사 등에 대한 어휘 포함 단어 선정에서 개선이 필요하다. 예를 들어 'apple'은 어휘에 포함되어 있지만, 유사한

개념의 단어인 'pear'는 어휘에 포함되지 않았다고 가정하자. "Take a bite. This apple is delicious"의 음성입력에 대해서 apple은 인식 가능하지만, 음성입력에서 apple이 pear로 바뀌게 되면, pear는 인식결과의 후보로서 검토가 되지 않고, 결과적으로 해당 단어는 인식이 불가능하게 된다

따라서 추후연구에서는 명사, 동사 등에 대해서 WordNet 등을 활용하여 의미상 유사단어군을 생성하고 이들의 어휘 반영 여부를 잘 선정할 수 있는 방법을 제시하고자 한다. 또한 개체명 인식의 연구 결과를 활용하여 고유명사의 구분을 세분화하고, 체계적으로 고유명사를 수집할 수 있는 방법을 제시하고자 한다.

참 고 문 헌

[1] M. Adda-Decker, L. Lamel, "The use of lexica in automatic speech recognition", *Lexicon Development for Speech and Language Processing,* Kluwer Academic, pp. 235-266, 2000.

[2] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data", *Proc. Eurospeech*, pp. 1763-1766, 1995.

[3] S. Adolphs, N. Shemitt, "Lexical coverage of spoken discourse", *Applied Linguistics*, Vol. 24, No. 4, pp. 425-438, 2003.

[4] P. Nation, R. Waring, "Vocabulary size, text coverage and word lists", *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge University Press, pp. 6-19, 1997.

[5] V. Vandeghinste, "Lexicon optimization: Maximizing lexical coverage in speech recognition through automated compounding", *Proc. International Conference on Language Resources and Evaluation*, pp. 1270-1276, 2002.

[6] K. Hwang, "Vocabulary optimization based on perplexity", *Proc. ICASSP*, pp. 1419-1422, 1997.

[7] J. Zhao, J. Gao, E. Chang, M. Li, "Lexicon optimization for Chinese language modeling", *Proc. International Symposium on Chinese Spoken Language Processing*, 2000.

[8] Y. Xiong, J. Zhu, "Toward a unified approach to lexicon optimization and perplexity minimization for Chinese language modeling", *Proc. International Conference on Machine Learning and Cybernetics*, pp. 3824-3829, 2005.

[9] R. Ordelman, A. van Hessen, F. de Jong, "Lexicon optimization for Dutch speech recognition in spoken document retrieval", *Proc. Eurospeech*, pp. 1085-1088, 2001.

[10] R. Garside, G. Leech, T. Varadi, "Manual of information for the Lancaster parsed corpus". Available at http://khnt.hit.uib.no/icame/manuals/LPC/LPC.PDF.

[11] R. Reppen, N. Ide, "The American National Corpus: Overall goals and the first release", *Journal of English Linguistics*, pp. 105-113, 2004.

접수일자: 2008년 8월 11일
게재결정: 2008년 9월 22일

▶ 임민규 (Minkyu Lim)
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 715-2715
E-mail: lmkhi@sogang.ac.kr

▶ 김광호 (Kwang-Ho Kim)
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 715-2715
E-mail: kimkwangho@sogang.ac.kr

▶ 김지환 (Ji-Hwan Kim) : 교신저자
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 705-8924
E-mail: kimjihwan@sogang.ac.kr