기능어용 음소 모델을 적용한 한국어 연속음성 인식

명주현, 정민화
서강대학교 컴퓨터학과

Korean Continuous Speech Recognition using Phone Models for Function words

JooHyun Myung, Minhwa Chung
Department of Computer Science, Sogang Univercity

요 약

의사형태소를 디코딩 단위로 하는 한국어 연속 음성 인식에서는 조사, 어미, 접사 및 짧은 용언의 어간등의 단어가 상당수의 인식 오류를 발생시킨다. 이러한 단어들은 발화 지속시간이 매우 짧고 생략이 빈번하며 결합되는 다른 형태소의 형태에 따라서 매우 심한 발음상의 변이를 보인다.

본 논문에서는 이러한 단어들을 한국어 기능어라 정의하고 실제 의사형태소 단위의 인식 실험을 통하여 기능어 집합1, 2를 규정하였다. 그리고 한국어 기능어에 기능어용 음소를 독립적으로 적용하는 방법을 제안했다. 또한 기능어용 음소가 분리되어 생기는 음향학적 변이들을 처리하기 위해 Gaussian Mixture 수를 증가시켜 보다 견고한 학습을 수행했고, 기능어들의 음향 모델 스코어가 높아짐에 따른 인식에서의 삽입 오류 증가를 낮추기 위해 언어 모델에 fixed penalty를 부여하였다. 기능어 집합1에 대한 음소 모델을 적용한 경우 전체 문장 인식률은 0.8% 향상되었고 기능어 집합2에 대한 기능어 음소 모델을 적용하였을 때 전체 문장 인식률은 1.4% 증가하였다. 위의 실험 결과를 통하여 한국어 기능어에 대해 새로운 음소를 적용하여 독립적으로 학습하여 인식을 수행하는 것이 효과적임을 확인하였다.

1. 서 론

한국어 음성 언어 처리의 경우는 사람의 자연스러운 음성을 그 입력으로 하기 때문에 인식 단위의 경계가 모호하고, 인식 결과 자체에 어느 정도의 오류가 포함되어 있어서 문자 기반의 언어 처리와는 달리 정확한 입력 형태를 기대하기 어렵다. 그러므로 보다 정교한 음성 언어 처리를 위해서 형태소 해석 단계에서부터 음성 인식과 자연어 처리 기술을 접목하는 접근이 필요하다. 문자 기반의 형태소에는 단음소, 단음절로 구성되어 있는 단위들이 많고 발화 지속 시간이 짧은 단위일수록 그 안에서 인식에 필요한 충분한 정보를 추출해 내기가 어려워 진다. 그래서 이를 위해 제안된 방법이 새로운 디코딩 단위인 의사형태소(Pseudo-Morpheme)[3]를 이용하는 것이다. 그런데 적절한 결합정보를 이용하여 의사형태소 단위로 결합을 하더라도 조사 및 짧은 어미 중 몇몇은 결합되지 않은 채 남아 있게 된다. 실제 실험에서 발생되는 인식 오류를 살펴 보면 상당수의 오류가 이러한 단어에 의한 것이다. 이들은 대체적으로 강세를 약하게 받고 생략이 빈번하며 주위 문맥(context)에 의해 심하게 왜곡되는 현상을 보인다.

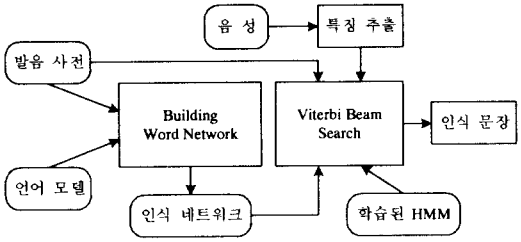
본 논문에서는 인식 오류의 상당 부분을 차지하는 이러한 단어들을 한국어에서의 기능어(function word)로 정의하고 기능어의 인식률 향상을 위해 새로운 기능어용 음소[1] 모델을 새로이 적용한 접근을 제안한다. 같은 phone 이라도 기능어와 기능어가 아닌 일반 표제어(content word)에서 발화될 때 음성학적인 특성이 다르게 나타난다. 따라서 기능어에 사용되는 음소들은 일반 표제어와 분리시켜 학습을 하여 기능어만의 음향학적 특성을 반영할 수 있는 파라미터를 얻어 내고 이를 적용하여 인식 실험을 수행한다.

2. 베이스라인 연속 음성 인식 시스템

2.1. 베이스라인 연속 음성 인식기의 구성

본 논문의 실험을 위한 베이스라인 연속 음성 인식기로 Entropic사의 HTK(HMM Tool Kit) 을 사용하였다[1]. 베이스라인 시스템의 구조는 [그림 1]과 같다.

[1] 원래 '음소'는 'phoneme'에 해당하는 말로 음운의 최소 단위를 의미한다. 이런 음소가 실제로 발음된 것을 '음성(音聲)'이라고 하고 이 말은 'phone'에 대응한다. 그런데 '음성'이라는 표기는 'speech'와 혼돈될 가능성이 있으므로 이후부터는 'phone'을 '음소'라고 표기한다.



[그림 1] 베이스라인 연속 음성 인식기 구성도

문장 단위로 입력되는 음성은 모두 16kHz 로 샘플링하였으며, 매 10ms 마다 25ms 의 크기를 갖는 해밍 윈도우를 프레임을 만들고 이어서 preemphasis 과정을 거친 후 FFT 를 수행하여 13차 MFCC 를 구하고 얻어낸 MFCC 의 1 차 및 2 차 차분 계수를 구하여 총 39 차의 특징 벡터를 생성한다. 기본적인 인식 단위로 좌우 문맥을 고려한 음소단위인 트라이폰을 사용한다. 다중발음열을 허용하였고 생성된 다중 발음사전을 이용하여 사전 표제어 내부에서 생성되는 트라이폰들만 사용을 하였다. 각 HMM 은 5 개의 상태를 가지는 left-to-right HMM 으로 구성하였다. 기본실험에서는 각 상태마다 Gaussian Mixture 를 1 개만 사용하였고 실험을 진행하며 보다 견고한 학습이 필요하여 수정된 실험환경에서는 각 상태당 2 개의 Gaussian Mixture 를 이용하였다. 본 실험에서 사용된 기본적인 인식단위의 음소 표기는 서강대 PLU[2]를 이용하였다. 또한 인식 성능을 높이기 위해 적절한 길이의 발성시간을 가지고 언어학적으로도 안정적인 의사형태소[4]를 디코딩 단위로 사용한다.

3.2. 언어모델

언어모델은 학습 코퍼스내에서 특정 단어열이 나타날 확률을 말한다. 이 언어모델 확률을 이용하여 현재까지 나온 단어열을 통해 현재 인식해내야 하는 단어에 대한 확률을 계산할 수 있다. 언어모델 확률값은 다음 수식에 의해 계산되어 진다.

$$P_{HW} = (\Pr(W \mid H))^{\omega} + \rho$$

여기에서 $\Pr(W \mid H)$ 는 현재까지의 단어 history H 에 대해서 현재 단어 W 가 나올 조건부 확률을 의미한다. 그리고 ω 는 언어모델 scale fac tor로, 언어모델 확률값에 상대적인 가중치를 부여하기 위해 사용되는 값이고, ρ 는 언어모델 penalty로써 인식시의 삽입 및 삭제 오류를 제어하기 위해 사용된다.[1] 이 언어모델 penalty 는 단어 네트워크상의 탐색을 수행할 때 한 노드에서 다른 노드로 토큰이 방출되는 순간에 해당하는 값만큼의 벌점을 주기 위해 사용되었다. 즉 탐색시 거쳐가는 경로상의 노드수가 많을수록 더 많은 벌점을 받게 된다.

[2] PLU(Phone-Like Unit)는' 유사음소 라고 하며 인식 단위의 음소표기이다.

3. 기능어용 음소 모델

3.1. 한국어의 기능어

영어권에서의 기능어는 전치사, 접속사, 대명사, 그리고 짧은 동사들로 구성되어 있다. 이런 단어들은 여러 가지 고유한 특성들로 인해 연속 음성 인식 과정에서 특유의 문제점을 보인다. 93% 의 내용어가 발화 중 강세를 받는 반면에 기능어는 단지 14% 만이 강세를 받는다[2]. 이런 기능어들의 음소는 정확한 지속시간 학습이 되지 못하고 주위 문맥에 의해서 심각한 영향을 받는 등 여러 형태로 변형되기 쉽다. 한국어의 조사, 어미 접사등의 단어들은 발화 지속시간이 매우 짧고 결합되는 다른 형태소의 형태에 따라서 매우 심한 발음상의 변이를 보인다. 본 논문에서는 한국어 연속 음성 인식에서 빈번한 인식 오류를 발생시키는 단어들을 기능어라 정의하는데 여기에는 언어학적인 측면에서의 대상인 형식 형태소들과 실험을 통해서 밝혀진 대상인 길이가 짧은 용언의 어간 등이 포함된다.

3.2. 기능어 집합

의사형태소를 디코딩 단위로 한 인식 실험을 수행하여 실제 발생하는 인식 오류의 패턴과 종류를 분석했다. 각 단어의 인식 오류의 횟수를 구하고 조사, 어미, 접사 및 짧은 용언의 어간등에 대해 15 회 이상의 오류를 보인 14 개의 단어들을 기능어 집합1로, 큰 오류 횟수를 보이지는 않지만 대용량으로 확장되고 자연스러운 문장으로 발화된다면 충분히 오류를 발생시킬 수 있는 단어들까지 확장한 28 개의 단어로 구성된 집합을 기능어 집합2로 정의하였다. 다음 [표1]은 기능어 집합2이고 표의 왼쪽 부분의 14 개의 단어가 기능어 집합1이다.

단 어	오류 횟수	단 어	오류 횟수
의	90	와	13
에	71	면	10
이	146	기	14
하	69	죠	12
을	51	여	11
은	41	없	12
해	42	요	9
가	30	들	8
는	52	주	14
ㄴ	29	으로	7
있	32	서	8
ㄹ	21	보	9
를	15	어	7
되	47	로	7

[표1] 기능어 집합

본 논문에서는 기능어로 새로이 추가되는 단어에 대해서 그 단어의 발음열에 해당하는 새로운 음소를 부여한다. 서강대 PLU set 은 대문자 표기를 원칙으로 한다. 그러므로 여기에서 새롭게 추가되는 음소는 같은 발음을 갖는 서강대PLU 의 소문자 표기를 사용한다. 기능어 집합1에 의해 추가되는 음소는 19 개 이고 기능어 집합 2 에 의해 추가되는 음소는 17개로 다음 [표2]에 나타나 있다.

기능어 집합1에 의해 추가되는 새로운 음소	기능어 집합2에 의해 추가되는 새로운 음소
aa d eh ey g hh hi iy kh l n ph r ss tq tt we wi ww	ax b ch jo jx kk m ow p pq s t uw wa z zh zz

[표2] 기능어 음소

3.3. 기능어 음소를 이용한 발음 사전 구성

정의된 기능어 집합을 통해서 기능어에만 사용되는 새로운 음소를 전체 음소집합에 추가하고 이를 발음 사전에 적용하여야 한다. 학습용 발화의 전사된 문장을 형태소 분석을 하고 의사형태소 결합규칙에 의해 의사형태소 태그를 붙인 후에 [6]의 시스템을 수정하여 해당 문장에 맞는 한국어 발음열을 생성하였다. 그리고 각 단어의 태그정보를 이용하여 기능어 여부를 판단한 뒤 기능어는 해당하는 기능어 음소로 변환하여 새로운 발음열을 생성하여 학습에 이용한다.

4. 실험 및 결과 분석

4.1. 실험 환경

본 논문의 실험을 위해서 사용한 음성 데이터베이스는 한국 과학 기술원 통신 연구실에서 제작한 3,000 단어급의 무역 상담용 연속 음성 데이터베이스이다. 이 중 HMM 모델 학습을 위해 남성 화자 50 명이 발화한 4910 개의 문장을 사용하였고, 인식 실험용으로는 남성 화자 11 명이 발화한 1000 개의 문장을 사용하였다.

4.2. 기본 실험

기본 실험에서는 기능어 음소1을 적용한 실험에서 기능어 음소를 추가하지 않은 경우보다 문장 correctness 와 단어 accuracy는 떨어지는 결과를 보였다. 전체 맞은 단어의 수도 줄었고 대체 오류와 삭제 오류는 감소하였으나 삽입오류가 급격히 증가하여 단어 correctness 만이 증가하였다. 기능어용 음소가 분리되어 독립적으로 학습이 이루어졌기 때문에 발생한 음향학적인 문제를 해결하기 위해 Gaussian mixture 수를 증가시켜 견고한 학습을 수행했고, 증가한 삽입오류를 제어하기 위하여 언어모델에 penalty 를 부여하였다.

4.3. 수정된 환경에서의 실험

Gaussian mixture 를 증가시키고 언어모델 penalty 를 부여한 수정된 환경에서 기능어 음소를 추가하여 실험을 수행하였다. 수정된 환경하에서 기능어 음소1을 적용하여 실험을 하였을 때 적용하지 않은 경우보다 문장 및 단어 인식률이 향상되었고 기능어 음소2을 적용하였을 때는 기능어 음소1을 적용하였을 때 보다도 더 향상된 결과를 얻었다. 인식결과는 다음 [표3]에 나타나 있다.

	기능어 음소 적용 없음	기능어 음소1 적용	기능어 음소2 적용
Sent Corr.	35.6%	36.4%	37%
Word Corr.	85.2%	86.31%	86.5%
Word Acc.	82.03%	82.84%	83.20%

[표3] 인식결과

5. 결론

본 논문에서는 한국어 연속 음성 인식에서 빈번한 인식 오류를 발생시키는 조사, 어미 및 짧은 용언의 어간등을 한국어 기능어라 지칭하고 이러한 기능어용 음소를 독립적으로 적용하는 방법을 제안하였다. 실제 인식 실험을 통해 기능어 집합을 정의하고 각각의 단어에 대한 기능어 음소를 적용한 실험을 수행하여 결과를 비교 분석하였다. 기본실험을 통하여 발생한 문제를 해결하기 위해 Gaussian Mixture 를 증가시키고 언어모델 penalty 를 부여한 새로운 실험환경을 적용하였다. 새로운 실험 환경하에서 기능어 집합1에 대해 기능어 음소를 적용하면 전체 문장 인식률은 0.8%, word correctness 는 1.29%, word accuracy는 0.79% 향상되었고, 기능어 집합1의 전체 오류 횟수는 708 에서 649 로 감소하였다. 또한 기능어 집합2에 대한 기능어 음소를 적용하였을 때 전체 문장 인식률은 1.4%, word correctness 는 1.48%, wor d accuracy 는 1.17% 증가하였고, 기능어 집합2의 전체 오류 횟수는 826 에서 772 로 감소하였다. 위의 실험 결과를 통해 빈번한 인식오류를 발생시키는 한국어 기능어에 대해 새로운 음소를 적용하여 독립적으로 학습하여 인식을 수행하는 것이 효과적이었음을 확인하였다. 또한 기능어 집합1과 기능어 집합2로 구분하여 실험해 본 결과 모두 인식률이 향상되므로 기능어 집합의 정의가 효과적이었음을 확인하였다.

6. 향후과제

실제 연속 음성 인식과정에서 언어모델과 음향 모델을 분리시켜 고려할 수는 없다. 현재 사용되는 단어 bigram 의 경우 코퍼스 내에서 디코딩 단위가 되는 단어들의 결합 패턴에 의해 그 확률값이 결정되므로 기능어에 대해 보다 정확한 처리를 하기 위해서는 태그 정보를 이용한 언어모델이 적용되어야 한다. class -based bigram 등의 언어모델을 통해 태그 정보를 이용하면 현 시스템의 성능을 개선시킬 수 있을 것으로 보인다. 또한 언어모델에 penalty 를 부여하여 증가된 삭제 오류를 줄이기 위한 접근이 필요하고 더 다양한 실험을 통하여 보다 정확한 기능어 집합의 정의가 요구된다.

7. 참고문헌

[1] Steve Young, *The HTK Book*, 1998.

[2] Waibel, A.H. *Prosody and Speech Recognition*, PhD thesis, Computer Science Department, Carnegie Mellon University, 1986.

[3] 권오욱, 박준, 황규웅, "의사형태소 단위 대어휘 연속음성 인식기 개발", *제 15 회 음성통신 및 신호처리 워크샵 논문집*, pp.320-323,1998.

[4] 이경님, "의사형태도 단위의 한국어 연속 음성 인식", 서강대학교 전자계산학과 석사학위논문, 1997.

[5] 이상호, 서정연, 오영환, "KTS: 미등록어를 고려한 한국어 품사 태깅 시스템", *제 12 회 음성통신 및 신호처리 워크샵 논문집*, pp.195~199, 1995.

[6] 전재훈, "형태음운학적 분석에 기반한 한국어 발음열 자동 생성", 서강대학교 전자계산학과 석사학위논문, 1997.