반복학습법에 의해 작성한 N-gram 언어모델을 이용한 연속음성인식에 관한 연구

Continuous Speech Recognition Using N-gram Language Models Constructed by Iterative Learning

오 세 진*, 황 철 준*, 김 범 국**, 정 호 열*, 정 현 열*
(Se-Jin Oh*, Cheol-Jun Hwang*, Bum-Koog Kim**, Ho-Youl Jung*, Hyun-Yeol Chung*)

* 영남대학교 전자정보공학부, ** 대구과학대학 정보전자통신계열
(접수일자: 2000년 5월 8일; 채택일자: 2000년 5월24일)

일반적으로 통계적 언어모델의 확률을 추정하는 방법은 대량의 텍스트 데이터로부터 출현빈도가 높은 단어를 선택하여 사용하고 있다. 하지만 특정 태스크에서 적용할 언어모델의 경우 시간적, 비용적 측면을 고려할 때 대용량의 텍스트의 사용은 비효율적일 것이다. 본 논문에서는 특정 태스크에서 사용하기 위해 소량의 텍스트 데이터로부터 효율적인 언어모델을 작성하는 방법을 제안한다. 즉, 언어모델을 작성할 때 출현빈도가 낮은 단어의 빈도를 개선하기 위해 같은 문장을 반복하여 학습에 참가시키므로 단어의 발생확률을 좀 더 강건하게 하였으며 제안된 언어모델을 이용하여 3명이 발성한 항공편 예약관련 200문장에 대하여 연속음성인식 실험을 수행하였다. 인식실험 결과, 반복학습에 의해 작성한 언어모델을 이용한 경우가 반복학습 적용 전에 비하여 평균 20.4%의 인식률 향상을 보였다. 또한 기존의 문맥자유문법을 이용한 시스템과 비교하여 인식률이 평균 13.4% 향상되어 제안한 방법이 시스템에 유효함을 확인하였다.
핵심용어: 반복학습법, n-gram 언어모델, 문맥자유문법, 단계적 탐색, HTK
투고분야: 음성처리 분야(2.7)

In usual language models(LMs), the probability has been estimated by selecting highly frequent words from a large text size database. However, in case of adopting LMs in a specific task, it is unnecessary to using the general method; constructing it from a large size text, considering the various kinds of cost. In this paper, we propose a construction method of LMs using a small size text database in order to be used in specific tasks. The proposed method is efficient in increasing the low frequent words by applying same sentences iteratively, for it will robust the occurrence probability of words as well. We carried out continuous speech recognition(CSR) experiments on 200 sentences uttered by 3 speakers using LMs by iterative learning(IL) in a air flight reservation task. The results indicated that the performance of CSR, using an IL applied LMs, shows an 20.4% increased recognition accuracy compared to those without it. This system, using the IL method, also shows an average of 13.4% higher recognition accuracy than the previous one, which uses context-free grammar(CFG), implying the effectiveness of it.
Key words: Iterative learning, N-gram language models, Context-free grammar, Multiple-pass search, HTK

I. 서 론

음성인식 시스템을 다양한 분야에서 적용하기 위해서는 무제한의 어휘와 자연스러운 형태로 발음된 음성을 처리할 수 있어야 한다. 하지만 고립단어 인식에 비하여 무제한 연속음성에 있어서는 단어 또는 음소 경계의 불확실성, 상호조음현상 등으로 인하여 인식에 많은 어려움이 따른다. 따라서 연속음성을 인식하기 위해서는 모든 가능한 음소 경계를 고려하고 효율적인 탐색과 정밀한 언어모델을 사용하여 가장 가능성이 높은 단어열을 찾는 것이 필수적인 요소이다. 또한 자연스러운 회화체 음성에서 나타나는 조음 결합을 반영하는 정밀한 음향학적 모델도 요구되고 있다[2,3].

연속음성인식시스템에 주로 사용되고 있는 대표적인 언어모델로서는 문맥자유문법(CFG)과 N-gram 언어모델이 있으며 문맥자유문법을 사용한 시스템의 경우 소량의 학습 데이터로 시스템을 빠른 시간에 효과적으로 구현할 수 있고 어휘 변화에 대한 유연성과 언어이해 부분과 쉽게 통합할 수 있는 장점이 있으나 인식 태스크의 적용범위가 좁고 복잡도(perplexity)가 크며 시스템의 유지 및 호환성 등의 단점으로 인해 음성인식 시스템의 성능을 저하시키는 원인이 되기도 한다[1]. 그러나 통계적 접근에

의한 N-gram 언어모델은 문맥자유문법과는 달리 대량의 학습 데이터가 요구되고 있으나 연속음성 및 회화체 음성 등을 태스크로 인식 시스템을 구성할 경우 효율적인 언어제약으로 탐색 시간을 최소화시킬 수 있어 대용량 연속음성인식 시스템에 널리 사용되고 있다[3,4].

일반적으로 음향학적 사건과 매칭되는 특정 단어열을 효과적으로 구하는 것이 음향학적 모델의 특징이라면 언어모델은 단어열 사이의 연관관계를 확률로서 추론하여 구하는 것이라고 할 수 있다. 즉, 음성인식에서 사용하는 언어모델의 작성 방법은 텍스트 문장 내에서 각 단어 사이의 상관성을 고려하여 출현빈도를 계산한 후 빈도 수가 높은 단어와 이 단어 다음에 출현하는 단어와의 관계를 확률로서 나타내는 것이라고 할 수 있다[5].

그러나 신뢰성 있는 음성인식 시스템을 구성하기 위해서는 다음과 같은 문제점을 고려하여 언어모델을 작성하여야 한다. 첫째로, 모델을 작성하기 위해 사용하는 텍스트 데이터의 단어수에 따라 파라미터의 차수는 단어수의 제곱이 되어 파라미터의 수가 무한히 커질 수 있다는 것을 고려해야 한다. 둘째로, 학습 데이터 중에서 나타나지 않는 단어열에 대해 확률을 0으로 추정하여 학습을 중단하거나 데이터 중에 출현하지만 출현빈도가 적은 단어열에 대해서는 통계적으로 신뢰성 있는 확률을 추정할 수 있어야 한다. 셋째로, 인식한 문장 중에서 '내일은' 으로 시작하는 문장의 끝이 '하였습니다'로 끝나거나, 문장 내에 부정구문이 들어간 문장의 끝이 부정으로 끝나지 않는 것과 같이 태스크에 적합하지 않는 텍스트를 이용하여 언어모델을 작성하는 경우 N-gram이 인접한 단어들 간의 관계를 잘 표현하여 탐색과정에서 용이하게 사용되지만, 인식결과에서 문장 전체가 틀리는 경우를 고려해야 한다. 이러한 문제점을 해결하기 위하여 언어처리 및 인식 분야에서 여러 가지 해결방법을 제시하고 있으나 아직도 많은 연구가 요구되고 있다[8,9,11,13-17].

또한 안정된 언어모델을 작성하기 위해서는 대량의 텍스트 데이터가 필요하며 출현빈도가 고르게 분포되어 있는 텍스트로부터 단어의 출현확률을 추정하는 것이 가장 효율적이라고 할 수 있다.

따라서 대용량의 텍스트를 수집하기 위해서 신문과 방송에서 사용하는 기사나 원고를 많이 이용하고 있으며 이러한 예로서, 국내에서는 인터넷을 통해 쉽게 구할 수 있는 신문기사나 방송기사가 이용되고 있고[16,17], 국외에서는 미국의 경우 Wall Street Journal(WSJ), DARPA에서 매년 수행하고 있는 Broadcast News의 벤치마크 테스트 등과 일본의 每日新聞, 日經新聞 기사 등이 많이 이용되고 있다[11,20].

특히 외국의 경우에 있어서는 수년간의 텍스트를 데이터베이스로 구축하여 언어처리나 다른 연구분야에서 사용하고 있으나 국내에는 아직 이러한 데이터베이스 구축이 미흡한 실정이므로 소량의 텍스트를 이용하여 효율적인 언어모델을 구성하기 위한 연구가 요구되고 있다[13,20].

따라서 본 논문에서는 이상에서 기술한 많은 문제점을 고려하여 연속음성인식 시스템의 인식 성능 향상을 목표로 특정 태스크에 적합한 소량의 텍스트 데이터로부터 강건한 언어모델을 작성할 수 있는 반복학습법을 제안하고, 안정된 언어모델을 작성하기 위하여 기존의 문맥자유문법과 N-gram 언어모델의 경우에 대해서 항공편 예약 관련 200문장을 대상으로 인식실험을 통하여 반복학습에 따른 언어모델의 유효성을 확인하고자 한다.

이를 위하여 제안한 반복학습을 통한 효율적인 언어모델의 작성에 있어서는 출현빈도가 낮은 단어의 빈도를 높이기 위해 같은 문장을 반복하여 학습하고 반복학습에 따른 인식률 변화를 검토한다. 또한 그 결과를 참고로 반복학습에 의해 N-gram 언어모델을 작성하고 그 유효성을 확인하기 위해 인식실험을 통하여 기존의 문맥자유문법을 이용한 경우와 비교 검토하고자 한다

논문의 구성은 다음과 같다. Ⅱ장에서는 기존의 통계적 언어모델 작성 방법과 본 논문에서 제안한 소량의 텍스트 데이터를 이용한 언어모델 작성방법에 대해 설명하고, Ⅲ장에서는 인식실험을 위해 사용되는 단계적 인식방법에 대하여 기술한다. Ⅳ장에서는 전체 시스템의 구성과 인식실험 및 결과에 대해서 서술한 다음, 마지막으로 Ⅴ장에서 본 논문의 결론을 맺는다.

II. 언어모델

2.1. 통계적 N-gram 언어모델

n 개의 단어로부터 생성되는 단어열 $w_1, \cdots, w_n$ 이 주어질 경우 단어열 $w_1, \cdots, w_n$ 의 생성확률은 식 (1)과 같이 정의되어 진다[5].

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

하지만 여러 가지 단어의 조합으로 인해 조건부 확률 $P(w_i|w_1 \cdots w_{i-1})$ 을 구하는 것은 실제로 불가능하게 된다. 만약 V개의 단어가 사용된다고 가정하면 확률 $P(w_i|w_1 \cdots w_{i-1})$ 을 완전히 구하기 위해서는 V^i개의 확률값을 구하여야 한다. 따라서 실질적으로는 모델 구성을 위해서 word history $w_1 \cdots w_{i-1}$ 을 같은 종류의 항으로 분할하고 모델의 파라미터 수를 감소시켜줄 필요가 있다. 이때 어느 시점에서 발생하는 사건의 확률은 바로 이전의 N개 시점까지 발생하는 사건에 영향을 받는 N중 마르코프 과정을 따르게 되는데 단어의 발생을 N-1중 마르코프 과정으로 근사화한 모델을 N-gram 모델이라고 한다[5]. N-gram 모델에서는 어느 시점에서의 단어 발생은 바로 이전의 N-1 단어에 의존하는 것으로 가정하면 다음 식 (2)와 같이 나타낼 수 있다.

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

또한 N-gram 확률은 학습 데이터 중에서 나타나는 단어 N개와 N-1개로부터 다음과 같이 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.

N이 너무 큰 값을 가진다면 학습 데이터로부터 신뢰성 있는 확률을 추정하기가 힘들게 된다. 따라서 음성언어처리에서 많이 사용되고 있는 N의 크기는 2, 3, 4가 대부분으로서 각각 2-gram, 3-gram, 4-gram 이라고 부른다. 또 N이 1인 경우는 단어의 생성확률도 되지만 unigram이라고 부른다[5,8].

예를 들어, 만약 '책을'이란 구절에 대해 '읽다', '사다', '쓰다', '빌리다'라는 어휘가 계속된다고 가정할 때, 이것을 확률로 표현한 것이 N-gram 모델이다. 이러한 모델을 구성하기 위해서는 텍스트 데이터를 이용하는데, 학습 데이터 중에 '책을' 이란 구절이 10회, '책을 사다' 라는 구절이 3회 출현한다고 하면, '책을' 이란 구절 이후에 '사다'가 나올 확률은 0.3이 된다. 따라서 식 (3)을 이용하여 '책을 사다'에 대한 확률을 계산하면 다음 식과 같이 나타낼 수 있다.

$$P(\text{사다}|\text{책을}) = \frac{C(\text{책을 사다})}{C(\text{책을})} \quad (4)$$

2.2. 반복학습에 의한 N-gram 언어모델의 추정

특정 태스크에 사용할 인식 시스템의 언어모델을 작성하는데 많은 량의 텍스트 데이터로부터 언어모델의 확률을 추정하는 경우 시간적, 비용적 측면에서 비효율적일 수 있다. 일반적으로 문장에서 단어 사이의 관계를 나타내는 언어모델의 발생확률 추정은 대용량의 텍스트로부터 출현빈도가 높은 단어를 이용하고 있다. 대용량의 텍스트 수집을 위해서는 주로 인터넷을 통한 신문기사 또는 방송원고를 이용하고 있다. 수집된 대용량의 텍스트로부터 출현빈도를 조사하여 빈도가 높은 단어를 선택하는 것이 가능하다. 그렇지만, 많은 텍스트 데이터에 대해 전처리를 통한 언어모델을 작성하는 작업은 많은 시간과 경비가 소요된다. 외국의 경우 신문기사와 방송원고는 특정 연구기관이나 집단에서 몇 년간의 텍스트를 수집하여 음성인식과 다른 연구분야에서 사용할 수 있도록 하고 있으나 이 또한 많은 경비와 시간이 요구된다.

따라서, 본 논문에서는 이러한 문제점을 해결하기 위해 소량의 텍스트 데이터만을 이용하여 특정 태스크에서 효율적으로 음성인식에 필요한 언어모델을 작성할 수 있는 방법으로서 반복학습법을 제안한다.

제안된 방법은 기존의 sparseness를 다루는 언어모델 작성 방법과 유사하지만 본 논문에서의 기본적인 아이디어는 sparseness를 다루는 언어모델을 작성하기 위한 텍스트 데이터에 비하여 매우 작은 소량의 텍스트로 언어모델을 작성하기 위해서 문장을 인위적으로 단어의 발생확률을 강건하게 높이기 위해 확장하였으며 확장된 텍스트를 반복하여 언어모델의 학습에 이용하는 방법이다. 즉, 인식될 수 있는 가능한 문장으로부터 제한된 태스크에서 임의로 만들 수 있는 문장으로 확장하였으며 이는 제한된 태스크에서의 문장만으로 언어모델을 만들 경우 발생할 수 있는 언어모델의 출현확률을 좀더 보강하기 위한 것이다.

또한 제안방법은 통상 이용되는 출현빈도가 높은 단어를 선택하지 않고 1번 이상 나타나는 모든 단어를 언어모델의 출현확률 추정에 이용하는 것이다. 원래의 텍스트만을 이용하여 언어모델을 작성할 경우에는 단어의 출현빈도가 적어 발생확률 추정이 불가능하여 언어모델을 생성하지 못하는 문제가 있으나, 제안 방법과 함께 언어모델의 smoothing 방법(discounting과 back-off 방법)[4,5,7,8]을 이용하여 동일 텍스트를 수회 반복하여 학습하게 되면 발생확률을 강건하게 하여 언어모델을 생성할 수 있게 된다. 학습문장의 확장에 대한 예는 4.2절에서 설명하기로 한다.

$$\hat{P}_i(w_3|w_1,w_2) = \begin{cases} P_i(w_1,w_2,w_3)\ , \text{for } trigram \\ B(w_1,w_2)\,P_i(w_3|w_2)\ , \text{for } bigram\ w_1,w_2 \\ P_i(w_3|w_2)\ , \text{otherwise} \end{cases} \quad (5)$$

만일 단어 3-gram에 대해서 제안방법을 적용한다면 식 (5)와 같이 표현[14]할 수 있으며, 임의로 만든 모든 문장에 포함된 단어들에 대해 단어 서로 간에 3-gram이 존재한다면 i번째 반복학습한 3-gram 발생확률 $P(w_1,w_2,w_3)$를 단어 $\hat{P}_i(w_3|w_1,w_2)$ w_1,w_2 다음에 발생할 수 있는 단어 w_3의 발생확률 에 대입하고, 추정하기 어려운 3-gram의 확률에 대해서는 2-gram w_1,w_2가 존재한다면 2-gram w_1,w_2의 확률에 Back-off 함수인 $B(w_1,w_2)$를 곱하여 확률값을 추정하게 되고 그 외에는 단어 w_2 다음에 단어 w_3이 발생할 확률로 추정하게 된다.

여기서 $\hat{P}_i(\cdot)$는 i번째 반복 학습된 언어모델 확률 추정값, $B(\cdot)$는 단어 3-gram에 대한 scaled 2-gram의 back-off[4,5] 함수를 나타낸다.

2.3. 복잡도

복잡도는 언어모델에 의해 주어진 제약을 나타내는 것으로 일반적으로 언어모델에 대한 객관적인 평가척도로 이용된다. 복잡도 PP와 엔트로피는 식 (6), 식 (7)로 정의된다[8].

$$PP = 2^{-H_a} \quad (6)$$

$$H_0 = \frac{1}{N}\sum_{i=1}^{N}\log_2\{\hat{P}(w_i|w_1,\cdots,w_{i-1})\} \quad (7)$$

여기서, $w_1,\cdots,w_N$는 문장의 경계 정보를 가지면서 서로 독립인 평가 데이터의 단어들을 나타내고 $\hat{P}(w_i|w_1,\cdots,w_{i-1})$은 단어열 $w_1,\cdots,w_{i-1}$ 다음에 발생하는 단어 w_i를 나타내는 언어모델의 확률을 나타낸다.

2.4. 발음사전 생성

발음사전은 화자들의 여러 가지 발성을 고려하고 한국어 음운규칙을 적용하여 작성하였다. 본 논문에서는 음운규칙을 적용한 발음사전 변환 방법으로서 한국어 처리 시스템(KPS)[18]를 이용하였다. 발음사전에 포함된 목록은 음향모델과 언어모델 모두에 존재하도록 구성하였으며 언어모델을 작성하기 위한 텍스트가 증가함에 따라 발음사전의 목록 수도 증가시켰다. 본 논문에서는 문장의 띄어쓰기를 기준으로 어절을 하나의 단어로 정하고 유사 음소단위(PLUs) 발음사전을 {표기}+{읽기}+{음소기호}로 구성하였으며 그 예를 표 1에 나타내었다.

표 1. 발음사전의 예
Table 1. Example of Pronunciation Lexicon.

KE122	[KE122] k eh ih ih ih l ih ih
KE704	[KE704] k eh ih ih ch ih l g~ ao ng s aa
거쳐서	[거쳐서] g axr ch axr s axr
경유하면	[경유하면] g jv ng ju hh~ aa m jv n
그런으로	[그런으로] g u p jv n u r ao
나고야까지는	[나고야까지는] n aa g~ ao ja gg aa z~ ih n u n
부산행은	[부산행은] b uh s aa n hh~ ae ng u n
비행기는	[비행기는] b ih hh~ ae ng g~ ih n u n
예약하겠습니다	[예약하겠습니다] je ja k aa g~ eh dp ss u m n ih d~ aa
왕복요금을	[왕복요금을] wa ng b~ ao g~ jo g~ u m u l
왕복으로	[왕복으로] wa ng b~ ao g~ u r ao
이코노미클래스	[이코노미클래스] ih k ao n ao m ih k u l r ae s u
편도만으로	[편도만으로] p jv n d~ ao m aa n u r ao

또한 대어휘에서는 많은 단어들이 동일한 접두사(prefix)를 가지는 경향이 있으므로 이 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. 그러나 인식을 위한 탐색과정에서 단어사전 접두사의 동일 부분을 목구조에서와 같이 공유하지 않는 경우 단어의 종단(leaf)에 도달하기까지 서로 다른 것으로 판정될 수 있다. N-gram 확률에서도 이와 같이 될 경우 주어진다면 언어모델 제약으로 활용하는데 문제가 된다. 이를 해결하기 위한 방법으로 언어모델의 확률을 분할하여 목구조 사전의 각 상태에 언어모델의 확률을 할당하는 인수분해(factoring)법[3,9]이 이용된다. 본 논문에서는 접두사를 공유하는 단어에 대한 N-gram 확률의 최대 값을 할당하고, 누적 값을 순서대로 적용하는 방법을 이용하였다. 2-gram 언어모델을 이용하는 경우 바로 이전 단어 v와 상태 s에 대한 2-gram 확률의 할당 값 $\pi(s|v)$는 다음 식과 같이 주어진다[3,9].

$$\pi(s|v) := \max_{w\in W(s)}\{p(w|v)\} \quad (8)$$

여기서, $W(s)$는 목구조 형태의 상태 s를 접두사로 공유하는 단어들을 나타내고, $p(w|v)$는 조건부 2-gram 확률을 나타낸다.

III. 단계적 인식 방법

인식방법으로서 많이 이용되는 디코딩 알고리즘은 관측된 음성에 대해 가장 적절한 단어 후보 열을 찾는 것이다. 대어휘 연속음성인식에서는 고차의 N-gram 언어모델과 triphone과 같이 정밀한 음향모델을 이용하는 것이 바람직하다. 그러나 모델이 너무 정밀하면 인식대상이 증가함에 따라 모델의 참조와 적용이 많아지므로 시스템에 많은 부하를 가져오게 된다. 따라서 1-pass 탐색만을 수행할 경우 전체적으로 처리 효율이 저하되므로 탐색 성능도 좋지 않다. 이를 해결하기 위해 탐색과정을 여러 개의 pass로 분할하여 간단한 모델에서 정밀한 모델까지 순서대로 적용하는 단계적 탐색법을 이용하고 있다[3,4,10,12]. 이 방법에서는 1-pass 탐색단계에서 입력음성 전체에 대해 간단한 모델을 이용하여 인식을 수행하고 단어 그래프(word graph) 형태의 중간 결과를 출력한 다음 2-pass 탐색단계에서는 1-pass 단계의 중간결과와 정밀한 언어모델을 동시에 적용하고 탐색공간을 제한하여 효율적인 재탐색을 수행한다. 2-pass 탐색에서는 1-pass 탐색에서 입력음성 전체에 대해 미리 구해둔 예측 정보를 정밀한 언어모델과 함께 이용하므로 1-pass 탐색의 중간 결과보다 속도도 빠르며 안정된 결과를 얻을 수 있다.

본 논문에서 사용한 multi-pass 탐색 알고리즘은 그 유효성이 입증된 JULIUS[1] 인식 시스템으로서 1-pass 탐색에서는 단어 2-gram을, 2-pass 탐색에서는 단어 3-gram을 이용하였다[11,12,13]. 1-pass에서는 시스템의 고속화를 위해 전향으로 프레임 동기형 빔 탐색 알고리즘을 이용하여 목구조 형태의 사전의 각 상태에 단어 2-gram 확률을 동적으로 분할하여 지정한 후 사전에 있는 모든 단어에 대해서 탐색을 수행한다. 2-pass에서는 단어단위의 best-first의 스택 디코딩 탐색을 수행한다. 언어모델은 단어 3-gram을 이용하고 단어단위의 탐색을 이용하는 것은 단어 레벨에서의 제약조건을 다루기 쉽고 단어단위의 정밀한 언어모델을 적용하는데 용이하기 때문이다. 1-pass에서 출력된 단어 그래프 형태의 중간결과로부터 얻은 정보를 이용하기 위해 2-pass에서는 1-pass와 반대방향인 후향으로 탐색을 수행한다.

1) distributed by Information-technology Promotion Agency, Japan, 1999.

IV. 반복학습법에 의한 언어모델을 이용한 연속음성인식 실험

4.1. 기본 인식 시스템

그림 1에 전체 인식 시스템의 구성을 나타내었다. 인식의 기본단위로는 향후 대어휘 인식으로의 확장성을 고려하여 묵음에 대한 모델을 포함한 51개의 유사음소단위(PLUs)를 이용한다. 유사음소의 로마식 표기를 표 2에 나타내었다.

본 논문에서 이용한 HMM 모델은 5상태 4천이의 7혼합 분포를 가지며 초기상태와 최종상태에는 확률분포가 없는 연속분포 HMM으로 구성하였다.

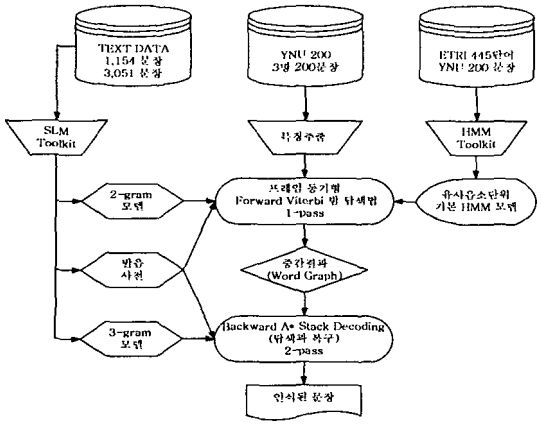

그림 1. 인식 시스템의 전체 구성도
Fig. 1. Overall block diagram of recognition system.

표 2. 51개의 유사음소단위
Table 2. 51 Phoneme-likely-units.

모음	aa, axr, ao, uh, u, ih, ae, eh, ja, jv, jo, ju, wa, wv, wee, we, wi, je, ya, wii
자음	b, d, g, b~, d~, g~, bp, d p, gp, z~, bb, dd, gg, zz, p, t, k, ch, s, ss, hh, hh~, r, n, m, ng, z, l
묵음	silS, silE, sp

4.2. 음성/텍스트 데이터

기본 HMM 모델을 작성하기 위한 음성 데이터는 한국전자통신연구원(ETRI)에서 작성한 PBW 445단어 음성 데이터베이스 중 19명 2회 발성 중 1회분 총 8,455 단어와 영남대학교(YNU)에서 작성한 연속음성 데이터베이스 중 8명의 200문장을 사용하였다. 수작업에 의해 이루어진 ETRI 445 단어의 유사음소단위 레이블 정보와 자동으로 구한 YNU 200 연속음성의 레이블 정보를 이용하여 HTK[21]에 의해 기본 HMM 모델을 작성하였다. 또한 평가용 데이터로는 YNU 200 문장 중 모델 작성에 참여하지 않은 3명의 200문장을 사용하였다.

본 논문에서 사용된 모든 음성 데이터는 16khz, 16bits로 샘플링하고 $1-0.97z^{-1}$의 필터로 pre-emphasis 처리한 후, 입력 음성의 각 프레임에 25msec의 Hamming windows를 곱하여 10msec 마다 분석하였으며, 이를 통해 추출한 12차의 MFCC(Mel-Frequency Cepstral Coefficient) 그리고 1차와 2차의 차분 MFCC와 각각의 차분 power로서 총 38차의 특징 파라미터를 구성하였다. 또한 모든 음성 채널을 고려하여 켑스트럼 평균 정규화(CMN)를 2차의 차분 power에 대해 적용하였다. 표 3에 음성 데이터와 분석조건을 나타내었다.

표 3. 음성 데이터와 분석조건
Table 3. Analysis conditions and speech databases.

음성 데이터			
발성형태	단어	연속음성	연속음성
화자	남성 19	남성 8	남성 3
사용단계	모델학습		인식
단어수/문장수	445	200	200
발성횟수	1	1	1
발성환경	방음부스		
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)		

언어모델을 작성하기 위해 사용한 텍스트는 항공편 예약관련 대화체 텍스트로서 채록환경과 발성형태에 따른 화자의 다양한 부분을 고려하고 또한 언어모델의 단어발생 확률을 보다 강건하게 만들기 위해 인식 태스크 200문장을 인위적으로 1,154문장과 3,051문장으로 각각 확장하여 사용하였는데 확장된 문장에 포함된 단어 수는 각각 6,243단어(평균5.4단어)와 16,271단어(평균 5.3단어)이다. 이렇게 확장한 텍스트 데이터의 예를 표 4에 나타내었다. 학습 텍스트 데이터는 전처리 과정을 통해 문장의 시작(<s>)과 끝(</s>)을 제외한 특수기호는 제외하였다.

언어모델의 작성은 CMU-Cambridge toolkit[14]을 이용하였으며, toolkit에서 제공하는 Good Turing discounting, Witten Bell discounting, Absolute discounting 그리고 Linear discounting의 smoothing 방법 중 Witten-Bell discounting 방법을 적용하였다[11-13]. 여기서 Good Turing discounting 방법은 단지 임의의 특정 회수보다 발생빈도가 작은 단어에 대해서 discounting을 적용하고, Absolute와 Linear discounting 방법은 모든 발생단어에 대해서 discounting을 적용하고 있다. Witten Bell discounting 방법의 경우에 있어서는 전체 발생단어에 대해서 discounting을 적용하지 않으며 discounting 비율(discounting 계수)이 발생단어의 회수에 의존하지 않고 있다. 그러나 특정 문맥을 따르는 단어 수에 의존하여

표 4. 확장한 학습 텍스트 데이터의 예
Table 4. Example of the extended learning text data.

```
<원래문장>
<s> 다시 말해 주세요 </s>
<s> 다시 말해 주시겠습니까 </s>
<s> 다시 말해 주시겠어요 </s>
<s> 다시 말해 주실 수 있나요 </s>
<s> 다시 말해 주십시오 </s>

<확장된 문장>
<s> 다시 불러 주세요 </s>
<s> 다시 불러 주시겠습니까 </s>
<s> 다시 불러 주시겠어요 </s>
<s> 다시 한번 말해 주시겠습니까 </s>
<s> 다시 한번 말해 주시겠어요 </s>
<s> 다시 한번 말해 주실 수 있습니까 </s>
<s> 다시 한번 말해 주십시오 </s>
<s> 다시 한번 불러 주세요 </s>
<s> 다시 한번 불러 주시겠습니까 </s>
<s> 다시 한번 불러 주시겠어요 </s>
<s> 다시 한번 더 말해 주세요 </s>
<s> 다시 한번 더 말해 주시겠습니까 </s>
<s> 다시 한번 더 말해 주시겠어요 </s>
<s> 다시 한번 더 불러 주세요 </s>
```

discounting을 적용하고 있다는 것이 위의 세 가지 방법과 다른 점이라고 할 수 있다[14]. 따라서 소규모 텍스트의 경우, 발생할 수 있는 단어의 빈도수가 적고 특정 문맥을 따르는 단어의 수에 의존하는 언어모델을 작성할 때 보다 유리하여 본 논문에서는 Witten Bell discounting 방법을 이용하였다. 표 5에 언어모델을 작성하기 위해 사용된 텍스트 데이터를 나타내고 있다.

표 5. 언어모델 작성용 텍스트 데이터
Table 5. Text database for language models.

문장종류	항공편 예약 대화체 문장	
문장 수	1,154	3,051
총 단어 수	6,243	16,271
평균 단어 수	5.4 단어	5.3 단어
smoothing 방법	Witten-Bell Discounting	

4.3. 연속음성 인식실험

소량의 텍스트 데이터를 이용하여 특정 태스크에 적합한 언어모델을 효율적으로 작성하고 반복학습법의 유효성을 확인하기 위해 항공편 예약 태스크를 대상으로 연속음성 인식 실험을 수행하고자 한다. 또한 반복학습에 의해 작성한 N-gram 언어모델의 유효성을 확인하기 위해 인식실험을 통하여 기존의 문맥자유문법을 이용한 경우와 비교 검토하고자 한다.

4.3.1. 반복학습에 따른 인식

반복학습으로 작성한 언어모델의 유효성을 검토하기 위하여 각각 1회에서 10회, 10회에서 100회까지 반복 학습한 언어모델의 복잡도와 2-pass에 대해 인식 실험을 수행하였다.

이때 인식실험을 위한 연속음성 데이터는 13명의 사용자가 가상으로 설정한 항공편 예약에 관련된 대화음성을 사용하며 텍스트 데이터는 4.2절에 나타낸 것과 같이 텍스트 200문장으로부터 1,154문장과 3,051문장으로 각각 확장하였다. 또한 이중 평가용 음성 데이터는 3명의 200문장으로 설정하고 나머지 10명의 음성 데이터 중 발성상태가 양호한 8명의 연속음성을 ETRI 445 14명 1회 발성의 단어와 함께 음향모델의 학습에 사용하였다.

반복학습에 따른 인식률의 변화를 표 6과 그림 2에 나타내었으며, 그림 3에 복잡도의 변화를 나타내었다.

표 6. 반복학습 회수에 따른 연속음성 인식률의 변화
Table 6. Change of continuous speech recognition rate according to iterative learning frequency.

반복학습 회수	1,154 문장	3,051 문장
	평균	평균
1	72.2 (68.8)	69.3 (66.5)
2	87.2 (80.5)	85.2 (76.8)
3	89.8 (81.0)	85.3 (76.4)
4	90.2 (81.7)	85.5 (75.2)
5	90.3 (81.2)	85.2 (74.8)
6	91.0 (81.8)	84.8 (74.5)
7	91.3 (82.2)	85.5 (74.3)
8	91.3 (82.0)	86.0 (75.7)
9	91.5 (82.3)	86.7 (76.0)
10	92.2 (82.7)	90.0 (80.2)
20	92.0 (83.8)	89.8 (80.7)
30	91.7 (83.8)	89.5 (80.2)
40	91.8 (85.3)	89.7 (81.2)
50	91.2 (84.2)	89.7 (80.7)
60	91.3 (84.2)	89.7 (80.7)
70	91.3 (84.3)	89.3 (80.7)
80	91.5 (84.2)	89.3 (81.0)
90	91.3 (84.2)	89.3 (80.8)
100	91.3 (84.2)	89.3 (80.8)

연속음성 인식률(%) : 2-pass with 3-gram (1-pass with 2-gram)

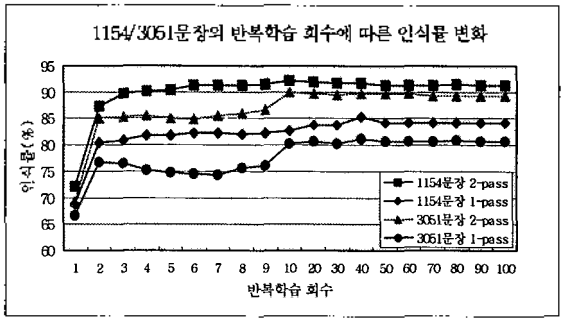

그림 2. 반복학습 회수에 따른 연속음성인식률의 변화
Fig. 2. Change of continuous speech recognition rate according to iterative learning frequency.

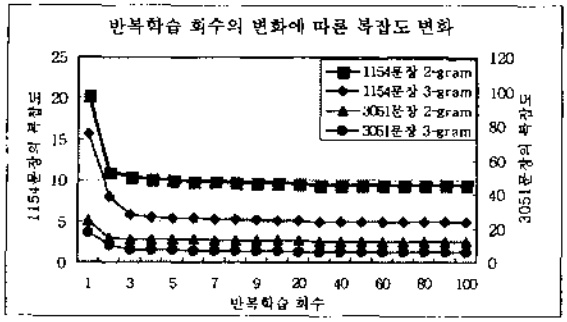

그림 3. 반복학습 회수에 따른 언어모델의 복잡도
Fig. 3. *Perplexity of LMs according to iterative learning frequency.*

표 6과 그림 2에 나타낸 것과 같이 평가용 3명의 200문장을 각 반복 학습 회수에 따른 언어모델에 대해 평가한 결과, 1,154문장의 경우 1-pass 탐색은 최고 평균 85.3%, 2-pass 탐색은 최고 평균 92.2%의 연속음성 인식률을 얻었다. 또한 3,051문장의 경우 1-pass 탐색은 최고 평균 81.2%, 2-pass 탐색은 최고 평균 90.0%의 연속음성 인식률을 얻었다. 그리고 반복학습 적용 전(학습회수 1회)과 후의 언어모델을 이용한 인식실험 결과, 적용 후의 인식률이 평균 20.4% 향상됨을 알 수 있었다.

또한 반복학습 회수가 증가함에 따라 2-pass 탐색의 인식률의 변화는 많지 않지만, 1-pass 탐색의 인식률은 반복학습 회수의 증가에 따라 인식률이 증가하다가 일정하게 유지됨을 알 수 있었으며 그림 3에서와 같이 언어모델의 복잡도의 변화도 반복회수가 증가함에 따라 서서히 줄어드는 것을 볼 수 있다.

전체적으로 볼 때, 1-pass의 경우보다 2-pass의 경우가 높은 인식률을 보였으며 또한 10회 정도의 반복학습 회수가 특정 태스크에 적합한 언어모델의 작성에 적합함을 알 수 있었다. 그리고 텍스트 데이터에 있어서는 3,051문장보다 1,154문장으로 확장한 경우가 인식률이 높게 나타났으나 그 원인으로서는 문장을 구성할 때 태스크의 특성상 시간과 날짜를 표현한 문장을 많이 첨가하여 단어 사이에 혼란이 발생하기 때문으로 생각되며, 향후 숫자를 포함한 문장에 대한 언어모델 작성 방법의 개선과 특정 태스크에 적합한 텍스트 데이터를 이용할 경우 인식률이 향상될 것으로 기대된다.

이상의 결과로부터 소량의 텍스트 데이터를 이용하여 특정 태스크에서 언어모델을 효율적으로 작성하기 위한 방법으로 반복학습에 의한 언어모델과 2-pass 인식방법이 유효성을 확인할 수 있었다.

4.3.2. 문맥자유문법(CFG)과 N-gram 모델의 비교[22]

제안한 반복학습법에 의해 작성한 N-gram 언어모델의 유효성을 검토하기 위해 인식실험을 통하여 기존의 언어모델로서 문맥자유문법을 이용한 경우와 비교 검토하였다.

이때 인식에 있어서는 저자들이 기존에 구축한 문맥자유문법을 적용한 인식 시스템의 특성을 고려하여 1-pass 방법을 사용하였으며, 언어모델인 문맥자유문법은 어휘수 413, 문법규칙 422, 종단기호 126, 비종단기호 197, word class 440으로 구성되어 있다. 평가 데이터에 있어서는 3명의 화자에 의한 100문장으로서 하나의 문장에 포함된 평균 단어수는 5.1개이고 복잡도는 43.2이다.

그림 4에 문맥자유문법을 적용한 기존의 인식 시스템과 반복학습으로 작성한 N-gram 언어모델을 적용한 시스템을 대상으로 인식실험한 결과를 나타내었다.

	문맥자유문법	n-gram 모델
3명 100문장	68.3%	81.7%
사용 언어모델	unigram	2-gram

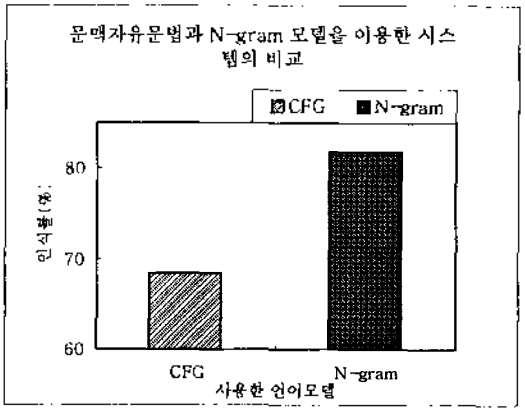

그림 4. 문맥자유문법을 적용한 기존 시스템과 N-gram 언어모델을 적용한 시스템의 비교
Fig. 4. Comparison of previous system adopted by CFG and recognition system adopted by N-gram LMs.

그림 4로부터 기존의 인식 시스템의 인식결과는 화자 적응화된 음향모델에 대해 3명의 100문장에 대해 평균 68.3%의 인식률을 보였으며, 제안된 방법의 경우 평균 81.7%의 인식률을 얻었다. 이상의 결과로부터 반복학습법에 의해 작성한 N-gram 언어모델을 이용한 인식률이 평균 13.4%의 향상된 인식률을 보여 본 연구에서 제안한 반복학습법으로 작성한 N-gram 언어모델이 기존의 문맥자유문법을 이용한 시스템보다 유효함을 확인할 수 있다.

V. 결 론

본 논문에서는 특정 태스크용 연속음성인식 시스템의 성능 향상을 위하여 특정 태스크에 적합한 소량의 텍스트 데이터를 이용하여 효율적으로 언어모델을 작성할 수 있는 반복학습법을 제안하고, 안정된 언어모델을 작성하기 위하여 기존의 문맥자유문법과 N-gram 언어모델의 경우 대해서 항공편 예약 관련 200문장을 대상으로 인식실험을 통하여 반복학습에 따른 언어모델의 유효성을 확인하였다.

제안한 반복학습법으로 작성한 언어모델을 이용한 인식실험 결과, 1,154문장의 경우 1-pass 탐색은 최고 평균 85.3%, 2-pass 탐색은 최고 평균 92.2%의 연속음성 인식률을 얻었다. 또한 3,051문장의 경우 1-pass 탐색은 최고

평균 81.2%, 2-pass 탐색은 최고 평균 90.0%의 연속음성 인식률을 얻었다. 그리고 반복학습 적용 전(학습회수 1회)과 후의 언어모델을 이용한 인식실험 결과, 반복학습을 적용한 후의 인식률이 평균 20.4% 향상됨을 알 수 있었다.

또한 반복학습 회수가 증가함에 따라 인식률이 증가하다가 10회 이후부터는 수렴함을 알 수 있어, 특정 태스크를 대상으로 언어모델을 작성할 경우 10회 정도의 반복학습이 효과적임을 알 수 있었다.

또한 반복학습법으로 작성한 N-gram 언어모델을 이용한 인식실험 결과, 기존의 인식 시스템의 경우 평균 68.3%의 인식률을 보였으며, 제안된 방법의 경우 평균 81.7%의 인식률을 얻었다. 이상의 결과로부터 반복학습법에 의해 작성한 N-gram 언어모델을 이용한 인식률이 평균 13.4%의 향상된 인식률을 보여 본 연구에서 제안한 반복학습법으로 작성한 N-gram 언어모델이 기존의 문맥자유문법을 이용한 시스템보다 유효함을 확인할 수 있다.

감사의 글

이 논문은 1998년도 한국과학재단 핵심전문연구(981-0913-065-2) 연구비에 의해 연구되었음.

참 고 문 헌

1. L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition," Prentice-Hall International, Inc. 1993.
2. S. Young, "A review of Large-Vocabulary Continuous Speech Recognition," IEEE Signal Processing Magazine, Vol. 13, No. 5, pp. 45-57, Sept. 1996.
3. H. Ney, S. Ortmanns, "Dynamic Programming Search for Continuous Speech Recognition," IEEE Signal Processing Magazine, pp. 64-83, Sept. 1999.
4. N. Deshmukh, A. Ganaparhiraju and J. Picone, "Hierarchical Search," IEEE Signal Processing Magazine, pp. 84-107, Sept. 1999.
5. 北 研二 외, "音聲言語處理," 1996.
6. 김득수, 황철준, 정현열, "음성인식 기능을 가진 주소입력 시스템의 개발과 평가," 한국음향학회지, Vol. 18 No. 2, pp. 3-10, 1999. 2.
7. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 35, No. 3, pp. 400-401, 1987.
8. R. Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach," Ph. D. thesis, School of Computer Science, Carnegie Mellon University, 1994.
9. S. Ortmanns, H. Ney and A. Eiden, "Language Model Look Ahead for Large Vocabulary Speech Recognition," In Proc. of ICSLP'96, pp. 2095-2098, 1996.
10. L. Nguyen, R. Schwartz, "Efficient 2-Pass N-best Decoder," In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1997.
11. K. Ohsuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui and K. Shirai, "Japanese Large Vocabulary Continuous Speech Recognition Using a Newspaper Corpus and Broadcast News," In Speech Communication, Vol. 28, pp. 155-166, 1999.
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. Kawahara et. al, "Shareable Software Repository for Japanese Large Vocabulary Continuous Speech Recognition," In Proc. of ICSLP'98, pp. 3257-3260, 1998.
14. P. Clarkson, R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," In Proc. of Eurospeech'97, pp. 2707-2710, 1997.
15. N. Yodo, K. Shikano and S. Nakamura, "Compression Algorithm of Trigram Language Models based on Maximum Likelihood Estimation," In Proc. of ICSLP'98, pp. 1683-1686, 1998.
16. Ha-Jin Yu, Joon-Mo Hong, Hoon Kim and Jong-Seok Lee, "Korean Broadcast Speech Recognizer Using Cross-Word Phone Network," In Proc. of ICSP'99, Vol. 1, pp. 311-314, 1999.
17. Gang-Seong Lee, Alex Waibel, "Korean Broadcast News Speech Recognition Using HMM," In Proc. of ICSP'99, Vol. 1, pp. 275-278, 1999.
18. 이상호, 오영환, 서정연, "한국어 문서 음성 변환 시스템을 위한 문서 분석기," 한국음향학회지, Vol. 15, No. 3, pp. 50-59, 1996.
19. Se-Jin Oh, Cheol-Jun Hwang, Ho-Youl Jung and Hyun-Yeol Chung, "A Study on Statistical Language Models for Large Vocabulary Continuous Speech Recognition System," Proc. of ICSP'99, Vol. 1, 1999.
20. In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1997-1998.
21. S. Young, J. Jansen, J. Odell, D. Ollason and P. Woodland, "The HTK Book," 1997.
22. 오세진, 김범국, 정현열, "연속음성인식 시스템의 성능개선," 한국음향학회 학술발표대회 논문집, Vol. 17, No. 2(s), 1997.

▲ 오 세 진(Se-Jin Oh)

1996년 2월 : 영남대학교 전자공학과 (공학사)

1998년 2월 : 영남대학교 대학원 전자공학과(공학석사)

1998년 3월~현재 : 영남대학교 대학원 전자공학과(박사수료)

※ 주관심분야 : 음성분석 및 인식, 언어처리

▲ 황 철 준(Cheol-Jun Hwang)

1996년 2월 : 영남대학교 전자공학과 (공학사)
1998년 2월 : 영남대학교 대학원 전자공학과(공학석사)
1998년 3월~현재 : 영남대학교 대학원 전자공학과(박사수료)
※ 주관심분야 : 음성분석 및 인식, 디지털 신호처리

▲ 김 범 국(Bum-Koog Kim)

1990년 2월 : 영남대학교 수학과(이학사)
1992년 2월 : 영남대학교 대학원 전자공학과(공학석사)
1998년 2월 : 영남대학교 대학원 전자공학과(공학박사)
1997년 3월~현재 : 대구과학대학 정보전자통신계열 전임강사
※ 주관심분야 : 음성분석 및 인식, 언어처리, 멀티모달 시스템

▲ 정 호 열(Ho-Youl Jung)

1988년 2월 : 아주대학교 전자공학과 (공학사)
1990년 2월 : 아주대학교 전자공학과 (공학석사)
1993년 2월 : 아주대학교 전자공학과 (박사수료)
1998년 : (프)리옹국립응용과학원 (INSA de Lyon) 전자공학전공 (공학박사)
1998년 4월~1998년 12월 (프)CREATIS 박사후 과정
1999년 3월~현재 : 영남대학교 전자정보공학부 전임강사
※ 주관심분야 : 음성·영상 신호처리, 인공지능, 디지털 워터마킹 등

▲ 정 현 열(Hyun-Yeol Chung)
현재 : 영남대학교 전자정보공학부 교수
한국음향학회지 16권 8호(1997) 참조