Journal of Korea Multimedia Society Vol. 13, No. 11, November 2010(pp. 1586-1593)

의미 분석과 형태소 분석을 이용한 핵심어 인식 시스템

안찬식†, 오상엽††

요 약

확률적 패턴 매칭과 동적 패턴 매칭의 어휘 인식 오류 보정 방법에서는 핵심어를 기반으로 문장을 의미론적으로 분석하므로 형태론적 변형에 따른 핵심어 분석이 어려운 문제점을 가지고 있다. 이를 해결하기 위해 본 연구에서는 음절 복원 알고리즘에서 형태소 분석을 이용하여 인식된 음소 열을 의미 분석 과정을 통해 음소의 의미를 파악하고 형태론적 분석으로 문장을 복원하여 어휘 오인식률을 감소하였다. 시스템 분석을 위해 음소 유사률과 신뢰도를 이용하여 오류 보정률을 구하였으며, 어휘 인식 과정에서 오류로 판명된 어휘에 대하여 오류 보정을 수행하였다. 에러 패턴 학습을 이용한 방법과 오류 패턴 매칭 기반 방법, 어휘 의미 패턴 기반 방법의 성능 평가 결과 2.0%의 인식 향상률을 보였다.

Key-word Recognition System using Signification Analysis and Morphological Analysis

Chan-Shik Ahn†, Sang-Yeob Oh††

ABSTRACT

Vocabulary recognition error correction method has probabilistic pattern matting and dynamic pattern matting. In it's a sentences to based on key-word by semantic analysis. Therefore it has problem with key-word not semantic analysis for morphological changes shape. Recognition rate improve of vocabulary unrecognized reduced this paper is propose. In syllable restoration algorithm find out semantic of a phoneme recognized by a phoneme semantic analysis process. Using to sentences restoration that morphological analysis and morphological analysis. Find out error correction rate using phoneme likelihood and confidence for system parse. When vocabulary recognition perform error correction for error proved vocabulary. system performance comparison as a result of recognition improve represent 2.0% by method using error pattern learning and error pattern matting, vocabulary mean pattern base on method.

Key words: Morphological Analysis(형태소 분석), Signification Analysis(의미 분석), Key-word Recognition(핵심어 인식), phoneme likelihood(음소 유사률)

1. 서 론

최근 인터넷의 발전으로 정보의 양이 방대하게 증가하여 컴퓨터 및 모바일 단말기를 이용한 정보의 효율적인 검색을 위해 어휘 인식에 대한 관심이 모아지고 있다. 어휘 인식에는 여전히 유사한 음소와 부정확한 어휘 제공에서 오류가 존재한다. 어휘의 입력으로부터 신호를 처리하여 오류를 보정하려는 관련

※ 교신저자(Corresponding Author) : 오상엽, 주소 : 경기도 성남시 수정구 복정동 산65(461-702), 전화 : 031)750-5798, FAX : 02)426-9159, E-mail : syoh@kyungwon.ac.kr
접수일 : 2010년 6월 15일, 수정일 : 2010년 9월 28일
완료일 : 2010년 10월 19일

† 정회원, 광운대학교 컴퓨터공학과 박사과정
(E-mail : absoluti@kw.ac.kr)
†† 종신회원, 경원대학교 IT대학 컴퓨터소프트웨어 교수
※ 본 연구는 2010년도 경원대학교 지원에 의한 결과임.

연구가 진행되고 있다. 화자 독립적인 시스템에서는 어휘 인식의 효율을 높이면서 오류를 보정하기 위해서는 신호 처리만으로 해결하기에 어려운 과정들이 따르게 된다. 따라서 어휘의 일반적인 신호 처리의 인식 결과에서 어휘 후처리를 이용한 오류 보정에 대한 연구를 진행하고 있으며[1], 기존의 방법에서는 확률적 패턴 매칭 기반의 오류 보정 방법과 동적 패턴 매칭 기반의 오류 보정 방법이 있다.

확률적 패턴 매칭 기반의 오류 보정 방법은 인식 과정에서 나타나는 오류를 정리하여 오류 패턴을 만들어 미리 학습을 수행하여 일정한 오류 패턴을 가지고 계산된 확률 값에 의해 처리하는 방법으로 확률 계산을 위해 많은 오류 패턴이 필요하고 시간이 오래 걸리는 단점이 있다[2].

동적 패턴 매칭 기반의 오류 보정 방법은 인식 대상열의 길이가 다를 경우 매핑 함수를 적용하여 정보 손실로 인한 패턴 매칭이 제대로 이루어지지 않아 인식할 문장을 의미적으로 분석하고 분석한 의미 정보를 특정 스트링에 포함하여 사용하므로 형태적인 변형이 일어나는 단점이 있다[3].

확률적 패턴 매칭과 동적 패턴 매칭은 오류 패턴 DB가 많이 필요하고 핵심어를 기반으로 문장을 의미론적인 분석을 수행하여 형태론적 변형에 따른 핵심어 분석이 어려운 문제점이 존재한다. 따라서 본 논문에서는 음절 복원 알고리즘에서 의미 분석과 형태소 분석을 이용한 핵심어 인식 시스템을 제안한다.

인식된 음소 열을 형태소 분석 과정을 거쳐 음소가 갖는 형태를 파악하고 음절 복원 알고리즘을 통해 형태적 변형이 일어나기 이전의 문자열로 복원한다. 형태소 분석 과정은 입력된 문장의 분석 후보를 다수 생성한 후 최적의 후보를 선택하는 방법이며 어절의 의미를 형태적으로 파악하여 문법적인 어절은 인식하고 비문법적인 어절은 오류 보정을 수행한다. 의미적으로 분석하기 어려운 핵심어 문장을 복원하여 전체적인 어휘의 오인식을 감소시켰다. 형태소 분석을 이용한 음절 복원 알고리즘에서 핵심어에 대한 오인식을 감소시켰으며 오류 패턴 DB의 양을 줄였다.

시스템 성능 평가를 위해 음소 유사률과 신뢰도를 이용하여 오류 보정률을 나타냈으며, 어휘 인식 과정에서 오류로 판명된 어휘에 대하여 오류 보정을 수행하였다. 에러 패턴 학습을 이용한 방법과 오류 패턴 매칭 기반 방법, 어휘 의미 패턴 기반 방법의 성능 평가 결과 2.0%의 인식 향상률을 보였다.

2. 관련연구

2.1 확률적 패턴 매칭 기반 오류 보정

확률적 패턴 매칭 기반의 오류 보정 방법은 인식 과정에서 나타나는 오류를 정리하여 오류 패턴을 만들고 사전 학습을 통하여 일정한 오류 패턴에 의한 계산된 확률 값에 의해 처리하는 방법으로 확률 계산을 위해 많은 오류 패턴이 필요하며 오류 패턴을 만들고 사전 학습을 진행하기 위한 시간이 오래 걸린다는 단점이 있다.

하지만 정보 검색 영역에서 사용되는 문장은 문장이 간결하고 사용자가 검색하고자 하는 핵심어로만 이루어진 경우가 많아 문장 전체의 오류 패턴을 만들고 사전 학습을 수행하는 것보다 적은 오류 패턴과 시간이 소요된다.

어휘에 대한 확률적 예측을 위해 은닉 마코프 모델을 이용한다. 은닉 마코프 모델(HMM:Hidden Morkov Model)은 숨겨져 있는 모델을 기존 모델의 확률을 이용하여 추정하는 방법이다[4].

HMM은 초기 $(t=0)$에 상태 i의 확률 $\pi_i = \Pr(s_0 = i)$, 상태 i에서 j로의 천이 확률 $a_{ij} = \Pr(s_t = i, s_{t+1} = j)$, 상태 j에서 심볼 k를 관측할 확률 $b_j(k) = \Pr(x_t = k | s_t = j)$로 표현되어 다음 식(1)과 같이 나타낸다.

$$P(q|\lambda) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdots \cdot a_{q_{N-1} q_N} \quad (1)$$

임의의 음성 특징벡터의 관측열 $O = (o_1, o_2, \ldots, o_T)$이 사실임을 가정할 때 주어진 N-states HMM 모델에서의 상태열이 $q = (q_1, q_2, \ldots, q_T)$ 라면 결국 관측열의 확률은 다음 식(2)와 같이 주어진다.

$$P(O|q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T) \quad (2)$$

관측열과 상태열의 결합 확률은 O와 q가 동시에 일어날 확률로 다음 식(3)과 같이 표현되고, 식 (1)과 식 (2)의 두 식의 곱으로 재구성된다.

$$\begin{aligned} P(O|q,\lambda) &= P(O|q,\lambda) \cdot P(q,\lambda) \\ &= \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T) \end{aligned} \quad (3)$$

HMM에 따른 관측열의 확률은 파라미터들로 구

성된 HMM이 허용하는 모든 가능한 상태열에서 열의 결합우로의 합으로 확장되어 다음 식(4)와 같이 표현된다.

$$P(O|\lambda)=\sum_{q\in Q}P(O,q|\lambda) \quad (4)$$

초기 상태 $t=1$에서 확률 $a_{s_0s_1}(\pi_{s_1})$로 천이가 시작되며, 관측 O_1는 출력 확률 $b_{s_1}(O_1)$로서 생성이 된다. 초기 상태 s_1에서 상태 s_2로의 천이는 천이 확률 $a_{s_1s_2}$로 이루어지며, 대응되는 상태 s_2에서의 관측 O_2를 생성될 확률은 $b_{b_2}(O_2)$가 된다. 이러한 과정은 상태 s_{T-1}에서 마지막 상태 s_T로 $a_{s_{T-1}s_T}$의 확률로 천이되어 기호 O_T를 출력 확률 $b_{s_T}(O_T)$로 생성할 때까지 계속된다. 이러한 과정을 정의에 의하여 직접 계산하면 모든 시간 $t=1,2,\cdots,T$에서는 진행 가능한 상태 수는 N개가 되어 계산 복잡도는 $O(N^T)$이 된다[5].

이러한 확률계산은 음성 구간에 따라 모델이 지수함수적으로 증가하는 상태 열을 갖기 때문에 쉽게 계산할 수 없고 계산량이 지나치게 방대해지므로 전향, 후향 알고리즘을 이용하여 HMM 모델의 관측열의 확률을 추정한다[6].

2.2 동적 패턴 매칭 기반 오류 보정

어휘 의미 패턴 오류 보정은 인식 문장을 의미적으로 분석하여 인식된 단어별로 의미 정보를 포함한 특정 스트링으로 대치하여 오류를 보정하는 방법이다. 기존 패턴에 대한 분석을 통해 어휘가 지니는 의미를 특정 스트링으로 구축하여 인식 시 입력되는 어휘가 기존 어휘에 존재하면 이미 구성되어진 특정 스트링으로 대치하여 사전에 오류를 보정한다. 기존 어휘의 분석을 통해 구축된 특정 스트링이 많을수록 오류를 저하시킬 수 있으므로 특정 스트링을 구축하기 위해 사전 어휘가 많이 필요하고 이에 따른 특정 스트링인 패턴이 많이 필요하며 오류 보정률은 패턴의 유무에 따라 많은 차이를 보이고 있다.

패턴 인식의 대상이 되는 패턴은 정적 패턴과 동적 패턴으로 나뉘고 고정된 영상의 패턴을 인식하는 경우와 시간에 따라 변화하는 패턴을 인식하는 경우이다. 패턴 인식은 패턴 매칭 방법에 의해 이루어지고 동적 패턴일 경우 시간상 늘이고 줄이는 신축을 허용하는 방식으로 패턴 매칭이 이루어지며 DTW (Dynamic Time Warping) 알고리즘을 이용한다[7].

인식 대상열의 길이가 다를 경우 매핑 함수를 적용하게 되며 신축 과정에서 인식 대상 열의 성분을 샘플링하고 신장 과정에서 보간하여 두 인식 대상 열의 길이를 같게 만든 후 비교하므로 인식 대상 열의 정보 손실로 인하여 패턴 매칭이 제대로 이루어지지 않아 비선형 패턴 매칭 방법을 사용하게 된다.

비선형 매핑 함수를 이용한 패턴 매칭 방법은 DTW 알고리즘을 사용한다. DTW 알고리즘은 인식 대상인 두 열의 각 성분에 대한 거리척도 값을 비용으로 설정하고 인식 대상 두 열이 이루는 격자상에서 각 열의 시작 성분에서 시작하여 끝 성분에 이르기까지 비용 테이블에 최소 비용을 순환적으로 택하여 저장하는 점화식을 이용하는 동적 계획법으로 매핑 함수를 찾아가면서 인식 대상 두 열을 비교하는 알고리즘이다. 최종적으로 끝 성분에서 비용 테이블에 저장되는 비용 값이 두 열에 대한 유사도로 표시된다. 매핑 함수의 궤적은 앞의 동적 계획법의 최적 탐색패스를 찾는 것과 같이 탐색 과정에서 최소 비용을 택하는 경로를 별도의 경로 테이블에 매 단계마다 저장하고 끝 성분에서 최종 최소 비용을 구한 후에 역추적하여 비교 열의 궤적을 찾는다[8].

어휘의 인식 시 일정한 패턴을 갖는 오류가 존재하므로 오류에 대한 카테고리를 미리 작성하여 오류 패턴을 미리 학습하여 발화된 문장과 인식할 문장을 비교하여 오류로 판단되는 문장에 대해 후처리 과정에서 오류 패턴을 보정하는 방법이다[9].

2.3 형태소 분석

기존의 형태소 분석기에 대한 연구는 형태소 분석 전처리 과정으로 입력되어진 어휘에서 발생하는 철자와 띄어쓰기에 대한 오류 교정에 중점을 두고 있었다. 철자와 띄어쓰기 오류를 해결하기 위해 통계적 방법을 사용하거나 Noisy Channel Model을 이용한 방법을 사용하였다. 통계적 방법은 기존 어휘에 대해 철자와 띄어쓰기에 대한 패턴을 구축하고 구축되어진 패턴에 의해 통계적 유사도를 측정하여 오류를 저하시키는 방법이며 Noisy Channel Model 방법은 통계적 유사도를 측정하는 방법은 비슷하나 기존 어휘에 대한 패턴을 구축하고 오류 패턴인 잡음 채널에 대한 모델을 구축하여 기존 어휘 모델에 대해 잡음 채널 모델을 제거하여 유사도를 측정한다[10].

패턴을 이용하는 형태소 분석 방법은 언어학적 표

현의 규칙과 사전적 정보를 이용하여 오류로 인식되어진 어휘를 오류가 아닌 어휘로 인식하기 위해 음소와 음절 단위의 변환 규칙을 작성하고 작성되어진 후보 어휘와 인식되는 어휘 간의 유사도를 측정하여 인식한다[11].

형태소 분석은 문장을 구성하는 각 어휘를 구성하고 있는 형태를 분석하는 방법이며 형태소 분석 결과는 어휘의 중의성을 해결하고 구문을 분석하며 의미역 부착, 기계 번역등과 같이 다양한 자연어처리 분야에서 활용되고 있으며 부정확한 분석 결과는 후행하는 모듈에 치명적인 영향을 미치게 되므로 높은 정확성을 필요로 한다. 하지만 특정 어절을 형성할 수 있는 형태소 조합이 하나 이상인 경우가 존재하기 때문에 형태론적 모호성이 발생하며 이로 인해 형태소 분석의 오류가 발생한다. 또한 여러 형태소들이 결합하여 어절을 형성할 때 발생하는 활용과 축약은 형태소 분석을 보다 어렵게 한다[12].

확률이나 통계를 활용하여 입력문을 분석하며 대량의 학습 말뭉치를 활용하여 추정한다. N-gram 모형이나 은닉 마코프 모형은 형태소 분석에서 가장 많이 쓰이는 통계 모형이다. 형태소 분석을 위한 N-gram 모형으로 이전 어절을 바탕으로 다음 어절의 형태소를 분석하는 방법을 사용하였다. N-gram 모형에서 unigram, bigram 단위로는 원거리 정보를 고려하지 못하여 부정확하게 분석하는 경우가 발생하지만, 이를 해결하기 위해 문맥을 늘리게 되면 자료부족문제가 심화된다는 문제점이 있다. 또한 어절 단위 확률 추정에 있어서 어절의 높은 생산성으로 인한 자료부족문제로 인해 어려움을 겪는다. 자료부족문제는 한정된 양의 말뭉치를 통해 확률을 추정하기 때문에 유발되며, 규칙 기반 방식에 비해 부정확한 결과를 보완하기 힘들다는 단점이 있다[13].

3. 시스템 모델

입력되어지는 어휘로부터 핵심어의 인식은 기존 패턴을 이용한 어휘의 의미 분석과 형태소 분석을 통하여 인식할 어휘와 확률적인 유사도를 측정하여 구한다. 핵심어의 경우 단순한 언어 모델로 구성되어 있기 때문에 문장 전체의 흐름을 파악하기 어렵고 기존 문장과의 패턴을 구축하여야 하므로 패턴과 오류 패턴 데이터를 필요로 한다.

문장이 핵심어로만 이루어진 경우 간결하고 사용자가 검색하고자 하는 어휘 또한 핵심어가 주류 이루기 때문에 핵심어의 대한 의미적으로 분석하기 힘들고 형태의 변형이 자주 일어난다. 이러한 단점을 보완하기 위해 핵심어에 대한 의미를 분석하고 핵심어의 형태를 분석하여 인식할 수 있는 시스템 모델을 그림 1과 같이 구성하였다.

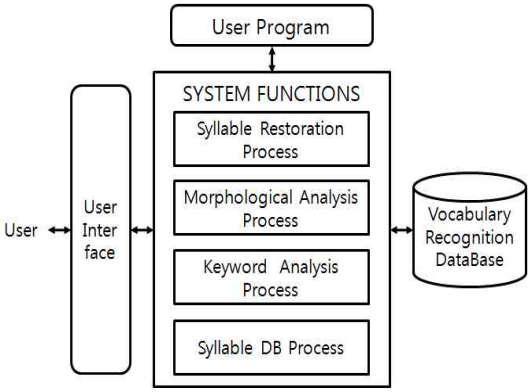

그림 1. 시스템 모델

3.1 음절 복원

음절 복원 처리에서는 어절을 구성하는 음소의 신뢰도와 음소에 대한 타 음소와의 유사률을 이용하여 각각의 음소에 대한 후보 음절을 생성한다. 생성된 후보 음절을 통하여 우선순위 어절을 선정한다.

사용자로부터 입력된 문장에 대하여 음절 복원을 먼저 수행하여 문장 전체의 의미가 어느 정도 일치하는지를 확인하며 복원된 문장은 패턴으로 정리되어 음절 데이터로 구축되며 재 복원을 위해 사용하게 된다. 복원된 문장에 핵심어를 따로 정리하고 오류를 파악하여 핵심어 패턴과 오류 패턴을 구축하기 위한 음절 복원을 수행한다.

인식되는 핵심어의 경우 오류의 패턴이 인식 후보 핵심어의 포함되어 있는 경우가 많으므로 후보 핵심어인 음절을 생성할 때에는 “산”이란 어휘인 경우 초성 자음 ‘ㅅ’과 모음 ‘ㅏ’, 종성자음 ‘ㄴ’에 대해 음소 유사률을 이용하여 구한다. 음절 복원 알고리즘은 다음과 같다.

```
if(itemA.size()==0) continue
question.m_context=0;
int n=itemA.size();
question.m_phoneA.resize(n);
        if(itemA[0].find("-*") != string::npos){
          question.m_context=-1;
          for(int i=0;i<n;i++)
           question.m_phoneA[i]=
                GetLeftContext(itemA[i]);
        }
        else if(itemA[0].find("*+") != string::npos) {
              question.m_context=1;
              for(int i=0;i<n;i++)
                  question.m_phoneA[i]=
                        GetRightContext(itemA[i]);
        }
        else{
             for(int i=0;i<n;i++)
                 question.m_phoneA[i]=
                          GetLabId(&itemA[i],false);
        }
```

인식할 핵심어로부터 후보 핵심어인 음절을 생성하여 초성 자음과 모음, 종성자음에 대해 각각의 음절을 비교하여 유사 패턴을 구한다. 알고리즘에서는 입력되어진 HasModel을 가지고 트리구조를 이용하여 초성, 중성, 종성을 생성하여 기존 list와 비교하여 m_maxLike를 구하는 형태로 구성하였다.

3.2 형태소 분석

형태소 분석 처리 과정은 형태소에 대한 규칙 기반 모형을 구축하여 구축되어진 분석 규칙을 이용하게 되는데 패턴의 부재와 기존 패턴과의 충돌로 인하여 분석을 실패하는 경우에는 확률 기반 형태소 분석 모형이 분석을 실패한 어절에 대해 다시 분석을 시도한다. 확률 기반 모형뿐만 아니라 다양한 종류의 형태소 분석 모형으로 대체하기 위한 패턴을 준비하여 다시 분석을 하여 대체 가능하게 한다. 기존의 형태소 분석에서의 분석 후보를 생성하는 것만을 목적으로 하는 일반 형태소 분석과 달리 주어진 핵심어에 대해 모든 형태에 대한 분석 후보를 생성한다. 각각의 후보에 대해 추가적으로 확률을 추정하여 순위를 만들어 필요시 순위에 의해 결정하는 것이 특징이다. 확률 추정은 어절 단위로 먼저 추정 하며 실패할 경우 핵심어 단위로 추정한다.

어절 단위 확률 모형에서 주어진 어절에 대하여 해당하는 분석 결과의 확률 추정 시 식 (5)와 같이 학습 문장으로부터 최대확률 유사도를 이용한다[14].

$$P(R|w)=\frac{frequence(R,w)}{frequence(w)} \quad (5)$$

R은 형태소 분석 결과이고 w는 어절을 나타낸다. 단순한 구조이면서 빠른 분석을 보장하지만 확률 모형은 규칙 모형과는 달리 하나 이상의 후보들을 생성한다.

형태소 단위 확률 모형은 형태소들이 결합할 때 발생하는 음운 현상을 고려하기 위하여 가능한 음운 현상 복원 후보를 모두 생성한 후, 각각의 복원 후보에 형태소 분석을 수행하여 결과를 취한다. 형태소 분석은 은닉 마코프 기반 모형을 이용하여 각 분석 후보를 생성하고 확률을 추정한다[15].

3.3 핵심어 의미 분석

핵심어 모델과 필러 모델 간의 구분이 명확한 경우 성능이 높게 나타나며 핵심어는 인식 대상 어휘로 사용되고 필러 모델은 핵심어를 제외한 어휘들로 핵심어 모델과 유사하지 않으면서 얼마나 모델링이 잘 되었느냐에 따라 성능이 좌우된다.

핵심어 검출 네트워크의 핵심어 모델들과 필러 모델들은 GMM(Gaussian Mixture Model)을 이용하여 모델링된다. GMM은 출력 확률 밀도 함수가 가우시안 밀도 혼합(Gaussian density mixture)인 1개의 상태만으로 구성된 CHMM(Continuous HMM)의 한 형태이다. 이러한 GMM은 다음과 같은 2개의 큰 특징을 지니고 있다. 첫째, GMM은 음향학적 클래스(Acoustic Class)의 집합을 모델링할 수 있다. 발성에 대응되는 음향공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적 클래스의 집합으로 잘 표현된다. 둘째, 단봉(unimodal) 가우시안 음소 모델은 평균 벡터(mean vector)와 공분산(covariance)으로 각 음소의 특징 벡터의 이산 집합으로 음소 분포를 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산 집합을 사용하여, 각각의 평균과 공분산을 가지게 함으로써 이들 두 모델의 특징을 혼합한 형태이다.

가우시안 혼합 밀도는 M성분(component)밀도의 가중합계로서 식 (6)에 의해 얻어진다.

$$p(x|\lambda)=\sum_{i=1}^{M}c_i b_i(x) \quad (6)$$

x는 d-차원 랜덤 벡터이며, $b_i(x), i=1,...,M$는 i번째의 성분(component)밀도이고, c_i, $i=1, 2, ..., M$는 i

번째 혼합 밀도 가중치(mixture weight)이며 각 혼합 밀도의 가중치는 다음 식 (7)과 같이 제한된다.

$$\sum_{i=1}^{M} c_i = 1 \tag{7}$$

각 성분 밀도는 평균 μ_i과 공분산 Σ_i를 가지는 d-변량(variate) 가우시안 함수이다. 가우시안 혼합 밀도는 모든 성분 밀도의 혼합 밀도 가중치와 공분산 행렬, 평균 벡터로 구성된다. 따라서 GMM의 파라미터를 구하면 아래 식 (8)과 같은 모델을 만든다.

$$\lambda = \{c_i, \mu_i, \Sigma_i RIGHT \quad i = 1, \ldots, M \tag{8}$$

그림 2는 이러한 GMM을 사용한 음소 모델을 나타내었다.

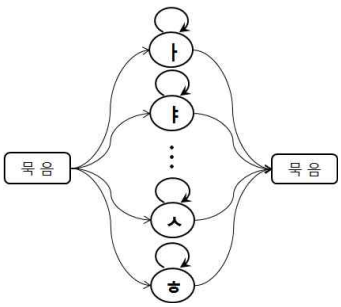

그림 2. GMM의 음소 모델 생성 과정

모델 학습은 주어진 학습 음성으로부터 학습 특징 벡터의 분포와 가장 잘 맞는 GMM 파라미터를 추정하는 것이다. 음소 인식 과정에서 음소별 GMM의 평균, 공분산과 CHMM의 상태 천이 확률을 이용한 연속 음소 인식 네트워크를 구성하고 최대 사후 확률을 갖는 음소열을 발생하며 후보 단어 선택 단계로 제공된다. 핵심어의 오검출을 방지하기 위해 핵심어가 실제 발성 되었는지를 검증하고 신뢰도를 계산한다. 핵심어 검출의 평가 기준은 미검출률과 FAR로 표현하는데 미검출률은 테스트 문장에서 출현한 핵심어 검출기가 제대로 검출하지 못한 경우를 표시하고, FAR로는 각 핵심어 당 FA의 출현횟수를 평가 시간으로 정규화한 FA/KW/HR을 사용하여 표시한다[16].

3.4 음절 DB 관리

음절 DB 관리에서는 입력된 핵심어에 대해 형태를 음소 단위로 분류하여 복원된 음절과 비교하기 위하여 음절 데이터를 구축하여 형태소별 패턴을 정리하며 음절로부터 핵심어 모델과 필러 모델을 구축하고 오인식의 일정한 패턴을 갖는 오류 패턴 DB를 그림 3과 같이 구축하였다.

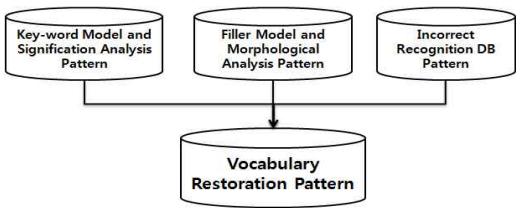

그림 3. 오류 패턴 DB

오인식의 일정한 패턴을 바탕으로 오인식 패턴을 정리하여 사전 DB를 구축한 오인식 패턴 DB는 오인식이 자주 발생하는 음소와 음소의 오인식 패턴을 쌍으로 묶어 오류 패턴 사전 DB를 구축하였다.

또한 핵심어의 의미를 분석한 패턴을 정리하여 DB로 구축하였으며 형태소 분석을 통한 형태 변형에 대한 DB를 구축함으로 복원 시 활용하였다.

구축되어진 데이터는 재인식 시 오류 패턴 DB로 활용하고 입력되어진 음절로부터 핵심어 모델과 필러 모델의 유사률과 오류 보정률을 계산하기 위하여 사용되며 의미적으로 사용자가 제시한 의미와 일치하는 패턴으로 재구성하여 음절을 복원시킨다.

4. 실험 결과 및 분석

본 논문에서 음절 복원 알고리즘에서 의미 분석과 형태소 분석을 이용한 핵심어 인식 시스템 모델을 구성하여 인식 실험을 수행하였다. 본 실험에서는 단말기에서 사용되는 단어를 선정하여 총 20개를 표 1과 같이 선별하였고 후처리 실험을 위해 5명의 사용자가 100회의 데이터를 구축하였다.

어휘는 실내 환경과 잡음 환경에서 이동기기 등에 내장되어 있는 내장형 마이크로폰을 사용하여 16kHz Mono로 녹음 하였고, 16bit PCM 양자화를 사용하였다. 실험 어휘는 실내 10명, 실외 5명 등 총 15명의 성인 남성이 참가하였다.

녹음된 데이터는 인식기 학습을 위해 특징 추출 방법을 사용하였고 인식기는 SITEC에서 개발한 ECHOS[17]를 이용하였다. ECHOS는 각 단어별 데

이터로 학습된 인식 모델을 이용하여 발화된 단어의 인식 가능한 단어들의 인식 가능 확률을 표현하고 최대값을 가지는 단어를 최종 결과로 선정한다.

표 1. 실험 단어

No.	문 장	No.	문 장
1	서울 잠실역 검색합니다.	11	서울 잠실역을 다시 검색합니다.
2	서울 송파역 검색합니다.	12	서울 송파역을 다시 검색합니다.
3	서울 논현역 검색합니다.	13	서울 논현역을 다시 검색합니다.
4	서울 구로역 검색합니다.	14	서울 구로역을 다시 검색합니다.
5	서울 여의도역 검색합니다.	15	서울 여의도역을 다시 검색합니다.
6	도착지로 선택합니다.	16	빠른 경로를 선택합니다.
7	도착지로 잠실역을 선택합니다.	17	최단 경로를 선택합니다.
8	도착지로 송파역을 선택합니다.	18	다른 경로를 선택합니다.
9	출발지로 선택합니다.	19	고속도로 경로를 선택합니다.
10	출발지로 잠실역을 선택합니다.	20	무료도로 경로를 선택합니다.

음소 유사률의 구성에 따른 음소 유사률의 정확성을 확인하기 위하여 특징 추출 방법을 사용하여 음소 유사률을 구성하고 어휘 인식 과정에서 의미 분석과 형태소 분석을 이용한 핵심어 인식 시스템 모델에 적용하였다.

표 2는 기존의 확률 패턴 매칭, 동적 패턴 매칭, 어휘 의미 패턴, 제안방법의 인식률과 오류 보정률을 비교 실험한 결과이다.

인식률 실험을 한 결과 보정 전 인식률에서 확률 패턴 매칭을 이용한 인식률은 80.2%, 동적 패턴 매칭의 경우 82.6%, 어휘 의미 패턴의 오류 보정의 경우 84.5%의 인식률을 보였으며 제한 방법의 인식률은 84.6%의 인식률을 보였다.

표 2. 오류 보정률 비교

오류 보정	보정 전 인식률(%)	보정 후 인식률(%)	보정률 (%)
확률 패턴 매칭	80.2	84.5	4.3
동적 패턴 매칭	82.6	88.1	5.5
어휘 의미 패턴 매칭	84.5	90.6	6.1
제한 방법	84.6	90.9	6.3

오류 보정을 수행하지 않은 인식률을 오류 보정을 통해 다시 인식률을 확인한 결과 확률 패턴 매칭은 84.5%, 동적 패턴 매칭의 경우 88.1%, 어휘 의미 패턴의 오류 보정의 경우 90.6%의 인식률을 보였으며 제한 방법의 인식률은 90.9%의 인식률을 보였다. 보정률로 비교하여 나타낸 경우 확률 패턴 매칭은 4.3%, 동적 패턴 매칭은 5.5%, 어휘 의미 패턴의 오류 보정은 6.1%의 보정률을 보였으며, 제안한 방법을 이용한 경우 6.3%의 보정률을 나타냈으며 확률 패턴의 인식률 보다 2.0%의 인식 향상률을 보였다.

5. 결 론

본 논문에서는 음절 복원 알고리즘에서 의미 분석과 형태소 분석을 이용한 핵심어 인식 시스템을 제안하였다.

인식된 음소 열을 형태소 분석 과정을 수행한 후 음소가 갖는 형태를 찾아 데이터로 정리하고 음절 복원 알고리즘을 통해 형태적 변형이 일어나기 이전의 문자열로 복원한다. 입력된 문장으로부터 형태소 분석 과정의 수행은 분석 후보를 다수 생성한 후 최적의 후보를 선택하는 방법을 취했으며 어절의 의미를 형태적으로 파악하여 문법적인 어절은 인식하고 비문법적인 어절은 오류 보정을 수행하였다.

의미적으로 분석하기 어려운 핵심어 문장을 복원하여 전체적인 어휘의 오인식을 감소시켰으며 형태소 분석을 이용한 음절 복원 알고리즘에서 핵심어에 대한 오인식을 감소시키므로 오류 패턴 DB의 양을 줄일 수 있는 장점을 얻었으며, 검색 시 속도와 인식률에서 기존 확률 패턴 매칭, 동적 패턴 매칭, 어휘 의미 패턴 오류 보정 시스템보다 나은 결과를 얻을 수 있었다.

시스템 성능 평가는 음소 유사률을 측정하였고 신뢰도를 이용하여 평가하였다. 어휘 인식 과정에서 오류로 판명된 어휘에 대하여 오류 보정을 수행한 결과 6.3%의 보정률을 나타냈으며 어휘 의미 패턴의 인식률 보다 2.0%의 인식 향상률을 보였다.

참 고 문 헌

[1] Eiichi Tanaka and Tamotsu Kasai, "Synchronization and Substitution Error-correcting codes for the Levenshtein Metric," *IEEE Trans. Information Theory,* Vol.IT-22, No.2, pp. 156- 176, 1976.

[2] E. K. Ringer and J. F. Allen, "A fertility channel model for post-correction of continuous speech recognition," Proc. ICSLP, pp. 897-900, Oct, 1996.

[3] 박미성, 김미진, 김계성, 최재혁, 이상조, "연속 음성인식 후처리를 위한 음절 복원 rule-based 시스템과 형태소분석기법의 적용," 대한전자공학회 논문지, 제36권, 제3호, 47-57쪽, 1999년.

[4] 문광식, 김회린, 정재호, 이영직, "가변어휘 단어 인식에서의 미등록어 거절 알고리즘의 성능비교," 신호처리합동학술대회 논문집, 제12권, 제1호, 305-308쪽, 1999년.

[5] 조시원, 이동욱, "음성 인식 후처리를 위한 연속 음절 문장의 키워드 추출 알고리즘," 대한전기학회, 학술대회 논문집, 심포지엄 논문집 정보 및 제어부문, 170-171쪽, 2008년 4월.

[6] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, 1993.

[7] 안찬식, 오상엽, "MLHF 모델을 적용한 어휘 인식 탐색 최적화 시스템," 한국컴퓨터정보학회 논문지, 제14권, 제10호, 217-223쪽, 2009년 10월.

[8] 송원문, 김명원, "문맥 및 사용 패턴 정보를 이용한 음성인식 후처리," 정보처리학회 논문지, 제13-B권, 제5호, 553-560쪽, 2006년.

[9] 김동주, 김한우, "문맥가중치가 반영된 문장 유사도 척도," 대한전자공학회 논문지, 제43권, 제6호, 496-504쪽, 2006년.

[10] 이승욱, 이도길, 임해창, "형태소 분석 및 품사 부착을 위한 말뭉치 기반 혼합 모형," 한국컴퓨터정보학회 논문지, 제13권, 제7호, 11-18쪽, 2008년 12월.

[11] D. Lee, H. Rim and D. Yook, "Automatic Word Spacing using Probabilistic Models Based on Character n-grams," *IEEE Intelligent Systems,* Vol. 22, No. 1, pp. 28-35, Jan.-Feb. 2007.

[12] 여상화, "한영 모바일 번역기를 위한 강건하고 경량화된 한국어 형태소 분석기", 한국컴퓨터정보학회 논문지, 제14권 제2호, 191-198쪽, 2009년 2월.

[13] S. Kang and C. Woo, "Automatic Segmentation of Words using Syllable Bigram Statistics," Proc. Natural Language Processing Pacific Rim Symposium, pp. 729-732, Nov. 2001.

[14] M. Ostendorf, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *Speech and Audio Processing, IEEE,* Vol. 4, pp. 360-378, 1996.

[15] 한동조, 최기호, "음성인식 후처리에서 음소 유사율을 이용한 오류보정에 관한 연구," 한국ITS학회 논문지, 제6권, 제3호, 77-86쪽, 2007년 12월.

[16] M. F. Gales, "Model-based techniques for noisie robust speech recognition," Ph. D. dissertation, University of Cambridge, Sept, 1995.

[17] 음성정보기술산업지원센터, "한국어 음성인식 플랫폼 사용자 매뉴얼(ECHOS Manual)," 135-308쪽, 2006년.

안 찬 식

2002년 광운대학교 컴퓨터공학과 석사
2004년 광운대학교 컴퓨터공학과 박사수료
관심분야 : 음성인식, 분산처리, 음성/음향 신호처리

오 상 엽

1999년 광운대학교 전자계산학과 박사
1993년~현재 경원대학교 IT대학 컴퓨터소프트웨어 교수
관심분야 : 소프트웨어공학, 버전관리, 소프트웨어 재사용, 형상관리, 객체지향, 음성인식, 분산 처리, 음성/음향 신호처리