문장 음성 인식을 위한 VCCV 기반의 언어 모델

박 선 희, 홍 광 석
성균관대학교 정보통신공학부
전화 : 031-290-7196 / 핸드폰 : 019-618-0516

A Language Model based on VCCV of Sentence Speech Recognition

Seon-Hee Park, Kwang-Seok Hong
School of Information and Communication Engineering, Sungkyunkwan University
E-mail : seonhp@dreamwiz.com

Abstract

To improve performance of sentence speech recognition systems, we need to consider perplexity of language model and the number of words of dictionary for increasing vocabulary size. In this paper, we propose a language model of VCCV units for sentence speech recognition. For this, we choose VCCV units as a processing units of language model and compare it with clauses and morphemes. Clauses and morphemes have many vocabulary and high perplexity. But VCCV units have small lexicon size and limited vocabulary. An advantage of VCCV units is low perplexity. This paper made language model using bigram about given text. We calculated perplexity of each language processing unit. The perplexity of VCCV units is lower than morpheme and clause.

Ⅰ. 서론

연속 음성인식에선 어휘모델, 음향모델, 그리고 언어 모델이 인식 성능을 좌우한다. 한국어의 특성을 반영하여 문장을 형태소 단위로 발음열을 생성하는 단계가 어휘모델이며 음향 모델은 사람이 발화한 음성으로부터 신호처리와 패턴 매칭을 통해 가능한 발음들을 추정하는 역할을 한다. 그리고 언어 모델의 역할은 음성의 모호함 때문에 음향학적으로 구별하지 못하는 부분을 언어 정보를 이용하여 탐색 공간을 줄이는데 있다.

일반적으로 연속음성 인식성능은 이러한 모델들이 어떻게 결합되어지느냐에 따라 좌우된다[1].

그 중에서 언어 모델로는 문법기반의 언어 모델과 통계적 언어 모델이 있으나, 일반적으로 통계적 언어 모델이 많이 사용되고 있다. 대표적인 통계적 언어 모델로는 N-gram을 들 수 있으며, 충분한 학습 데이터가 존재할 경우에 매우 좋은 성능을 보여준다. 본 논문에서 언어 모델로 bigram모델을 사용하였다[2].

언어 모델의 언어처리 기본 단위로는 한국어의 경우에 어절이나 형태소를 주로 많이 사용한다. 어절은 한글의 띄어쓰기 단위로서 발성의 지속시간이 길기 때문에 N-gram에서 어절 단위를 인식 단위로 사용하는 경우에는 많은 양의 학습 데이터가 존재해야 좋은 성능을 나타낼 수 있다. 형태소를 사용하게 되면 단음소나 단음절을 가진 형태소가 많아서 인식 오류가 증가하게 되므로 인식 성능 개선을 위해 적절한 길이의 발성시간과 적절한 수의 사전 표제어를 가질 수 있게 하는 형태소 결합이 필요하게 된다.

VCCV를 인식 단위로 사용하게 되면 어절 단위보다 모델 구성에 있어서 학습 데이터를 많이 필요로 하지 않기 때문에 사전의 크기가 줄어들게 된다. 그리고 형태소처럼 단음소나 단음절을 가진 형태가 많지 않다.

그리고 형태소나 어절보다 복잡도가 적기 때문에 인식 성능이 좋게 된다.

본 논문에서는 실험 텍스트를 VCCV, 어절, 형태소 단위로 나누어서 어휘 수를 조사하였고 각각의 복잡도를 구하여 비교하였다.

II. 언어 모델

2.1 언어 모델의 개요

언어 모델은 발성된 음성에 대해 기본 단위의 인식 후 발음 사전으로부터 언어적 정보를 추출하여 인식하기 위한 것이다. 보통 발성된 음성에 대한 기본단위 인식은 발화된 기본 단위들 모두를 인식결과로 제공하지 못한다. 즉 기본단위에서의 완전한 인식 결과 없이 단어나 문장을 인식하려면 단어나 문장 구조에 대한 정보가 반드시 요구된다.

따라서 언어 모델은 음성 인식 시스템이 단어나 구문 혹은 문장의 전체 구조에서 올바른 음성 인식을 수행하도록 하기위한 단계이다.

언어 모델은 문장 음성을 인식할 때 탐색해야 할 단어의 수를 줄임으로써 문장 구성을 위한 탐색 시간과 인식률을 높이는 역할을 한다[3].

음성 인식에 주로 사용되는 언어 모델로는 구 구조 문법에 기반한 모델과 통계적 언어 모델을 들 수 있다. 구 구조 문법의 경우 정규 문법이나 문맥 자유 문법을 사용하여 문장의 탐색과 동시에 parsing을 수행하여 문법의 구조에 어긋난 탐색 공간을 제어하게 된다. 비교적 단순한 문법의 경우 FSN을 사용하여 쉽게 표현이 가능한 장점을 가지고 있지만 단어 수가 늘어나면 Network의 state수가 급격히 증가하게 되어 탐색 시간이 오래 걸리는 단점이 있다[2].

이해 비해 통계적 언어 모델은 보통 주어진 영역이 많은 텍스트 문장으로부터 쉽게 추출이 가능하고, 입력 문장 전체를 parsing하지 않고 문장의 발생 확률만을 계산하므로 학습 문장과 부분적으로 다른 문장도 인식 할 수 있는 장점이 있다

2.2 통계적 언어 모델

통계적 언어 모델의 목적은 주어진 인식 영역에 맞는 단어열 W의 확률을 예측하는 것이다. 단어열 W는 $w_1, w_2, \cdots, w_Q$ 로 이루어진 단어열이라고 가정하면 P(W)는 식(1)과 같이 계산된다.

$$P(W) = p(w_1, w_2, \cdots, w_Q) \tag{1}$$

$$= p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\cdots p(w_Q|w_1, \cdots, w_{Q-1})$$

그렇지만 주어진 언어에서 모든 가능한 단어에 대한 P(W)를 계산하기 용이하지 않다. 따라서, Markov가정에 의해 P(W)를 근사하는 N-gram을 사용한다. N-gram 방식은 앞의 N-1개의 단어를 바탕으로 현재의 단어에 대한 확률을 계산하는 방법으로 식 (2) 이다.

$$P(w_j|w_1w_2\cdots w_{j-1}) \approx P(w_j|w_{j-N+1}\cdots w_{j-1}) \tag{2}$$

일반적으로 N은 2(bigram) 또는 3(trigram) 이외에는 계산이 어렵기 때문에 2나 3 이외에는 많이 사용하지 않는다.

언어 모델 P(W)는 보통 정해진 영역의 많은 텍스트 코퍼스로부터 추출한다. 텍스트 코퍼스로부터 P(W)를 추정하는 방법은 다음과 같다. 실제적인 관점에서 P(W)는 식(3)과 같이 근사적으로 표시한다.

$$P_N(W) = \prod_{j=1}^{Q} P(w_j|w_{j-1}, w_{j-2}\cdots w_{j-N+1}) \tag{3}$$

이것을 N-gram Language Model 이라고 한다[2]. 이 조건부 확률 $P(w_j|w_{j-1}, w_{j-2}\cdots w_{j-N+1})$ 은 식 (4)에 의해 추정될 수 있다.

$$P(w_j|w_{j-1}, w_{j-2}\cdots w_{j-N+1}) = \frac{F(w_jw_{j-1}\cdots w_{j-N+1})}{F(w_{j-1}\cdots w_{j-N+1})} \tag{4}$$

여기에서 F는 훈련 코퍼스에서 발생된 문자열의 전체 수이다. 그런데 이것이 잘 추정되기 위해서는 훈련 코퍼스가 매우 커서 단어열이 충분히 나타나야 할 것이다. 실제로 $F(w_jw_{j-1}\cdots w_{j-N+1}) = 0$ 이 되는 단어열이 많이 나타날 것이다. 이러한 문제점을 극복하는 하나의 방법은 N이 3이라고 했을 때 trigram, bigram, unigram의 확률을 이용하는 것이다. 식은 (5)와 같다.

$$P(w_3|w_1, w_2) = P_1 \frac{F(w_1, w_2, w_3)}{F(w_1w_2)} + P_2 \frac{F(w_1, w_2)}{F(w_1)} + P_3 \frac{F(w_1)}{\sum F(w_1)} \tag{5}$$

여기에서 $P_1+P_2+P_3=1$이고 $F(w_j)$는 코퍼스 전체의 크기이다[4].

2.3 Smoothing

언어 모델에서 나올 수 있는 경우가 드는 사전에 대한 확률의 올바른 추정을 위해서는 smoothing 기술이 필수적이다.

현재 음성 인식 이외의 분야에서 Good -turing 방법이 널리 사용되고 있으나 음성 인식에서는 Linear interpolation, Witten-Bell discounting, Absoute discounting, Backing-off 방법 등이 많이 쓰인다.

III. VCCV 언어 모델

본 절에서는 앞 절에서 설명한 통계적 언어 모델을 사용하여 VCCV 단위로 언어 처리를 하는 방법에 대해서 설명한다. 먼저 표준 발음 변환에 대해 설명하고 VCCV 단위에 대한 개념과 모델링 하는 방법 그리고 복잡도에 대해 설명한다.

3.1 표준 발음 변환

음성 인식을 위해서는 음운 변화가 적용된 인식 단위를 처리해야 하므로 음운 변화 현상이 반영된 발음으로 변환을 해야 한다. 발음하는 사람에 따라 발음이 다르긴 하지만 가장 일반적인 표준 발음법을 적용하여 발음 변환을 하였다.

예를 들어 "강에 살고 있습니까"란 문장에 대해서 표준 발음법을 적용하게 되면 있어 ㅆ이 ㄷ으로 바뀌고 습에 ㅂ이 ㅁ으로 바뀌어서 "강에 살고 읻슴니까"로 바뀌게 된다.

본 논문에서 사용한 표준 발음법은 문교부 고시 제 88-2 호에 규정되어 있는 것을 사용하였다.

3.2 VCCV 단위

인식 단위의 경계 추출은 무성음과 유성음 사이에서 뚜렷하게 나타나는 경우가 있지만 유성음, 무성은 혹은 유성음과 무성음들 사이에서 상호간의 조음현상 때문에 정확한 경계를 찾기가 어려울 뿐만 아니라, 이러한 경계 영역의 천이 구간에서 음성 데이터들은 별다른 의미를 갖지 못한다. 왜냐하면 이러한 천이 구간의 데이터는 조음 현상에 의하여 인접 음소의 영향을 받으므로 인접 음소에 따라 특성이 달라지기 때문이다.

따라서 이러한 음소의 경계를 정확히 찾는 것보다 특성의 변화가 적은 안정된 대표구간, 즉 모음 영역을 찾는 것이 더 효율적이라 할 수 있다[5].

이러한 인식 단위를 음향 모델 뿐만 아니라 언어 모델의 언어 처리 단위로도 적용하기 위해서 텍스트를 VCCV(vowel consonant consonant vowel) 로 분할하여 사용한다. 언어학적 특성인 모음 정보를 이용하여 VCCV 단위로 분할하는 예를 보여주기 위해 표 1의 코퍼스를 예로 들었다.

표 1. Corpus의 예

Corpus
1. 강에 살고 읻슴니까
2. 날개가 인는 곤충 임니까

표 1처럼 코퍼스가 주어졌을 때 코퍼스의 음절 수와 모음 정보를 이용하여 코퍼스를 VCCV 단위로 분할하는데 강에는 이음절로 이루어졌고 그 중에서 모음은 아와 에의 두 개로 이루어져 있다. 이것을 VCCV로 나누면 아와 에를 경계로 가, 앙에, 애의 세 부분으로 나누어지게 된다. 표 1에 코퍼스를 VCCV 단위로 나눈 예가 표 2 이다.

표 2. VCCV 단위의 Corpus

VCCV 단위의 Corpus
1. 가/CV 앙에/VCV 애/V | 사/CV 알고/VCCV 오/V | 이/V 읻스/VCCV 음니/VCCV 이까/VCV 아/V
2. 나/CV 알개/VCCV 애가/VCV 아/V | 이/V 인느/VCCV 은/VC | 고/CV 온추/VCCV 웅/VC | 이/V 임니/VCCV 이까/VCV 아/V

본 논문에서는 코퍼스가 분할되어 나타날 수 있는 단위인 V(vowel), VC(vowel consonant), VCV(vowel consonant vowel), VV(vowel vowel), VCCV(vowel consonat consonat vowel), CV(consonant vowel) 등을 한꺼번에 묶어 VCCV 단위라 부른다. 전체 텍스트에서 나올 수 있는 CV의 개수는 총 350개이고 VC의 개수는 166개이다. CV와 VC를 연결하여 구성할 수 있는 전체 VCCV의 개수는 58100개이다.

3.3 VCCV 단위의 언어 모델링

본 논문에서는 코퍼스가 주어지면 그 코퍼스에 대해서 발음 변환을 한 후 VCCV 단위로 분할한 다음 bigram으로 언어 모델을 구성하였다. 구성하는 방법은 첫 번째로 문장이 VCCV로 나누어지면 각 문장의 처음과 끝에 구분 기호를 부여한다. 두 번째로 문장에서 VCCV 회수가 bigram VCCV 회수를 센다. 마지막으로 그 회수를 가지고 bigram 확률값을 계산한다.

3.4 복잡도

음성 언어 처리에서 취급하는 언어 모델은 자연 언

어의 근사 모델이다. 이 경우 언어 모델이 자연언어를 얼만큼 정확하게 근사하고 있는가라는 것이 문제로 된다. 이를 위한 척도로 cross entropy가 일반적으로 이용되고 있다. 복잡도는 언어 모델에 의해 주어진 제약을 나타내는 것으로 일반적으로 객관적인 평가 척도를 말한다.

언어 L에 대한 모델 M을 생각하고 언어 모델 M에 의한 단어열 $w_1, w_2, \cdots, w_n$ 의 생성 확률을 $P_M(w_1 \cdots w_n)$ 으로 나타내기도 한다. 이때 언어 모델 M에 대한 언어 L의 cross entropy와 복잡도 PP를 식 (6),(7) 과 같이 정의 할 수 있다[6].

$$H_0(L,M) = -\sum_{w_1 \cdots w_n} P(w_1 \cdots w_n) \times \log P_M(w_1, \cdots, w_n) \quad (6)$$

$$PP = 2^{H_0} \quad (7)$$

Ⅳ. 실험 및 결과

본 논문에서의 DB는 스무고개 게임에 적용하기 위해서 스무고개에서 나올 수 있는 질문 중에서 동물 영역의 데이터를 가지고 구성하였다. 음성 DB로 구축되어 있는 1000문장에 관한 텍스트와 거기에 추가된 문장이 746개로 총 DB에 구성된 문장은 1746문장이다.

1746 문장에 대해서 표준 발음법을 적용한 프로그램을 사용하여 발음 변환을 한 후 VCCV 단위로 나누었다. VCCV 단위와 비교하기 위하여 발음 변환 후 형태소 단위와 어절 단위로 나누었다. 1746 개의 문장에 대한 어절과 VCCV 그리고 형태소의 개수는 표 3과 같다.

표 3. 어절과 VCCV 그리고 형태소 단위의 개수

문장		1746
어절		1955
VCCV	CV	154
	VCCV	1404
	VC	40
형태소		1543

표 3의 결과에서 어절 단위의 개수에 비교 했을 때 VCCV 단위의 개수가 더 적게 나타나는 것을 볼 수 있다.

동물 1746 문장에 대한 어절과 형태소 그리고 VCCV 단위 bigram 언어 모델을 구성하고 각각의 복잡도를 계산한 결과는 표 4와 같다.

표 4. 어절과 형태소 그리고 VCCV 단위의 복잡도

단위	복잡도
어절	3.98
형태소	2.57
형태소	2.35

표 4의 결과에서 VCCV 단위가 가장 복잡도가 적게 나오는 것을 볼 수 있다.

Ⅴ. 결론

본 논문에서는 VCCV 단위로 언어 모델을 구성하여 기존의 언어 모델 언어처리 단위와 비교하였다. bigram 언어 모델을 적용하여 코퍼스를 형태소와 어절 그리고 VCCV로 나누어서 실험한 결과 VCCV가 이휘수도 적게 나오면서 가장 복잡도가 적게 나온 것을 볼 수 있었다. 복잡도가 너무 크게 되면 인식률을 저하시키게 되는데 VCCV 단위는 복잡도가 적기 때문에 인식에 유용한 단위로 쓰일 수 있다. 향후 과제로는 현재 bigram의 언어 모델만 적용해서 텍스트를 테스트해 보았는데 trigram의 언어 모델도 구성하여 테스트해보고 임의의 텍스트에 대해서 VCCV 단위로 문장 음성 인식을 할 것이다.

< acknowledgement >
본 연구는 한국과학재단 목적기초연구 (R05-2002-000-01007-0) 지원으로 수행되었음.

참고문헌

[1] Laurence Rabiner and Bing-Hwang Juang, "Fundamentals of Speech Recognition", Printice-Hall, EngleWood Cliffs, NJ, 1993.

[2] 오영환 ,"음성언어정보처리", 홍릉과학출판사, 1997.

[3] 이건상, 양성일, 권성헌 공저, "음성인식", 한양대학교 출판부, 2001.

[4] 이진석, 박재득, 이근대, "K-SLM Toolkit을 이용한 한국어의 통계적 언어 모델링 비교", 한국전자통신연구원, 1999.

[5] 윤재선, "한국어 음성 인식 dictation system의 구현", 성균관대학교 전기 전자 및 컴퓨터 공학과 박사 학위 논문, 2001.

[6] 北 研二, "音聲言語處理", 森北出版株式會社, 1998.