음성인식을 위한 의사(擬似) N-gram 언어모델에 관한 연구

A Study on Pseudo N-gram Language Models for Speech Recognition

오 세 진, 황 철 준, 김 범 국, 정 호 열, 정 현 열

Se-Jin Oh, Chul-Joon Hwang, Bum-Koog Kim, Ho-Youl Jung, Hyun-Yeol Chung

요 약

본 논문에서는 대어휘 음성인식에서 널리 사용되고 있는 N-gram 언어모델을 중규모 어휘의 음성인식에서도 사용할 수 있는 의사(擬似) N-gram 언어모델을 제안한다. 제안방법은 ARPA 표준형식 N-gram 언어모델의 구조를 가지면서 각 단어의 확률을 임의로 부여하는 비교적 간단한 방법으로 1-gram은 모든 단어의 출현확률을 1로 설정하고, 2-gram은 허용할 수 있는 단어시작기호 <s>와 WORD 및 WORD와 단어종료기호 </s>의 접속확률만을 1로 설정하며, 3-gram은 단어시작기호 <s>와 WORD, 단어종료기호 </s> 만의 접속을 허용하며 접속확률을 1로 설정한다. 제안방법의 유효성을 확인하기 위해 사전실험으로서 국어공학센터(KLE) 단어음성에 대해 오프라인으로 평가한 결과, 남성 3인의 452 단어에 대해 평균 97.7%의 단어인식률을 구하였다. 또한 사전실험결과를 바탕으로 1,500단어의 중규모 어휘의 증권명을 대상으로 온라인 인식실험을 수행한 결과, 남성 20명이 발성한 20단어에 대해 평균 92.5%의 단어인식률을 얻어 제안방법의 유효성을 확인하였다.

ABSTRACT

In this paper, we propose the pseudo n-gram language models for speech recognition with middle size vocabulary compared to large vocabulary speech recognition using the statistical n-gram language models. The proposed method is that it is very simple method, which has the standard structure of ARPA and set the word probability arbitrary. The first, the 1-gram sets the word occurrence probability 1 (log likelihood is 0.0). The second, the 2-gram also sets the word occurrence probability 1, which can only connect the word start symbol <s> and WORD, WORD and the word end symbol </s>. Finally, the 3-gram also sets the word occurrence probability 1, which can only connect the word start symbol <s>, WORD and the word end symbol </s>. To verify the effectiveness of the proposed method, the word recognition experiments are carried out. The preliminary experimental results (off-line) show that the word accuracy has average 97.7% for 452 words uttered by 3 male speakers. The on-line word recognition results show that the word accuracy has average 92.5% for 20 words uttered by 20 male speakers about stock name of 1,500 words. Through experiments, we have verified the effectiveness of the pseudo n-gram language modes for speech recognition.

Keywords : Pseudo N-gram language models, Context-Free Grammar, Finite State Network Multi-Pass Search, HTK, Julius

I. 서 론

음성은 인간이 사용하는 가장 기본적인 의사소통을 위한 수단이며, 편리함과 경제성의 측면에서 다른 방법에 비해 우수한 특성을 갖는다. 그동안 음성과 관련된 연구들은 각기 개별적으로 이루어져 왔으며, 그러한 분야들은 언어학, 음성학, 음운학, 생리학, 해부학 등 다양한 학문적인 배경 하에 진행되어 왔다. 그러한 결과들이 신호처리 기술, 기억장치의 대용량화, 그리고 고속의 정보처리 기술의 발달 등 급격한 기술의 발전으로 인해 단순히 시험적인 결과가 아닌 실용적인 측면에서 그러한 결과들을 활용하는 연구가 활발히 진행되어 왔으며, 한편으로는 계산이론적인 측면에서 음성처리와 관련된 다양한 연구들이 이루어지게 되었다[1]. 또한 음성을 이용하여 응용 프로그램을 동작시키고 음성인식 시스템을 이용하여 인터넷상에서의 정보검색, 전화망을 이용한 증권 거래 및 시세를 조회하는 등 음성인식이 우리 생활의 일부로써 활용되기 시작하고 있다[2].

일반적으로 음성인식에 사용되는 기본적인 인자로는 음향모델과 언어모델을 예로 들 수 있다[3]. 음향모델이란 음성신호의 신호적인 특성을 모델링하는 것으로 입력신호와 직접적으로 매칭되는 표준패턴에 해당되며, 언어모델은 음향적인 신호가 아닌 언어적인 단위로서 인식어휘에 해당되는 단어나 음절들간의 언어적인 순서관계를 나타내는 것이다. 특히 언어모델은 인식 대상이 음절이나 단어인 경우에는 의미가 없지만, 인식대상이 구, 절 또는 문장인 경우에는 언어모델의 필요성은 매우 커지게 된다.

음성인식에서 사용되는 언어모델은 통계에 기반한 언어모델과 유한상태 네트워크(Finite State Network; FSN) 형태에 기반한 모델링 방법들이 널리 사용되고 있다[4]. 통계적 언어모델링은 수집된 음성인식 대상의 텍스트 코퍼스(Corpus)를 대상으로 수집된 자료로부터 텍스트를 구성하는 단어들간의 접속관계를 확률적으로 모델링한 것을 말한다. 통계적 언어모델은 1,000~20,000 단어를 대상으로 한 대용량 음성인식을 수행하는데 경우에 일반적으로 적용된다. 이와 반대로, 유한상태 네트워크(FSN) 형태의 경우 인식대상 영역이 제한적이고, 어휘수가 크지 않으며, 화자의 발성이 형식적인 경우에 사용된다.

최근의 음성인식에 관한 연구는 소용량에서 대용량까지 취급할 수 있도록 다양한 분야에서 활발한 연구가 수행되고 있다. 따라서 음성인식 엔진도 다양하게 개발되고 있다. 하지만, 현재 개발된 대부분의 음성인식 엔진은 사용되는 언어모델에 따라 소용량 인식엔진, 대용량 인식엔진 등과 같이 특정분야에서만 사용할 수 있도록 제작되고 있다. 이러한 경우 인식성능이 좋은 소용량 인식엔진을 대용량 인식엔진으로 적용하거나, 이와 반대로 대용량 인식엔진을 소용량 인식엔진에 적용하는데는 많은 어려움이 따른다. 여기서 언어모델의 경우 대용량 인식엔진에서 사용되는 통계적 N-gram 언어모델을 중규모 어휘를 갖는 한정된 태스크에 적용할 경우, 언어모델 작성을 위한 텍스트 데이터의 부족으로 인해 언어모델의 신뢰성을 보장할 수 없는 문제점이 발생된다.

따라서 본 논문에서는 이와 같은 문제점을 해결하기 위해, 좋은 성능의 대용량 음성인식엔진과 중규모 어휘의 한정된 태스크에서도 적용할 수 있는 언어모델로서 의사(擬似) N-gram 언어모델을 제안한다. 제안방법은 통계적 방법과 같이 ARPA 표준형식[8]의 N-gram 언어모델 구조로 구성되어 있고, 각 단어의 확률을 임의로 부여하는 비교적 간단한 방법으로서 유한상태 네트워크(FSN) 형태로도 해석이 가능한 방법이다.

제안방법의 유효성을 확인하기 위해, 유한상태 네트워크(FSN), 문맥자유문법(CFG), 의사 N-gram 언어모델을 452 단어 규모에 대해 각각 작성하여 단어인식실험을 수행하고, 본 연구실에서 구축한 1,500단어 규모의 음성인식 증권조회 시스템에 적용하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 언어모델의 필요성과 유한상태 네트워크 언어모델, 통계적 언어모델에 대해서 설명하고, 3장에서는 본 논문에서 제안한 의사 N-gram 언어모델에 대해서 설명한다. 4장에서는 제안한 언어모델을 이용한 음성인식 실험 및 고찰에 대해서 설명한 후, 마지막으로 5장에서 결론을 맺는다.

II. 음성 언어 모델링

2.1 언어모델의 필요성

음성인식에 있어서 언어모델은 발음사전에 등록된 어휘에 대한 구문 제약으로 사용되며, 음성인식에 적용할 경우의 장점[5,6]을 요약하면 다음과 같다.

첫 번째는 다중 지식을 사용함으로써 인식오류를 줄일 수 있다는 점이다. 음성인식은 크게 음향적 지식과 언어적 지식이 사용되는데, 음향적 지식은 음성의 신호적인 특성의 오류를 줄이기 위해 사용되며, 언어적 지식은 인식된 인식 어휘의 열들에 대한 오류를 줄이는데 사용된다는 점에서 다중 지식이 필요하다고 할 수 있다.

두 번째는 인식 어휘들간의 구문 제약은 문장 인식오류를 줄인다는 점에서 언어모델이 필요하다는 점이다. 음성인식기는 발성된 음성을 인식한 후 사용자가 원하는 처리를 수행하기 위해서는 문장 또는 구, 절 형태의 인식을 수행하여야 한다. 문장 또는 구, 절에 대해 어휘들 사이의 구문 제약이 존재하게 되며, 구문 제약을 규정하기 위해 언어모델이 사용된다. 그리고 언어모델은 음성신호의 음향적인 분석결과에 대한 오류를 보정하여 인식오류를 줄이기 때문에 필요하다고 할 수 있다.

2.2 유한상태 네트워크(FSN) 언어모델

형식언어인 유한상태 네트워크(FSN)에 기반한 언어모델은 소규모 어휘를 대상으로 명령 및 제어를 위한 짧은 문장을 인식하기 위해 사용된다. 만약 문장, 구 또는 절을 문맥자유문법(Context-Free Grammar; CFG) 형태로 표시할 경우 음성인식은 유한상태 네트워크(FSN) 형태로 변환하여 인식을 수행함으로써 단어 앞/뒤의 순서관계가 명확하여 인식오류를 줄일 수 있게 된다. 유한상태 네트워크(FSN)를 이용할 경우, 인식대상 어휘로 구성된 문법을 이용해서 하나의 큰 탐색 네트워크를 만들고, 이 탐색 네트워크와 입력된 음성신호에 대한 매칭을 수행하게 된다. 이때, 탐색 네트워크가 인식 어휘들간의 순서관계를 나타내는 언어모델로 사용되게 된다. 유한상태 네트워크(FSN) 형태의 언어모델로는 하나의 인식 어휘에 올 수 있는 어휘수가 인식기에서 인식할 총 어휘수와 동일한 no-gram 문법, 그리고 단어간에 올 수 있는 제약을 이진값으로 제약하는 word-pair 문법이 있다[7].

2.3 통계적 언어모델

언어모델은 n 개의 단어들로 구성된 단어열 W =

$\{w_1, \cdots, w_n\}$ 이 주어질 경우 언어모델의 확률은 식 (1)에 의해 구할 수 있다[8].

$$P(w_1 \cdots w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1 \cdots w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_1 \cdots w_{i-1}) \quad (1)$$

하지만 여러 가지 단어의 조합으로 인해 조건부 확률 $P(w_i|w_1 \cdots w_{i-1})$ 을 구하는 것은 실제로 불가능하다. 만약 V개의 단어가 사용된다고 가정하면 확률 $P(w_i|w_1 \cdots w_{i-1})$ 을 완전히 구하기 위해서는 V^i개의 확률을 계산하여야 한다. 따라서 실제로는 모델 구성을 위해서 단어 이력(history) $w_1 \cdots w_{i-1}$ 을 같은 종류의 항으로 분할하고 모델의 파라미터 수를 감소시킬 필요가 있다. 이때 어느 시점에서 발생하는 사건의 확률은 바로 이전의 N개 시점까지 발생하는 사건에 영향을 받는 N중 마르코프 과정을 따른다. 이러한 단어의 발생을 N-1중 마르코프 과정으로 근사화한 모델을 N-gram 모델이라고 한다[8]. N-gram 언어모델에서는 어느 시점에서의 단어 발생은 바로 이전의 N-1 단어에 의존하는 것으로 가정하면 다음 식 (2)와 같이 쓸 수 있다.

$$P(w_n|w_1 \cdots w_{n-1}) = P(w_n|w_{n-N+1} \cdots w_{n-1}) \quad (2)$$

식 (2)에서 나타낸 것과 같이 현재 단어의 예측을 위해 이전의 한 단어를 고려한다면 2-gram 이라 하고, 이전의 두 단어를 고려한다면 3-gram 이라고 한다. 또한 현재 단어 자체를 예측한다면, 단어 생성확률도 되지만 1-gram 이라고 하며, 현재 음성인식에서 널리 사용되고 있다.

N-gram 언어모델의 확률은 학습 데이터 중에서 나타나는 단어 N개와 N-1개로부터 다음 식에 의해 추정할 수 있다.

$$P(w_n|w_{n-N+1} \cdots w_{n-1}) = \frac{C(w_{n-N+1} \cdots w_n)}{C(w_{n-N+1} \cdots w_{n-1})} \quad (3)$$

여기서 C는 단어열 $w_1 \cdots w_n$ 이 학습 데이터 중에 출현하는 회수를 나타낸다.

2.4 Back-off Smoothing

N-gram 언어모델을 추정할 때 추정에 사용된 코퍼스(Corpus) 중에 출현하지 않는 N 단어가 존재하게 되는데 이때 추정 확률은 0이 된다. 또한 코퍼스에서 출현하더라도 출현빈도가 적은 단어열에 대해서는 통계적으로 신뢰성 있는 확률을 추정하는 것은 어려운 일이다. 따라서 이러한 문제를 해결하기 위해 여러 가지 스무딩(Smoothing) 방법[9]이 제안되고 있지만, 여기서는 그 중에서 대표적인 Back-off 스무딩에 대해 기술한다.

Back-off 스무딩은 코퍼스 중에 출현하지 않는 N-gram의 확률을 낮은 차수의 (N-1)-gram의 확률로 추정하는 방법을 말한다. 이 방법에서는 N개의 조합으로 구성된 단어의 출현회수 C를 각각 사용하게 되며, 출현회수를 보정한 회수 C^* ($C^* < C$) 를 사용한다. 이와 같이 회수 C를 C^*에 보정한 것을 디스카운팅(discounting)이라고 한다. 디스카운팅에 있어서 존재하는 N개의 조합으로 구성된 단어의 출현회수가 적기 때문에 확률의 총합이 1이 된다면 이때 출현하지 않는 N개의 조합으로 확률값을 분배하게 된다.

Back-off 스무딩을 이용한 단어 2-gram의 추정은 $C(w_1w_2) > 0$ 인 경우에는 식 (4)와 같다.

$$P(w_2|w_1) = \frac{C^*(w_1w_2)}{C(w_1)} = \frac{C^*(w_1w_2)}{C(w_1w_2)} \frac{C(w_1w_2)}{C(w_1)} = d_{C(w_1w_2)} \frac{C(w_1w_2)}{C(w_1)} \quad (4)$$

또한 $C(w_1w_2) = 0$ 인 경우는 식 (5)와 같이 정의된다.

$$P(w_2|w_1) = \zeta(w_1)P(w_2) \quad (5)$$

여기서,

$$\zeta(w_1) = \frac{1 - \sum_{w_2: C(w_1w_2)>0} P(w_2|w_1)}{\sum_{w_2: C(w_1w_2)=0} P(w_2)} \quad (6)$$

을 나타내며, 식 (5)의 $\zeta(w_1)$ 은 식 (6)에 의해 계산된다.

스무딩에 의한 보정회수 C^*를 구하는 방법에는 Linear discounting, Good-Turing discounting, Witten-Bell discounting 방법 등이 제안되고 있다[9].

Ⅲ. 의사(擬似) N-gram 언어모델

본 장에서는 대어휘 연속음성인식 시스템에서 널리 사용되고 있는 N-gram 언어모델을 중규모 어휘 음성인식에서도 사용할 수 있도록 본 논문에서 제안한 의사 N-gram 언어모델에 대해 요약한다.

본 논문에서 제안한 의사 N-gram 언어모델은 유한상태 네트워크(FSN)에 기반한 언어모델과 같이 인식대상 영역이 제한적이며, 사용되는 어휘수가 그리 크지 않고, 발성되는 문장 또는 단어의 형태가 패턴화 되어 있는 경우에 사용할 수 있다.

제안방법의 기본적인 아이디어는 먼저 한정된 단어를 인식대상으로 할 경우, 패턴화된 하나의 단어를 사용자가 발성한다고 가정하면, 발성된 단어에 대해 언어모델의 구조를 일반적인 단어 N-gram 형태로 표현할 수 있다는 점이다. 이때 발성된 단어에는 단어의 시작과 종료 시점에 묵음(Silence)이 포함되게 되는데 이를 언어모델의 구문적 제약으로 이용한다면 [시작기호] → [임의의 단어] → [종료기호] 형태로 표현할 수 있다. 따라서 단어의 시작과 종료기호를 하나의 단어로 취급하고 이와 같이 표

현할 수 있는 단어의 접속만을 허용하고, 그 외의 단어 접속은 허용하지 않는다고 가정한다면 ARPA 표준형식 [8]의 단어 N-gram 언어모델을 나타낼 수 있게 된다.

그림 1은 한정된 영역의 단어인식을 위한 의사 N-gram 언어모델의 구조를 나타낸 것이다. 그림 1에서 순방향 2-gram과 역방향 3-gram이 있는데 이는 최근의 대용량 인식엔진은 Multi-Pass 탐색[12,13]을 수행하여 인식을 수행하는 것이 일반적이므로 이를 위한 것이다. 그림 1에 나타낸 것과 같이 단어 1-gram의 경우, 모든 단어의 출현확률은 1(대수우도 0)로 설정하고, 2-gram의 경우 단어 시작기호 <s>와 WORD 그리고 WORD와 단어 종료기호 </s>로의 접속확률만을 1로 설정한다. 또한 3-gram은 단어 시작기호 <s>와 WORD, 단어 종료기호 </s>의 접속확률을 1로 설정하고, 이외의 접속확률은 아주 작은 값으로 설정하여 실제로 인식후보에는 출현하지 않도록 한다.

순방향 2-gram	역방향 3-gram
\data\ ngram 1=3 ngram 2=2 \1-grams: 0.0 <s> -99.0 0.0 </s> -99.0 0.0 WORD -99.0 \2-grams: 0.0 <s> WORD 0.0 WORD </s> \end\	\data\ ngram 1=3 ngram 2=2 ngram 3=1 \1-grams: 0.0 <s> -99.0 0.0 </s> -99.0 0.0 WORD -99.0 \2-grams: 0.0 </s> WORD-99.0 0.0 WORD <s> -99.0 \3-grams: 0.0 </s> WORD <s> \end\

그림 1. 단어인식을 위한 의사 N-gram.
Fig. 1. Pseudo N-gram for word recognition.

그림 2는 연속(연결) 단어인식에 대한 의사 N-gram 언어모델의 구조를 나타낸 것이다. 연속(연결) 단어인식을 위한 의사 N-gram 언어모델은 연속 단어에 포함된 모든 단어에 대해서 접속할 수 있도록 구성되어 있다.

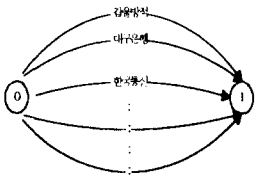
순방향 2-gram	역방향 3-gram
\data\ ngram 1=3 ngram 2=1 \1-grams: 0.0 <s> 0.0 0.0 </s> 0.0 0.0 WORD 0.0 \2-grams: 0.0 WORD </s> \end\	\data\ ngram 1=3 ngram 2=1 ngram 3=1 \1-grams: 0.0 <s> 0.0 0.0 </s> 0.0 0.0 WORD 0.0 \2-grams: 0.0 </s> WORD0.0 \3-grams: 0.0 </s> WORD <s> \end\

그림 2. 연속(연결) 단어인식을 위한 의사 N-gram.
Fig. 2. Pseudo N-gram for continuous(connect) word recognition.

그림 3은 의사 N-gram 언어모델을 이용할 경우 인식 어휘에 대한 발음사전 구조를 나타낸 것이다. 발음사전은 HTK[10] 형식을 이용하는데, HTK와의 차이점은 언어모델의 단어심벌을 나타내는 제1항에 의사 N-gram 언어모델의 대표 언어심벌인 WORD만을 표기하여 이 WORD에 모든 인식 어휘가 포함되도록 구성한 것이다. 그림 3과 같은 발음사전을 이용할 경우 의사 N-gram 언어모델을 유한상태 네트워크(FSN)와 문맥자유문법(CFG)에 기반한 형식언어 형태로 표현할 수 있으며 그림 4에 각각 나타내었다. 그림 4의 (a)는 0상태에서 1상태로 전이하는 동한 인식어휘들이 모두 출현하고, 0상태에 단어시작기호, 1상태에 단어종료기호가 각각 해당된다. 그림 4의 (b)는 의사 N-gram 언어모델을 비종단 기호인 START, WORDS, SIL_B, SIL_E, WORD와 종단기호인 word, <s>, </s>에 의해 나타낸 것이다. 이와 같이 의사 N-gram 언어모델은 형식언어로 해석할 수 있으며, 구조가 간단하여 태스크에 따라 쉽게 적용이 가능하다.

인식을 위한 발음사전의 구조		
<s>	[]	silB
</s>	[]	silE
WORD	[갑을방적]	g aa b+ U l b~ aa ng z~ axr g+
WORD	[대구은행]	d ae g~ uh U n hh~ ae ng
WORD	[한국통신]	hh aa n g~ uh g+ t ao ng s ih n
⋮	⋮	⋮

그림 3. 발음사전의 구조.
Fig. 3. Structure of pronunciation lexicon.



START SIL_B WORDS SIL_E
WORDS WORD
WORD word
⋮ ⋮
SIL_B <s>
SIL_E </s>

(a) 유한상태 네트워크 (b) 문맥자유문법
(a) Finite state network (b) Context-free grammar

그림 4. 의사 N-gram 언어모델을 유한상태 네트워크와 문맥자유문법 언어모델로 표현한 예.
Fig. 4. Example of finite state network and context-free grammar according to Fig. 1.

Ⅳ. 의사 N-gram 언어모델을 이용한 단어인식 실험

4.1 기본 인식 시스템

그림 5에 본 논문에서 이용한 Julius 인식엔진[11]의 전체 구성을 나타내었다. 인식 엔진은 음향모델, 언어모델, 발음사전, 그리고 인식 부분으로 나뉜다.

우선 음향모델을 작성하는데 사용된 기본단위는 확장성을 고려하여 묵음에 대한 모델을 포함한 50개의 유사음소단위(Phoneme Likely Units; PLUs)를 이용하였으며 표 1에 나타내었다. 음향모델의 학습은 HTK를 이용하여 5상태 4천이의 1, 4, 8, 12, 16혼합 분포를 각각 가지며 초기상태와 최종상태에는 확률분포가 없는 문맥독립 연속분포 HMM으로 구성하였다.

언어모델은 452 단어 및 증권명 조회를 위해 인식대상으로 선정한 1,500 단어에 대해 본 논문에서 제안한 의사 N-gram을 이용하여 의사 2-gram과 의사 3-gram 언어모델을 각각 작성하였다. 이때, 성격이 유사한 단어를 구분하기 위해 증권명의 유사한 정도에 따라 15개의 카테고리로 나누어진다.

표 1. 50개의 유사음소단위.

Table 1. Phoneme likely units of 50 units.

모음	aa, axr, ao, uh, U, ih, ae, eh, ja, jv, jo, ju, wa, wv, wE, we, wi, je, ya, Wi
자음	b, d, g, b~, d~, g~, b+, d+, g+, z~, bb, dd, gg, zz, p, t, k, ch, s, ss, hh, hh~, r, n, m, ng, z, l
묵음	silB, silE

발음사전은 화자들의 여러 가지 발성을 고려하고 한국어 음운규칙을 적용하여 작성하였다. 본 논문에서는 음운규칙을 적용한 발음사전 변환 방법으로서 한국어 처리 시스템(KPS)[11]을 이용하였다. 그리고 어휘수가 큰 경우 유사한 단어들은 동일한 접두사(prefix)를 가지기 때문에 접두사를 공유하는 목구조(tree-structure) 형태의 사전으로 구성함으로서 발생 가능한 상태의 수를 감소시킬 수 있다. 이 구조는 어휘가 증가할수록 현저한 효과가 나타나게 되어 대어휘 연속음성인식에 많이 이용된다[12,13].

인식 방법은 Multi-Pass 탐색[12,13]법으로서 작성한 음향모델과 언어모델, 발음사전을 이용하여 입력음성에 대해 1)1-pass 탐색에서는 순방향 의사 2-gram 언어모델을 이용하여 프레임 동기형 Viterbi Beam Search를 수행한 후 중간결과를 출력하고 2) 2-pass 탐색에서는 1-pass의 중간결과와 역방향 의사 3-gram 언어모델을 이용하여 A* best-first stack decoding 탐색을 수행한 후 인식결과를 출력한다.

4.2 언어모델 및 음향모델

언어모델은 본 논문에서 제안한 의사 N-gram 언어모델을 이용하였다. 인식대상 어휘수는 오프라인의 경우 452단어이며, 온라인의 경우 1,500여 개의 증권명으로서 성격이 유사한 카테고리를 15개로 분류하고 각 카테고리에 따라 의사 N-gram 언어모델을 작성하였다. 본 논문에서 분류한 카테고리를 표 2에 나타내었다.

기본 HMM 음향모델의 작성을 위해 국어공학센터(KLE)에서 작성한 PBW(Phoneme Balanced Words) 452단어 음성 데이터베이스 중 남성 38명의 2회 발성 중 35명의 1회와 1, 2회 발성 단어에 대해 HTK를 이용하여 혼합수 1, 4, 8, 12 및 16을 가지도록 작성하였다. 평가용 음성 데이터는 사전실험을 위한 오프라인의 경우 KLE의 학습에 참가하지 않는 나머지 3명 1회의 452단어를 사용하였고, 온라인의 경우 사무실 환경의 개인용 컴퓨터에서 일반 헤드셋 마이크를 통해 20명의 화자가 발성한 20단어를 사용하였다.

모든 음성 데이터는 16kHz, 16bits로 샘플링된 후 $1-0.97z^{-1}$의 필터를 통과시킨다. 입력 음성의 각 프레임에 25ms의 해밍 윈도우를 곱하여 10ms 마다 분석된다. 인식을 위한 특징 파라미터로서는 12차의 MFCC(Mel Frequency Cepstral Coefficient), 1차 및 2차의 차분 MFCC와 각각의 차분 power를 합한 총 38차의 계수로 구성하였다. 또한 모든 음성 채널을 고려하여 2차의 차분 power에 대해 켑스트럼 평균 정규화(CMN)를 적용하였다. 표 2에 사용된 음성 데이터 및 분석조건을 나타내었다.

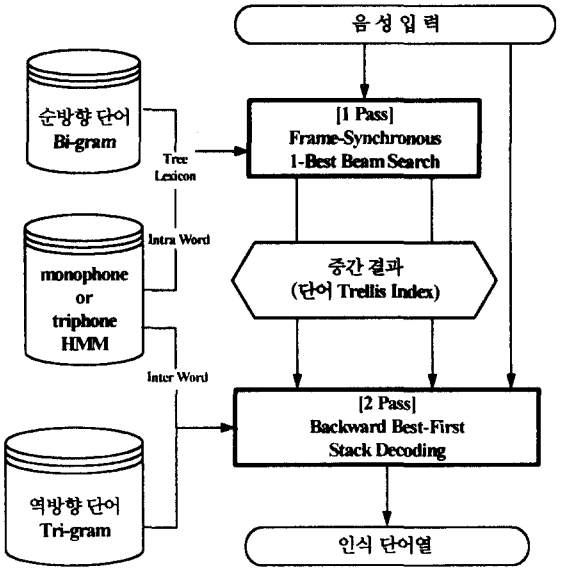


그림 5. 인식 시스템의 전체 구성도.

Fig. 5. Overall diagram of recognition system.

표 3. 증권종목의 분류.
Table 3. Classification of stock name.

순번	카테고리명	순번	카테고리명
1	건설에너지수도	9	자동차운송
2	고무화학제약	10	전자통신컴퓨터
3	금속기계장비	11	종이출판가구
4	도소매숙박서비스	12	코스닥관리투자유의
5	섬유의복	13	코스닥뮤추얼펀드
6	은행증권보험	14	코스닥벤처기업
7	음식농광어업	15	코스닥일반기업
8	일반관리대상		

표 3. 음성 데이터 및 분석조건.
Table 3. Analysis conditions and speech database.

음성 데이터			
발성형태	단어	단어	단어
화자	남성 35	남성 3	남성 20
사용단계	모델학습	인식	인식
단어수	452	452	20
발성횟수	2	1	1
발성환경	방음부스		사무실환경
분석조건			
샘플링 주파수	16khz		
분해능	16bits		
Pre-emphasis	$1-0.97z^{-1}$		
Window	Hamming window(25msec)		
분석 주기	10msec		
특징 파라미터 (38 차)	MFCC(12) + △MFCC(12) + △Power(1) + △△MFCC(12) + △△Power(1)(CMN)		

표 4. 오프라인 단어인식실험의 비교.
Table 4. Comparison of off-line word recognition experiments

문법	FSN		CFG		의사 N-gram	
엔진	HTK One-Pass		Julian Multi-Pass		Julius Multi-Pass	
혼합수	탐색	평균	탐색	평균	탐색	평균
1	1pass	85.4	1pass	82.5	1pass	85.2
			2pass	85.7	2pass	85.2
4	1pass	95.6	1pass	87.8	1pass	95.3
			2pass	95.6	2pass	95.3
8	1pass	97.7	1pass	89.0	1pass	97.6
			2pass	97.8	2pass	97.6
12	1pass	97.4	1pass	90.3	1pass	97.7
			2pass	97.8	2pass	97.7

4.3 단어인식 실험

본 논문에서 제안한 의사 N-gram 언어모델의 유효성을 확인하기 위한 사전실험으로 남성 3인이 발성한 452 단어에 대해 오프라인 평가를 수행하였다. 음향모델은 혼합수 1, 4, 8, 12의 유사음소단위(PLUs)이며, 언어모델은 형식언어인 유한상태 네트워크(FSN)와 문맥자유문법(CFG), 그리고 본 논문에서 제안한 의사 N-gram 언어모델을 사용하였다. 먼저 유한상태 네트워크 언어모델의 경우에는 One-Pass 탐색을 수행하는 HTK 음성인식엔진을 이용하였고, 문맥자유문법 언어모델의 경우에는 Multi-Pass 탐색을 수행하는 Julian 음성인식 엔진[14]을 이용하였으며, 본 논문에서 제안한 의사 N-gram 언어모델의 경우에는 Multi-Pass 탐색을 수행하는 Julius 음성인식 엔진을 이용하였다. 각각의 실험결과에 대한 평가결과를 표 4에 나타내었다.

표 4의 실험결과로부터 알 수 있는 바와 같이 혼합수가 증가함에 따라 인식률의 증가함을 확인할 수 있으며, 유한상태 네트워크(FSN)와 문맥자유문법(CFG) 언어모델의 경우 Multi-Pass 탐색을 수행한 경우의 인식률이 One-Pass 탐색을 수행한 경우보다 조금 높은 것을 알 수 있다. 또한 유한상태 네트워크(FSN)와 문맥자유문법(CFG) 그리고 의사 N-gram 언어모델을 이용한 인식결과에서는 유한상태 네트워크(FSN)의 1-pass 인식률과 문맥자유문법(CFG)의 2-pass 인식률에 비해 조금 낮은 결과를 보이고 있지만, 문맥자유문법(CFG)과 의사 N-gram 언어모델을 비교할 경우 문맥자유문법(CFG)의 1-pass 인식률에 비해 의사 N-gram 언어모델의 1-pass 인식률이 평균 6.5%의 향상된 성능을 보임을 알 수 있다. 만약 Multi-Pass 탐색을 수행하는 인식엔진에 대해 의사 N-gram 언어모델과 전향 1-pass 탐색만을 적용하여 시스템을 구성할 경우 후향 2-pass 탐색에 대한 시간을 줄일 수 있어 시스템에 유효할 것으로 기대된다. 표 4의 오프라인 사전실험 결과를 통하여 본 논문에서 제안한 의사 N-gram 언어모델의 유효성을 확인할 수 있었다.

오프라인 사전실험의 유효성을 바탕으로 온라인 평가실험에 대해 본 연구실에서 구축한 음성인식 증권조회 시스템에 적용하여 실험을 수행하였다. 그림 6에 나타낸 음성인식 증권조회 시스템은 워크스테이션에서 구동하는 Julius Multi-Pass 음성인식 엔진을 개인용 컴퓨터에서 일반 마이크로 입력된 음성을 인식할 수 있도록 비주얼 C++로 포팅한 것이다.

표 3에 나타낸 증권종목 분류에 따라 각 카테고리마다 작성한 의사 N-gram 언어모델과 국어공학센터의 452단어를 남성 38인이 2회 발성한 음성 데이터베이스에 대해 HTK를 이용하여 혼합수 16의 음향모델을 작성하였다. 각 카테고리마다 작성한 의사 N-gram 언어모델은 카테고리 명도 인식대상으로 포함시켰으며, 인식을 수행할 때 대표이름을 발성하면 인식된 대표이름의 언어모델에 포함된 증권명이 인식대상이 되도록 시스템을 구성하였다.

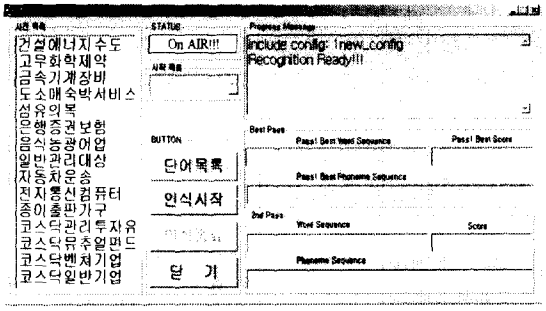


그림 6. 음성인식 증권명 조회 시스템.
Fig. 6. Stock name search system employing speech recognition.

표 5. 온라인 단어인식 실험 결과.
Table 5. Experimental result of on-line word recognition.

화 자	인식대상	평 균
남성 20명	각 카테고리마다 20 단어	92.5

표 5의 각 언어모델의 카테고리에 포함된 어휘수는 40단어에서 350단어로 다양하게 구성되어 있으며, 남성 20명이 각각의 다른 카테고리에 대해 20단어를 한번 발성한 후 인식을 수행한 결과 92.5%의 인식률을 얻었다. 전체적으로 오프라인 평가인 사전실험에서의 인식결과와 비교하여 낮은 인식률을 보이고 있는데, 이는 온라인 평가를 위해 작성한 음향모델의 학습에는 인식대상 어휘의 음성 데이터가 사용되지 않았으며, 마이크로 입력된 음성과 모델 학습에 사용된 음성 데이터의 녹음환경의 차이라고 생각된다. 이 경우에는 인식대상의 음성 데이터를 음향모델 학습에 사용한다면 인식률의 향상을 기대할 수 있을 것으로 생각된다.

이상의 오프라인/온라인 인식실험 결과로부터 제안방법이 유한상태 네트워크(FSN)와 문맥자유문법(CFG) 언어모델을 이용한 경우와 비교하여 인식성능이 비슷하거나 조금 상회하는 결과를 나타내어 제안방법의 유효성을 확인할 수 있었다.

V. 결 론

본 논문에서는 음성인식 시스템의 실용화를 위한 기초적 연구로서 중규모 어휘의 한정된 태스크에서도 적용할 수 있는 언어모델로서 의사(擬似) N-gram 언어모델을 제안하였다. 제안방법은 ARPA 표준형식 N-gram 언어모델의 구조를 가지면서 각 단어의 확률을 임의로 부여하는 비교적 간단한 방법으로 1-gram은 모든 단어의 출현확률을 1로 설정하고, 2-gram에서 허용할 수 있는 단어 시작기호 <s>와 WORD 및 WORD와 단어종료기호 </s>의 접속확률만을 1로 설정하며, 3-gram은 단어시작기호 <s>와 WORD, 단어종료기호 </s> 만의 접속을 허용하며 접속확률을 마찬가지로 1로 설정하는 방법이다.

제안방법의 유효성을 확인하기 위해 국어공학센터(KLE) 452 단어와 1,500 단어규모의 음성인식 증권명 조회 시스템에 대해 남성 3인의 452 단어와 남성 20명의 20단어에 대해 각각 오프라인과 온라인 단어인식실험을 각각 수행하였다. 인식실험결과, 오프라인 평가인 경우 남성 3인의 452 단어는 평균 97.7%의 단어인식률을 얻었으며, 온라인 평가인 경우 남성 20인의 20단어는 평균 92.5%의 단어인식률을 얻어 제안방법의 유효성을 확인할 수 있었다.

접수일자 : 2001. 7. 11 수정완료 : 2001. 7. 18

· 본 논문은 한국과학재단 목적기초연구 (과제번호: 2000-1-30300-003-3) 지원으로 수행되었음.

참 고 문 헌

[1] 오영환, "음성언어 정보처리 연구의 동향," 정보과학회지, 제16권 제2호, pp. 5-11, 1998. 2.

[2] 김순협, "음성인식 기술의 현황과 연구동향," 2000년도 한국음향학회 정기총회 및 학술대회 논문집, 제19권 제2(s)호, 2000.

[3] X. Huang, A. Acero, and H-W. Hon, "Spoken Language Processing: a guide to theory, algorithm, and system development," Prentice Hall, 2001.

[4] 中川聖一, "確率モデルによる音声認識," 日本電子情報通信学会, 1988.

[5] F. Andry, and S. Thornton, "A Parser for Speech Lattices Using a UCG Grammar," Proc. of second European Conference on Speech Communication and Tech., pp. 219-222, 1991.

[6] S. Austin, R. Schwartz, et al., "The Forward Backward Search Algorithm," Proc. of ICASSP'91, pp. 697-700, 1991.

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

[8] P. Clarkson, and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proc. of Eurospeech'97, pp. 2707-2710, 1997.

[9] D. Jurafsky, and J. Martin, "Speech and Language Processing: an introduction to natural language processing, computational linguistic and speech recognition," Prentice Hall, 2000.

[10] S. J. Young et al., "The HTK Book," 1997.

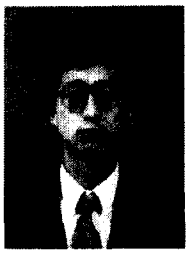
[11] 이상호, 오영환, 서정연, "한국어 문서 음성변환 시스

템을 위한 문서 분석기," 한국음향학회지, 제15권 제3호, pp.50-59, 1996.
[12] A. Lee, T. Kawahara, and S. Doshita, "Large Vocabulary Continuous Speech Recognition based on Multi Pass Search Using Word Trellis Index," In IEICE, Vol. J82-D-II, No. 1, pp. 1-9, 1999.
[13] T. Hori, "A Study on Large Vocabulary Continuous Speech Recognition," Ph. D thesis, Yamagata University, Japan, 1999.
[14] H. Kashima, and T. Kawahara, "Speech Understanding Based on Key Phrase Spotting and Combined Language Models," Technical Report of IEICE, pp. 115-120, 2000. 12.
[15] 오세진, 황철준, 김범국, 정호열, 정현열, "반복학습법에 의해 작성한 N-gram을 이용한 연속음성인식에 관한 연구," 한국음향학회지, 제19권 제6호, pp. 62-70, 2000.
[16] 오세진, 황철준, 김범국, 정호열, 정현열, "의사 N-gram 언어모델을 이용한 단어인식에 관한 연구," 한국음향학회 하계학술발표대회 논문집, 제20권 제1(s)호, pp. 179-182, 2001.

오 세 진(Se-Jin Oh)
正會員
1996년 영남대학교 전자공학과
1998년 영남대학교 전자공학과 (공학석사)
1998년 3월-현재 영남대학교 전자공학과 박사수료
관심분야 : 음성분석 및 인식, 언어처리

황 철 준(Chul-Joon Hwang)
正會員
1996년 영남대학교 전자공학과
1998년 영남대학교 전자공학과 (공학석사)
2000년 영남대학교 전자공학과 박사수료
2000년 3월-현재 대구과학대학 정보전자통신계열 전임강사
관심분야 : 음성분석 및 인식, 디지털 신호처리

김 범 국(Bum-Koog Kim)
正會員
1990년 영남대학교 수학과
1992년 영남대학교 전자공학과 (공학석사)
1998년 영남대학교 전자공학과 (공학박사)
1997년 3월-현재 대구과학대학 정보전자통신계열 조교수
관심분야 : 음성분석 및 인식, 언어처리, 멀티모달 시스템



정 호 열(Ho-Youl Jung)
正會員
1988년 아주대학교 전자공학과
1990년 아주대학교 전자공학과 (공학석사)
1993년 아주대학교 전자공학과 박사수료
1998년 (프)리옹국립응용과학원 전자공학전공(공학박사)
1998년 4월 ~ 1998년 12월 (프)CREATIS Post-Doc
1999년 3월 ~ 현재 영남대학교 전자정보공학부 조교수
관심분야 : 음성 · 영상 신호처리, 인공지능, 디지털 워터마킹 등

정 현 열(Hyun-Yeol Chung)
正會員
1975년 영남대학교 전자공학과
1989년 일본 동북대학교 정보공학과 (공학박사)
1989년 3월-현재 영남대학교 전자정보공학부 교수
1992년 7월-1993년 7월 미국 CMU Robotics 연구소 객원 연구원
1994년 12월-1995년 2월 일본 토요하시기술과학대학 외국인 연구자
2000년 6월-2000년 8월 미국 Qaulcomm Inc. 수석 엔지니어
관심분야 : 음성인식, 화자인식, 음성합성 및 DSP 응용분야