대용량 한국어 연속음성인식 시스템 개발

On the Development of a Large-Vocabulary Continuous Speech Recognition System for the Korean Language

최 인 정*, 권 오 욱*, 박 종 렬*, 박 용 규*, 김 도 영*, 정 호 영*, 은 종 관*
(In Jeong Choi*, Oh Wook Kwon*, Jong Ryeal Park*, Yong Kyu Park*, Do Yeong Kim*, Ho Young Jeong*, Chong Kwan Un*)

요 약

본 논문에서는 연속분포 HMM을 이용한 대용량 한국어 연속음성인식 시스템에 관하여 기술한다. 인식 시스템의 성능을 개선하기 위하여 음성 모델링 단위의 선정, 단어간 모델링, 탐색 알고리듬, 문법에 관하여 연구하였다. 기본 인식단위로 트라이폰을 사용하며, 학습성을 개선하고 기능어에서의 에러 발생을 줄이기 위하여 일반화된 트라이폰과 function word-dependent phone을 사용한다. 단어 사이에는 묵음 모델과 null transition을 사용하여 선택적으로 묵음을 추가하였다. 언어모델로는 단어 클래스에 근거한 word pair 문법과 bigram 모델이 이용된다. 또한 지식 정보들을 효율적으로 활용할 수 있도록 N개의 후보 문장들을 탐색할 수 있는 알고리듬을 구현하였다. 후처리기에서는 word triple 문법을 사용하여 N개의 최적 문장을 재정렬하여 최종적인 인식 문장을 결정하며, 마지막으로 후치사와 관련된 사소한 에러들을 수정한다. 3천단어의 연속음성 데이타베이스에 대한 인식실험에서, 후처리로 word triple 문법을 사용하여 93.1%의 단어 인식률과 73.8%의 문장 인식률을 얻었다.

ABSTRACT

This paper describes a large-vocabulary continuous-speech recognition system using continuous hidden Markov models for the Korean language. To improve the performance of the system, we study on the selection of speech modeling units, inter-word modeling, search algorithm, and grammars. We used triphones as basic speech modeling units, generalized triphones and function word-dependent phones are used to improve the trainability of speech units and to reduce errors in function words. Silence between words is optionally inserted by using a silence model and a null transition. Word pair grammar and bigram model based on word classes are used. Also we implement a search algorithm to find N-best candidate sentences. A postprocessor reorders the N-best sentences using word triple grammar, selects the most likely sentence as the final recognition result, and finally corrects trivial errors related with postpositions. In recognition tests using a 3,000-word continuous speech database, the system attained 93.1% word recognition accuracy and 73.8% sentence recognition accuracy using word triple grammar in postprocessing.

I. 서 론

음성은 대부분의 사람들에게 정보교환을 위한 가장 자연스럽고 효율적인 수단이다. 인간은 청각보다는 시각을 통해 외부로부터 수동적으로 더 많은 자극을 받지만, 상호 통신에 있어서는 음성을 사용하는 것이 더 효과적이다. 이것은 음성파형 자체가 언어적인 정보뿐만 아니라 화자의 어조와 감정까지 전달하기 때문이다.

음성이 인간과 기계사이의 효율적인 통신 수단이 되기 위해서는 연속음성인식의 기술 개발이 필수적이다. 외국에서는 자동통역 전화를 발표하는 등 음성대화 시스템의 개발 및 실용화에 박차를 가하고 있다. 최근 국내에서도 호텔예약, 증권정보 안내, 생활정보 안내 등의 응용 분야에서 연속어 및 연결어 인식 연구가 활발히 진행되고 있다[1][2][3][4].

본 논문에서는 무역상담을 태스크로 하는 3천단어 규

*한국과학기술원 전기 및 전자공학과 통신연구실
접수일자 : 1995년 5월 15일

모의 대용량 한국어 연속 음성인식 시스템에 대하여 기술한다. 일반적으로 연속음성의 인식은 단어 사이의 조음화 결합과 단어 경계의 불명확성, 그리고 문법적 제약 등의 특성에 의해 고립단어의 인식보다 훨씬 어렵다[5]. 개발된 시스템에서는 이러한 연속음성의 특성을 고려하기 위하여 개선된 음성신호의 모델링, N개의 후보 문장을 찾을 수 있는 탐색 기법, 통계학적 언어 모델을 사용하였다. 조음화 현상을 모델링하기 위하여 트라이폰(triphone)을 인식단위로 사용하였으며, 단어 사이에는 묵음을 선택적으로 추가할 수 있도록 하였다. 지식정보를 효율적으로 활용하기 위하여 N개의 후보 문장을 찾을 수 있는 탐색 알고리듬을 구현하였으며, 단어 클래스에 근거한 확률적인 언어모델을 문법으로 사용하였다.

본 논문의 전체적인 구성은 다음과 같다. 2장에서는 개발된 인식 시스템의 특성과 인식 성능을 향상시키기 위하여 채택된 기법에 관해 살펴본다. 3장에서는 시스템의 성능 평가를 위해 사용된 음성 데이타베이스와 실험 결과에 대해 서술하며, 마지막으로 4장에서 결론을 맺는다.

Ⅱ. 대용량 연속음성인식 시스템

2.1 개 요

연속음성인식 시스템은 크게 음성신호의 특징추출부, 단어단위 정합부, 문장단위 정합부 등 세부분으로 구성되어 있다[6]. 음성신호의 특징추출부에서는 음성신호를 시변적인 특성을 대표할 수 있는 특징벡터의 열로 변환한다. 단어단위 정합부는 입력 특징벡터열과 시스템의 단어 모델사이의 유사도를 측정하여 가장 유사한 단어를 결정한다. 여기서 단어 모델은 발음사전에 의한 subword 모델의 결합으로 얻어질 수 있다. 문장단위 정합부는 언어 모델을 이용하여 문법에 맞는 최대 확률을 내는 단어열을 인식 결과로 결정한다.

음성신호는 표본화, 끝점 검출, preemphasis 과정 등을 포함하는 전처리 과정을 거쳐, 인식에 사용될 수 있도록 특징벡터의 열로 변환된다. 음성은 프레임 단위로 분할되어 처리되는데 각 프레임은 30 msec의 길이를 가지며 10 msec씩 중첩된다. 본 논문에서 사용된 특징벡터는 12차의 캡스트럼 계수와 에너지, 그리고 이의 1차 미분계수들로 구성된다. 한 프레임이 26차의 벡터로 표현되어진다.

음성신호를 모델링하기 위해 사용한 방식은 left-to-right 형태의 연속분포 HMM(hidden Markov model)[7]이다. Subword 단위로 HMM을 구성하였으며, 각 subword 모델은 3개의 state로 구성된다. 그러나 문장의 처음과 끝, 그리고 단어 사이의 묵음 모델은 1개의 state를 갖는 HMM으로 모델링하였다. 출력 확률분포를 추정하기 위해서는 4개의 mixture를 갖는 Gaussian 혼합 밀도함수를 사용하였다.

Subword 단위의 모델 학습을 위하여, 모든 문장들을 subword 단위들의 나열로 나타낸 후 segmental k-means 알고리듬[8]을 사용하여 학습하였다. 이 알고리듬은 단어나 문장 등의 음성을 미리 지정해 준 발음사전과 함께 입력으로 받아 새로 얻어진 파라미터들을 이용하여 자동적으로 음성을 분할하여 학습한다. 이러한 과정을 반복하여 충분히 학습된 HMM 파라미터를 추정한다.

한정된 양의 학습 데이타, 상세한 인식단위의 사용 등으로 인하여 HMM 파라미터들이 부족 추정되어질 가능성이 있다. 이러한 문제를 해결하기 위하여 보간법을 이용하여 연속분포 HMM에서 중요한 파라미터인 공분산과 mixture 가중치를 평활화하였다[9].

인식을 위해 사용된 탐색 알고리듬은 one-pass 탐색 알고리듬[10]이다. 전체 탐색 공간에 대한 고려로 생길 수 있는 계산량의 비효율성을 개선하기 위하여 빔(beam) 탐색 기법을 채택하였다. 빔 탐색에서는 매 프레임에서 모든 후보 경로들을 계산하지 않고 확률이 높은 후보들만을 계산한다. 먼저 활성 state들 중에서 최대 likelihood 값을 찾고, 최대값보다 임계치 이하인 likelihood 값을 가지는 state는 더이상 고려하지 않는다. 또한 인식시 단어의 첨가와 삭제의 균형을 맞추기 위해 word insertion penalty를 사용하였다.

2.2 음성 모델링의 개선

음소는 좌우에 위치하는 음소에 크게 영향을 받으며, 이러한 영향을 고려하기 위하여 세분화한 인식단위가 트라이폰이다. 트라이폰은 조음화 현상을 모델링하는 인식단위로서 큰 장점을 지니지만 일반적으로 그 갯수가 많아 충분히 학습되기 어렵다. 따라서 한정된 학습 데이타를 사용하여 트라이폰과 같은 상세한 인식단위를 학습하려면 학습성의 결여 문제가 발생하게 된다. 학습성 결여 문제를 해결하기 위하여 제안된 방법들로서 문맥 통합(context merging)[5], 보간법, 인식단위 감축 규칙(unit reduction rule) 등이 있다. 본 논문에서는 인식단위 감축 규칙과 함께 문맥 통합 기법을 적용하였다. 문맥 통합은 트라이폰이 모든 문맥을 다르게 취급하는 비효율적인 면을 개선하기 위한 것이다. 실제적으로 많은 음소는 이웃하는 음소에 비슷한 영향을 주므로 유사한 문맥을 가지는 모델을 통합하면 모델의 상세성을 유지하면서 학습성을 향상시킬 수 있다. 이렇게 얻어진 인식단위가 일반화된 트라이폰이다. 문맥간의 유사성을 측정하기 위하여 엔트로피와 정보손실의 개념을 적용하였다. 손실된 정보량을 측정하기 위하여 통합 전후의 HMM의 엔트로피를 사용한다. 전이확률은 무시하며, 출력 확률분포에서의 정보량을 HMM의 엔트로피로 정의한다. 특히 다음과 같이 정의된 differential entropy를 연속 확률 변수의 엔트로피로 간주한다.

$$H(X) = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \qquad (1)$$

두 트라이폰 t_1과 t_2를 통합하였을 경우의 정보손실은 식 (2)와 같이 주어진다.

$$L(t_1, t_2) = N(m)H(m) - N(t_1)H(t_1) - N(t_2)H(t_2) \quad (2)$$

여기서 m은 트라이폰 t_1과 t_2가 통합된 경우의 일반화된 트라이폰을 나타내며, N(t)는 트라이폰 t에 분할된 특징 벡터의 수이다. 같은 음소이면서 문맥이 다른 두 트라이폰을 통합했을 때의 정보 손실을 구하며, 최소의 정보손실을 내는 트라이폰 쌍을 찾아 묶어준다. 원하는 갯수의 인식단위를 얻거나 중단 조건이 만족될 때까지 새로이 구성된 일반화된 트라이폰에 대해 위의 과정을 반복한다.

연속음성을 발음할 경우 단어와 단어를 띄어서 발음하거나 연이어서 발음을 할 수 있다. 그러므로 연속음성 인식에서는 단어 사이의 묵음을 잘 모델링하여야 한다. 본 논문에서는 묵음 모델과 null transition을 사용하여 묵음을 선택적으로 추가할 수 있도록 하였다.

영어권에서 개발된 대부분의 연속음성인식 시스템에서 에러의 많은 부분들이 *a, the, are, for, in* 등과 같은 비교적 짧고 내용을 포함하지 않는 일부 단어 집합으로부터 발생한다. 이러한 단어들은 기능어(function word)라고 불리며, 연속음성에서 강조되지 않거나 단축 내지 생략되며, 특히 이웃하는 문맥에 의해 심한 영향을 받는다[5]. 이러한 단어들에 의한 에러를 줄이기 위하여 제안된 방법이 기능어 모델링 방법이다. 한국어의 경우 *은, 는, 이, 가* 등과 같은 조사나 *및, 행, 개* 등과 같은 단어들이 기능어에 해당된다. 본 논문에서는 이러한 단어들을 function word-dependent phone을 사용하여 모델링하였다. Function word-dependent phone은 word-dependent phone과 유사하나, 단지 기능어에 대해서만 구별된 phone 모델을 사용한다는 점이 다르다. 3,000단어의 어휘로부터 29개의 기능어를 선택하고, function word-dependent phone을 사용하여 해당 단어들의 모델을 수정하였다. 이러한 인식단위들은 그 수가 제한되어 있고 학습데이타에서 자주 발생하므로 충분히 모델링될 수 있고, 또한 인식 성능이 상당히 개선될 것으로 기대된다.

2.3 언어 모델

음성인식 시스템에서 언어모델의 역할은 대상 언어의 실제 확률 분포를 개략화하는 것이다. 이 개략화는 주어진 발음 문장을 인식할 때 탐색될 필요가 있는 가능한 문장의 수를 줄이는 역할을 한다. 음성인식에서 사용되는 언어모델은 크게 형식언어이론에 바탕을 둔 문법과 통계학에 근거한 확률적인 문법으로 나눌 수 있다. 형식언어에서의 문법 중의 하나인 FSN(finite state network)으로 표현되는 regular grammar를 음성인식에 사용할 경우 단어수가 증가함에 따라 FSN에서의 state 수가 급격히 증가하게 되어 인식시간이 오래 걸리는 단점이 있다. 확률적 문법은 이러한 문제를 해결하기 위하여 입력된 문장을 파싱하지 않고 그 문장의 발생 확률만을 계산한다.

본 논문에서는 단어 클래스에 근거한 word pair 문법과 bigram 모델을 사용한다. 이 모델은 먼저 각 단어들을 단어 클래스에 할당한 후 단어 클래스에 근거한 가능한 문맥을 찾아 확률값을 부여한다. 단어 클래스에 의한 언어 모델은 문법적 제약을 위해 필요한 파라미터의 수가 적으므로, 학습 데이타를 효율적으로 활용할 수 있다는 장점이 있다. 단어 클래스는 원래 명사나 형용사, 관형사 등과 같은 형태소적 범주와 회사명, 국가명 등과 같은 특수한 의미상의 범주 등에 의해 분류된다. 한국어에서는 문법의 틀이 불완전하고, 특히 구어적 표현에서는 더욱 심하므로 본 논문에서는 회사명, 선적명 등 무역상담 태스크에서의 특수한 의미상 범주와 날짜, 숫자 등과 같은 일반적 단어 범주 등을 주로 사용하였다. 또한 단어 범주를 세분화하기 위하여 주어진 단어의 앞뒤에 올 수 있는 단어 클래스들을 조사하여, 비슷한 상황에 있는 단어들을 하나의 단어 범주로 묶어준다. 초기에 단어 범주들이 텍스트에 심하게 의존하지 않도록 수작업을 통해 비슷한 성격의 단어들을 클래스로 모아 주었다. 초기의 단어 클래스를 시작으로 하여 주어진 단어 클래스에 인접할 수 있는 단어 클래스의 종류를 조사하고, 다른 단어 클래스의 상황과 비교하여 그 정합 정도를 계산하며, 주어진 단어 클래스의 발생 빈도수를 고려하여 최종적인 정합의 정도를 정량화한다. 두 단어 클래스 w_1, w_2의 정합의 정도 $D(w_1, w_2)$는 식 (3)과 같다.

$$D(w_1, w_2) = \frac{1}{N(w_1) + N(w_2)} \sum_{all\ w} [P(w|w_1)P(w|w_2) + P(w_1|w)P(w_2|w)] \quad (3)$$

여기서 $P(w_1|w_2)$는 단어 클래스 w_2와의 w_1의 co-occurrence 확률이며, N(w)는 단어 클래스 w의 발생빈도수이다. 발생 빈도수가 적은 단어 클래스들을 우선적으로 묶어주기 위해 발생 빈도수의 역수를 가중치로 곱한다. 정합의 정도가 가장 큰 단어 클래스 쌍을 찾아 같은 단어 클래스로 묶어준다. 이러한 과정을 반복하여 원하는 단어 클래스 수나 다른 수렴 조건이 만족될 때까지 계속한다. 표 1은 단어 클래스간 정합 정도를 이용하여 얻어진 결과의 일부를 보여주고 있다.

표 1. 단어 클래스의 예
Table 1. Example of word classes

단어 클래스	단어 리스트
1	백, 이백, 삼백, 사백, 오백, 육백, 칠백, 팔백, 구백
2	달러화, 원화, 엔화, 마르크화, 파운드화
3	기계류, 섬유류, 신발류, 식품류, 완구류, 철강류
4	기입, 기재, 명사
5	어떠세요, 어떤가요, 어떻습니까, 어떻겠습니까
6	관하여, 관해, 관해서, 대하여, 대해, 대해서
...	.....

2.4 탐색 알고리듬

빔 탐색에서 임계값을 크게 하면 탐색의 정확도는 향상되나 계산량이 증가하며, 작게 하면 그 반대가 된다. 또한 프레임이 진행될수록 likelihood 값들 사이의 차이가 커진다. 따라서 정확도와 계산량 감소를 적당히 타협하기 위해서는 임계값을 매 프레임마다 적절하게 변화시키는 것이 필요하다. 매 프레임마다 탐색 공간을 적당한 범위에서 유지하기 위하여 이전 프레임에서의 임계값과 활성 state의 수를 고려하여 일정 범위 이내를 유지하도록 임계값을 변화시켰다. 임계값을 결정하기 위하여 사용된 방법은 다음과 같다.

if $s(t) > s(upper)$ and $bs(t-1) > bs(lower)$,

$$bs(t) = bs(t-1) * (1-\alpha_d \frac{s(t)-s(upper)}{s(upper)})$$

else if $s(t) > s(lower)$ and $bs(t-1) < bs(upper)$,

$$bs(t) = bs(t-1) * (1-\alpha_u \frac{s(t)-s(lower)}{s(lower)})$$

else

$$bs(t) = bs(t-1)$$

end if

여기서 $bs(t)$ = 프레임 t에서의 임계값
$s(t)$ = 프레임 t에서의 활성 state의 수
$s(lower)$, $s(upper)$ = 활성 state 수의 하한치와 상한치
$bs(lower)$, $bs(upper)$ = 임계값의 하한치와 상한치
α_d, α_u = 임계값 증가폭을 결정하기 위한 하강 및 상승 비율

많은 형태의 지식정보에 대한 효율적 활용, discriminative training 등을 위해서는 N개의 최적 문장을 찾을 수 있는 탐색 알고리듬의 구현이 중요하다[11]. 본 논문에서는 음성인식 단계를 여러 단계의 모듈로 나누어 지식정보를 적절히 사용할 수 있게 하였다. 대용량 연속어 인식에서 메모리의 부담을 줄이고, N개의 최적 문장을 찾기 위하여 단어내에서는 각 state에서 하나의 경로만 저장하고 문법 node에서는 N개의 가능한 경로를 저장하는 방식을 채택하였다. 저장한 경로수의 제한에 의해 최대 likelihood의 문장은 최적이나 나머지 (N−1)개의 가능한 문장은 준최적이라는 단점이 있다. 전체적인 인식 알고리듬은 7개의 주요 모듈로 구성되어 있으며, 인식 절차는 그림 1에서 보여주고 있다.

초기화 모듈에서는 모든 state와 문법 node에 대한 초기화를 수행하며, likelihood 계산 모듈에서는 단어모델의 모든 state에서 특징벡터를 관측할 확률을 구한다. 단어내 decoding 모듈(*Intra-DP*)에서는 Viterbi decoding에 의해 최적 경로를 찾는다. 문법 레벨 decoding 모듈(*Grammar-DP*)에서는 누적 likelihood와 문법적 제약에 근거하여 N개의 가능한 경로를 찾는다. 갱신 모듈(*Prune/Update*)에서는 빔 탐색에 의해 가능성이 희박한 경로들을 삭제한 후, 누적 likelihood와 경로 정보를 갱신한다. 마지막으로 역추적과 후처리 모듈에서는 N개의 후보 문장을 선택하고 추가된 지식정보를 활용하여 최종 인식 문장을 결정한다.

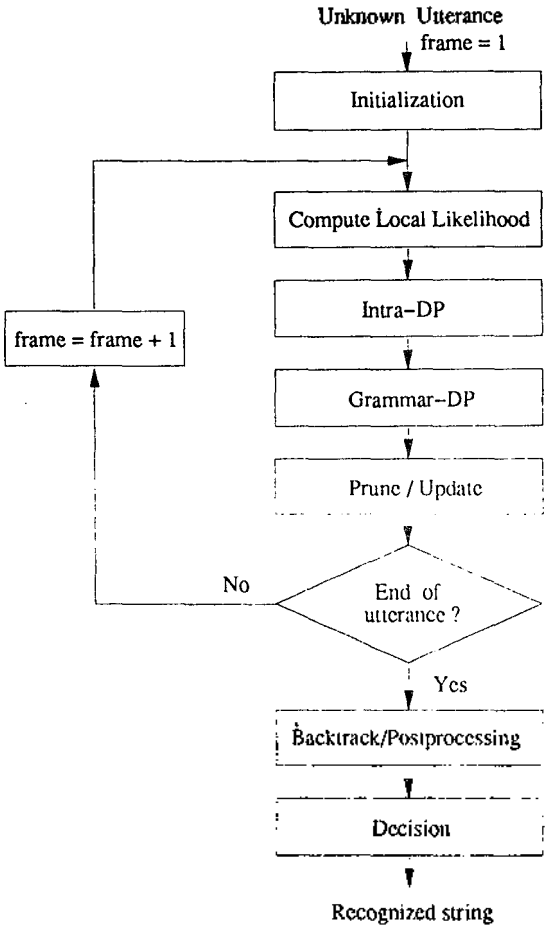

그림 1. N-best 탐색 알고리듬의 블럭도.
Fig 1. A block diagram of the N-best search algorithm.

그림 2는 N개의 후보 문장을 탐색하는 알고리듬을 효율적으로 활용하는 예를 보여주고 있다. 여러 지식정보를 영향력과 복잡성에 따라 정렬하며, 더 적은 계산량으로 많은 제한을 가할 수 있는 지식정보(*KS-1*)들은 N개의 최적 문장 탐색에서 먼저 사용된다. 선정된 N개의 후보 문장은 나머지 지식정보(*KS-2*)에 의해 재정렬되며, 최종적인 인식 문장을 결정한다. 본 논문에서 사용된 지식정보는 표 2와 같다.

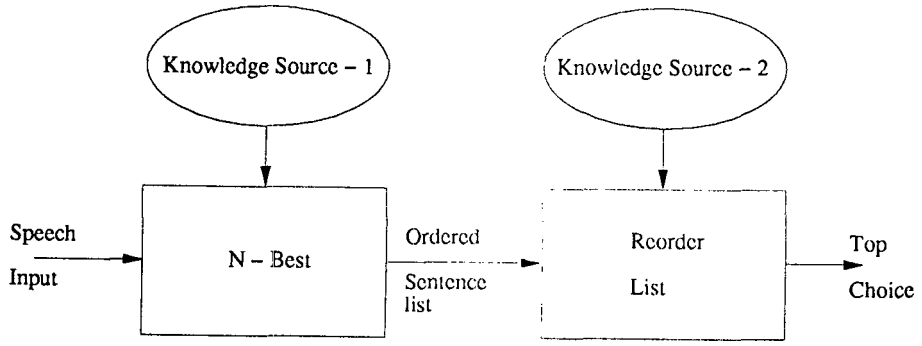

그림 2. N-best 탐색의 활용례.
Fig 2. N-best search paradigm.

표 2. 사용된 지식정보
Table 2. Knowledge Sources

지식 정보	내 용
KS-1	bigram
KS-2	word triple 문법 후처리 기법 (종성모음 + '를', '는') (종성자음 + '을', '은') (종성모음, 공성 'ㄹ' + '로', "로는") (종성자음 + "으로", "으로는")

Ⅲ. 인식 실험 및 결과

3.1 데이타베이스

음성인식 시스템의 학습과 성능 평가를 위하여 사용된 음성 데이타베이스는 한국과학기술원 통신연구실에서 제작한 무역상담용 연속음성 데이타베이스[12]이다. 문장은 2,920개의 어휘로 구성되어 있으며, 남성 100명, 여성 50명이 평균 98개의 문장씩 자연스럽게 발음하였다. 녹음은 조용한 사무실 환경에서 이루어졌으며, 16 kHz, 16 bit로 표본화되었다. 한 문장당 평균 단어수는 8.4개이며, 발음속도는 평균 일분당 166.5개의 단어이었다. 학습에는 남성화자 75명이 발음한 7,352개의 문장을 선정하여 사용하였으며, 성능 평가에는 학습에 참가하지 않은 남성 8명이 발음한 799개의 문장이 사용되었다.

3.2 실험 결과 및 고찰

연속어의 인식률은 단어 인식률과 문장 인식률로 나타내며, 단어 에러에는 치환, 첨가, 삭제의 에러가 포함된다. 인식 실험에서 정확도와 계산량 감소를 적당히 타협하기 위하여 이전 프레임에서의 빔 크기와 활성 노드의 수에 따라 빔 크기를 변화시키는 방법을 사용하였다. 또한 인식 실험을 통해 word insertion penalty 값이 30일 때 첨가와 삭제 에러가 균형을 이루었다. 단어간의 전이마다 누적된 likelihood 값에 word penalty를 추가하는 방법을 택하였다.

표 3은 인식 단위의 상세성과 단어 사이의 묵음 모델 사용이 인식 성능에 미치는 영향을 잘 보여주고 있다. 사용된 문법은 단어 클래스 pair 문법으로서 언어 복잡도(perplexity)가 약 51이다. 기본 음소의 갯수는 32개이며, 트라이폰은 1,895개이다. 트라이폰의 경우 학습성 결여 문제를 극복하기 위하여 인식단위 감축 규칙을 적용한 것이며, 임계값으로 30을 사용하였다. 먼저 단어 사이의 묵음을 고려하지 않은 경우 음소와 트라이폰을 인식단위로 사용하여 각각 65.3, 89.5%의 단어 인식률을 얻었다. 그러나 단어 사이를 묵음 모델과 null transition을 이용하여 모델링한 경우 단어 인식률은 각각 69.4, 90.7%로 향상되었다. 이 결과에서 보듯이 연속어 인식에서는 단어 사이의 모델링이 인식 성능에 큰 영향을 미친다. 기능어에서의 에러 발생을 줄이기 위하여 29개의 기능어에 대한 모델을 별도로 만들고 학습 및 인식 실험을 수행하였다. 기능어 모델링을 하였을 경우 likelihood는 증가하였으나, 단어 인식률은 90.9%로서 예상만큼 증가하지는 않았다.

표 3. Subword 단위와 단어 사이의 모델링에 대한 실험 결과
Table 3. Test Results for Subword Unit and Inter-Word Modeling

인식 단위	단어 인식률	문장 인식률
음소	65.3	24.8
음소 + 선택적 묵음	69.4	28.8
트라이폰	89.5	58.0
트라이폰 + 선택적 묵음	90.7	61.2
+ 기능어 모델링	90.9	61.2

모델의 상세성을 유지하면서 학습성을 향상시키기 위하여 문맥 통합 기법을 적용하였다. 이 방법에 의해 얻어진 일반화된 트라이폰을 인식단위로 하여 학습 및 인식 실험을 하였다. 실험 결과는 표 4에 나타나며, 모델의 상세성과 학습성의 균형점을 찾기 위하여 인식단위의 갯수를 변화시키면서 실험을 수행하였다. 문맥 통합 과정은 인식단위 감축 규칙을 적용하여 얻어진 1,895개의 트라

이폰으로부터 출발하여, 원하는 인식단위의 갯수가 얻어질 때까지 그 과정을 반복하였다. 인식 결과를 살펴보면, 트라이폰을 인식단위로 사용한 경우보다 오히려 인식 성능이 떨어짐을 볼 수 있다. 그 원인은 학습 데이타가 부족하지 않고, 이미 인식단위 감축 규칙을 적용하여 학습성이 충분한 상태에서 인식단위의 수를 줄임으로서 모델의 상세성만 떨어졌기 때문일 것으로 짐작된다. 그러나 대용량 인식 시스템에서는 인식 성능이 크게 떨어지지 않는다면 인식 단위의 갯수를 줄임으로서 인식시간을 크게 단축시킨다는 이점이 생긴다.

표 4. 일반화된 트라이폰을 인식단위로 사용한 인식 결과
Table 4. Test Results Using Generalized Triphones as Speech Unit

일반화된 트라이폰의 갯수	단어 인식률	문장 인식률
800	85.9	50.8
1000	89.5	57.5
1200	90.6	60.3
1895	90.7	61.2

본 논문에서는 단어 클래스에 근거한 문법적 제약을 사용하였다. 먼저 단어의 형태소적 범주와 의미상 범주를 이용하여 단어들을 분류한 후, 텍스트에서의 문맥성과 발생 빈도수를 고려하여 세분화 된 단어 클래스를 형성하였다. 표 5는 사용된 단어 클래스의 수와 단어 클래스에 근거한 word pair 문법을 적용한 경우의 언어 복잡도, 인식 결과를 보여주고 있다. 이 실험에서는 트라이폰을 인식단위로 사용하였으며, 단어 사이의 묵음 모델은 고려하지 않았다. 이 실험 이외의 다른 실험에서는 1,000개의 단어 클래스에 근거하여 문법을 작성하고 인식 실험을 하였다. Word pair 문법 이외에 bigram 모델과 word triple 문법을 사용하였을 때의 언어 복잡도와 인식 결과는 표 6에 나타나 있다. 이 실험에서는 1,000개의 일반화된 트라이폰을 인식단위로 사용하였으며, 단어 사이는 묵음과 null transition으로 모델링되었다. 결과로 부터 문법적 제약을 심하게 가할수록 가능한 문장 패턴의 수가 감소하는 반면 인식 성능은 개선됨을 볼 수 있다.

표 5. 단어 클래스 수가 언어 복잡도와 인식 결과에 미치는 영향
Table 5. Effect of the Number of Word Classes on Perplexity and Recognition Results

단어 클래스의 수	학습집합의 언어 복잡도	단어 인식률	문장 인식률
500	644.0	75.7	26.4
800	143.5	85.7	48.4
1000	51.4	89.5	58.0
1200	28.2	91.4	63.1
1500	22.7	91.9	64.8

표 6. 세 종류의 문법에 대한 언어 복잡도와 인식 결과
Table 6. Perplexity and Recognition Results for Three Kinds of Grammars

문법	언어 복잡도	단어 인식률	문장 인식률
word pair	51.4	89.5	57.5
bigram	29.6	90.7	60.8
word triple	19.4	94.6	79.0

지식정보를 효율적으로 활용하기 위하여 N개의 최적 문장을 찾을 수 있는 탐색 알고리듬을 구현하였다. 이 탐색 알고리듬을 사용하였을 때의 인식 결과는 표 7에 보여지며, 1,000개의 일반화된 트라이폰을 인식단위로 사용하였다. 3개의 후보 문장을 고려한 인식 결과는 word pair 문법을 적용했을 때 단어 인식률이 93.7%이었으며, bigram 모델을 사용했을 때의 단어 인식률은 94.6%이었다. N개의 최적 문장을 찾을 수 있는 탐색 알고리듬은 그림 2에서와 같이 활용될 수 있다. 이 실험에서는 첫번째 지식정보(*KS-1*)로 bigram을 이용하여 3개의 후보 문장을 탐색하였으며, 후처리와 word triple 문법을 두번째 지식정보(*KS-2*)로 활용하여 최종적인 인식 문장을 결정하였다. 여기서 사용된 후처리 기법은 '을', '를'이나 '은', '는'과 같이 의미상에 어떤 변화를 주지 않으면서 에러가 자주 발생하는 8개의 단어에 대하여 음운 규칙에 따라 교정해 주는 것이다. 후처리와 word triple 문법을 두번째 지식정보로 활용하여 93.1%의 단어 인식률과 73.8%의 문장 인식률을 얻었다.

표 7. N-best 탐색에서의 인식률
Table 7. Recognition Rate in N-best Search

rank	word pair		bigram 모델	
	단어 인식률	문장 인식률	단어 인식률	문장 인식률
1	89.5	57.5	90.7	60.8
2	92.6	72.1	93.6	74.5
3	93.7	75.3	94.6	78.2

표 8. 추가된 지식정보를 활용한 실험 결과
Table 8. Test Results Utilizing Additive Knowledge Source

KS-2 (*KS-1: bigram*)	단어 인식률	문장 인식률
None	90.7	60.8
후처리	91.2	64.2
후처리 + word triple	93.1	73.8

Ⅳ. 결 론

본 논문에서는 무역상담을 태스크로 한 3천단어 규모의 대용량 한국어 연속 음성인식 시스템의 특성과 실험결과를 기술하였다. 개발된 인식 시스템은 조음화 현상을 모델링하기 위하여 트라이폰을 인식단위로 하였으며, 단어 사이에는 묵음 모델과 null transition을 사용하여 묵음을 선택적으로 추가할 수 있도록 하였다. 학습성을 개선하기 위해서 보간법에 의한 HMM 파라미터의 평활화, 인식단위 감축 규칙, 문맥 통합 기법 등을 사용하였

다. 또한 기능어에 의한 에러 발생을 줄이기 위하여 function word-dependent phone을 사용하였다. 지식정보를 효율적으로 활용하는 동시에 메모리에 대한 부담을 가능한 줄일 수 있도록 N개의 후보 문장을 탐색할 수 있는 알고리듬을 구현하였다. 마지막으로 언어모델로는 단어 클래스에 근거한 word pair 문법과 bigram 모델을 사용하였다.

2,920 단어로 구성된 무역상담에 관한 연속음성 데이타베이스에서 남성화자 8명이 발음한 799개의 문장에 대해 화자독립 인식실험을 하였다. Bigram 언어모델을 적용하여 3개의 후보 문장을 선정하고, 후처리와 word triple 문법을 추가된 지식정보로 활용하여 93.1%의 단어 인식률과 73.8%의 문장 인식률을 얻었다. 앞으로는 인식 성능과 시간을 개선하기 위하여 학습 방법, 인식 단위, 그리고 고속탐색 알고리듬 등에 관하여 계속 연구해 나갈 것이다.

참 고 문 헌

1. 은 종관, *연속음성 인식시스템 개발 연구 최종보고서*, 한국과학기술원, 1994.
2. 김 희린, 황 규웅, 권 남용, 이 용주, "자동통역을 위한 호텔 예약 Task의 연속음성인식 검토," *제10회 음성통신 및 신호처리 워크샵 논문집*, pp. 256-261, 1993.
3. 구 명완, "N개의 최적문장을 찾을 수 있는 한국어 연속음성 인식 시스템," *제11회 음성통신 및 신호처리 워크샵 논문집*, pp. 48-51, 1994.
4. 김 도영, 박 용규, 권 오욱, 은 종관, 박 성현, "연속분포 HMM을 이용한 한국어 연속음성 인식 시스템 개발," *한국음향학회지*, Vol. 13, No. 1, pp. 24-31, 1994.
5. K. F. Lee, *Automatic Speech Recognition*, Kluwer Academic, 1989.
6. C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, Vol. 6, No. 2, pp. 103-107, 1989.
7. L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4-16, Jan. 1986.
8. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Segmental K-Means Training Procedure for Speech Recognition," *IEEE Trans. on ASSP*, Vol. 38, No. 12, pp. 2033-2045, Dec. 1990.
9. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," *IEEE Trans. on ASSP*, Vol. 1, No. 3, pp. 345-361, July 1993.
10. H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Vol. 32, No. 2, pp. 263-271, Apr. 1989.
11. R. Schwartz and Y. L. Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," *Proc. of ICASSP*, pp. 81-84, 1990.
12. 최 인정, 권 오욱, 박 종렬, 김 도영, 정 호영, 은 종관, "자동통역용 한국어 음성 데이타베이스," *제11회 음성통신 및 신호처리 워크샵 논문집*, pp. 287-290, 1994.

▲최 인 정 : 제 13권 5호 참조
▲권 오 욱 : 제 13권 1호 참조
▲박 용 규 : 제 13권 1호 참조
▲김 도 영 : 제 13권 1호 참조
▲은 종 관 : 제 10권 3호 참조

▲박 종 렬 1963년 3월 14일생
1985년 2월 : 한양대학교 전자공학과 졸업(공학사)
1987년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
1987년 3월~현재 : 한국통신 품질보증단 근무
1993년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정

▲정 호 영(Ho Young Jeong) 1970년 3월 8일생
1993년 2월 : 경북대학교 전자공학과 졸업(공학사)
1995년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
1995년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정