N개의 최적문장을 찾을 수 있는 한국어 연속음성인식 시스템

○
구 명완
한국통신 소프트웨어연구소 자동통역연구팀

A Korean Continuous Speech Recognition System for finding N-best Sentence Hypotheses

Myoung-Wan Koo
Automatic Interpreting Telephony Team, S/W Research Laboratories, Korea Telecom

요 약

본 논문은 N개의 최적문장을 찾을 수 있는 한국어 연속음성인식시스템 개발과 성능실험에 관한 것이다. 개발된 연속음성인식 시스템은 462개의 단어로 이루어지며 언어 복잡도가 30인 연속문장을 인식할 수 있으며 언어처리, 기계번역 등과 같이 쉽게 정합되어 문장 인식률을 향상시킬 수 있도록 N개의 최적문장도 찾을 수 있다. 또한 인식의 기본단위로 triphone을 사용하였으며 단어간, 단어내의 조음현상도 모델링하였다. 남성화자 3인에 대한 화자독립 실험 결과 단어 인식율은 95.7%을 얻었으며 문장 인식율은 한개의 최적문장인 경우 88.5%, 5개의 최적문장을 고려하면 98.6% 문장 인식률을 얻었다.

1. 서 론

연속음성은 사람이 말을 할 때 가장 자연스러운 형태이므로 연속음성을 인식하는 것은 궁극적으로 음성인식의 최종 목표중의 하나이다. 고립단어에 비해 연속음성의 가장 큰 특징은 문법이 있다는 것이다. 실제 언어는 단어가 문법적으로 배열되어야만 올바른 의미전달이 되므로 연속음성에서 문법은 매우 중요하다고 할 수 있다. 연속음성을 인식할 때 문법은 인식률을 향상시키는 도구로 이용되고 있다. 그러나 대화체인 경우는 문법적으로 틀린 문장이 많으므로 인식하기에 더욱 어렵다[1].

또다른 연속음성의 특징은 조음화 현상이다. 고립단어인 경우 조음화 현상은 단어내(intra-word)에만 존재한다. 그러나 연속음성은 단어내 뿐만 아니라 단어 간(inter-word)에도 존재한다. 또한 단어사이의 묵음도 사람에 따라서 생략하는 경우가 있어 묵음을 모델링하기가 쉽지 않다[2]. 마지막으로 연속음성은 고립단어에 비해 발성 지속시간이 길어지기 때문에 인식시간이 많이 걸린다. 이를 위해 고속탐색 알고리즘이 개발되어 인식시간을 감소시키고 있다[3].

본 논문에서는 위에서 기술한 연속음성의 특징을 고려하여 N개의 최적문장을 찾을 수 있는 한국어 연속음성인식시스템의 개발과 성능평가에 대해 기술한다. 먼저 연속음성인식시스템의 개요를 전반적으로 설명하고 특징추출, 음소모델에 대해 기술한다. 그리고 조음화 현상을 모델하는 방안에 대해 검토하고 탐색 알고리즘과 언어처리 알고리즘에 대해 알아본다. 또한 연속음성 인식의 성능평가를 위해 음성 데이타베이스 구성 및 실험결과에 대해 서술한다. 마지막으로 결론을 맺는다.

2. 연속음성인식시스템

2.1 개요

일반적인 연속음성인식 시스템의 개요도가 그림 1.에 그려져 있다. 음성이 입력되면 음성의 특징이 추출이 되고 추출된 특징을 이용하여 단어단위의 비교가 이루어 진다. 음성인식 기본단위로 서브워드를 사용하는 경우는 서브워드의 데이타베이스와 발음사전을 이용하여 입력음성의 특징과 비교되며 결과는 후보단어의 열이 된다. 문장인식은 구문 및 의미를 표현할 수 있는 언어모델을 이용하여 수행된다. 결과는 문장이 되며 필요시는 N개의 최적문장을 얻을 수 있다.

2.2 왜 N개의 최적문장 ?

연속음성을 인식하기 위해서는 많은 형태의 지식정보(knowledge source)가 필요하다. 예를 들면 음성특징, 발음사전, 문법 및 의미 등의 정보는 연속음성을 인식하는데 중요한 역할을 한다. 연속음성 인식시스템은 이러한 정보를 적절히 활용해야 한다. 현재 여러종류의 지식 정보를 처리해 주는 방식으로 종합처리(integrated approach)와 모듈처리(modular approach) 방식이 있다.

종합처리 방식은 모든 지식정보를 한꺼번에 이용해서 음성을 인식하는 방식이다. 즉 음성특징, 발음사전, 구문 및

의미 정보가 하나의 finite state로 표현된다. 현재 이러한 방식은 소용량 단어 및 단순한 문법으로 구성된 연속음성인식시스템에 사용되며 인식시간이 적게 걸리며 시스템이 단순하다는 장점이 있다. 그러나 모든 지식정보가 한번에 종합화 될 수 없는 경우가 발생될 경우 이러한 방식을 이용하기가 어렵다. 예를 들면 prosody 혹은 trigram과 같은 정보는 쉽게 finite state 형태로 변형시키기 어렵다. 또한 단어의 갯수가 증가되면 그에 따라 finite state가 복잡해지므로 실제 대용량 음성인식 시스템에는 적합하지 않다.

반면 모듈처리 방식은 음성인식 단계를 여러모듈로 나누고 매 모듈에서는 모듈정보를 활용하여 최종 결과를 찾는 방식이다. 예를 들면 음성특징을 이용하여 단어를 인식하고 그 다음에 구문정보를 이용하여 문장을 1차로 인식하며 최종적으로 의미 정보를 활용하여 인식된 문장을 결정한다. 그러므로 언어처리, 음성처리 등의 알고리즘이 독립적으로 개발될 수 있으며 단어가 많아지거나 문법이 복잡하여도 모듈단위로 계산이 이루어지므로 대용량 연속음성 인식시스템에 적합하다. 모듈처리 방식은 매 모듈마다 새로운 정보를 이용하여 이전 모듈 결과 중 잘못된 것을 제거하므로 매 모듈의 출력은 여러개가 나와야 한다. 대표적인 알고리즘으로 tree-trellis[4]와 forward-backward 탐색[5] 알고리즘이 있다. 이 알고리즘은 인식결과를 후 처리 할 수 있도록 N개의 최적문장으로 나타낸다. 모듈처리 방식을 채택한 인식시스템은 일반적으로 N개의 최적문장을 나타내 주는 인식시스템이다.

2.3 특징 추출

음성신호는 8㎑, μ-law 8bit로 sampling되고 $1-0.95^{-1}$ 전달함수를 갖는 필터를 사용하여 pre-emphasize된다. 이 음성은 frame 단위로 분할되어 처리되는데 각 frame은 20msec의 길이를 가지며 10msec 중첩된다. 매 frame은 LPC 분석이 수행되고 이 LPC 계수를 이용하여 cepstral 계수가 구해진다. t 프레임에서 구한 LPC 계수 C_t는 다음과 같이 weighting window W_c에 의해 weighting된다.

$$W_c(m) = 1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right) \qquad 1 \le m \le Q$$

여기서 Q는 LPC 차수이며, Weighted LPC cepstral 계수 외에도 그들의 delta(difference), 이차delta(second order difference), 로그파우워의 일차, 이차delta 값이 사용된다. 이 계수들은 다시 백터 양자화되어 아래와 같이 3개의 코우드 북에 저장된다.

(1) 12개의 LPC cepstral 계수

(2) 12개의 LPC cepstral 계수의 일차delta, 로그 파우워의 일차delta

(3) 12개의 LPC cepstral 계수의 이차delta, 로그 파우워의 이차delta

2.4 음소 모델

HMM 에 근거한 음성인식 시스템은 음성인식의 기본단위가 필요한데 우리는 음소와 유사한 단위(phoneme-like unit)를 선택하였다. 기본 유니트 갯수는 56개를 사용하였으며 조음화 현상을 고려하여 문맥종속 음소를 구하였다. 음소 모델은 7개의 상태와 12개의 전이를 갖으며 그 전이들은 3개의 그룹으로 묶을 수 있으며 같은 그룹의 전이는 같은 출력확률을 갖게 된다. 그림 2.에는 음소 모델의 topology가 그려져 있다[6].

2.5 조음화 현상 모델

연속음성에서 고려해야할 조음화 현상은 다음과 같이 세 종류가 있다.

(1) 묵음현상

연속음성을 발음할 경우 단어 사이의 묵음은 사람에 따라서 지키지거나 안이어서 발음을 할 수 있다. 그러므로 연속음성을 발음할 경우 묵음을 잘 모델링하여야 한다. 본 논문에서는 null transition을 만들어서 묵음을 모델링하였다.

(2) 단어내 조음화 모델

단어내의 조음화 현상을 모델링하기 위해 triphone을 사용하였다. Triphone 갯수를 결정할 때는 양이 너무 커지지 않게하기 위해서 unit reduction rule 을 사용하였다[2].

(3) 단어간 조음화 모델

연속음성에서 단어간 조음화 현상은 매 단어의 앞과 뒤에서 발생된다. 특히 Triphone을 이용한 연속음성인식시스템인 경우 단어간의 조음화 모델은 매우 복잡하다. 본 연속음성인식시스템은 훈련시에는 단어간 조음화 현상을 하나의 triphone으로 모델링하였으며 인식단계에서는 가능한 모든 triphone으로 모델링하였다. 그림 3.에는 단어간 조음화 현상을 고려한 triphone 구성이 그려져 있다.

본 논문에서는 위와 같은 세 종류의 조음화 현상을 고려하여 300개의 문맥종속 음소를 구하였다.

2.6 언어처리 알고리즘

연속음성에서 언어처리 알고리즘은 인식시간 및 성능에 중요한 역할을 한다. 현재 언어처리 알고리즘으로 finite state 문법이 많이 사용되고 있는데 문법이 복잡해지고 단어량이 증가하면 처리하기가 어렵다. 본 논문에서는 구문분석 모델로 bigram 문법을 사용하였다. Bigram 문법은 단어 w_i 다음에

단어 w_2가 올 확률값 $P(w_2/w_1)$을 훈련시 구하고 인식단계에서는 이 값을 단어 천이 확률값으로 사용한다[8]. 의미모델은 하지 않는다.

2.7 탐색 알고리즘

탐색 알고리즘으로 Viterbi 알고리즘을 사용하였으며 인식시간을 향상시키기 위해서 beam 탐색 알고리즘을 사용하였다. N개의 최적문장은 5개만 고려하였다. N개의 최적문장을 찾는 알고리즘은 다음과 같이 세 종류가 있다[9].

(1) Sentence-dependent 알고리즘

모든 state에서 N개의 가능한 path를 저장한다.

(2) Lattice 알고리즘

단어 내에서는 매 state에서 하나의 path만 저장하고 문법 node에서는 가능한 모든 path를 저장한다.

(3) Word-dependent 알고리즘

Sentence-dependent 알고리즘과 lattice 알고리즘의 중간으로 단어 내의 매 state에서 N개 보다 적은 n개의 가능한 path를 저장한다.

본 논문에서는 word-dependent 알고리즘을 구현하였다.

3. 인식 실험

3.1 데이타 베이스

남성 22인이 460단어로 구성된 문장을 일 인당 약 150문장씩 발음한다. 화자는 전화기 handset을 사용하여 발음하고 발음된 문장은 San SPARC 컴퓨터의 A/D 변환기를 통해 컴퓨터에 저장된다. 발음한 문장의 복잡도는 30이다. 표 1.에는 훈련에 사용된 화자와 성능평가에 사용된 화자의 분류가 나타나 있다.

표 1. 음성 데이타 베이스 구성

	훈 련 용	성능평가용
사 람 수	19	3
문장갯수	2885	453

3.2 실험 결과

표 2.에서는 연속음성의 성능결과가 나타나 있다. 단어 인식률은 다음과 같은 수식으로 구해진다.

$$\text{단어인식률} = (1 - \frac{(\text{치환}+\text{삭제}+\text{첨환})}{\text{전체 단어 갯수}}) \times 100\%$$

문장 인식률은 문장을 이루고 있는 모든 단어가 정확히 인식된 문장 갯수의 비를 나타낸다. 남성 화자에 대한 화자독립 단어 인식률은 95.7%이었으며 문장 인식률은 88.5% 였다. 5개의 최적문장을 고려할 때 단어 인식률은 98.6%, 문장 인식률은 76.5%로 향상되었다.

4. 결 론

본 논문에서는 한국통신 소프트웨어 연구소에서 개발한 한국어 연속 음성인식 시스템을 소개하고 성능평가 결과를 기술하였다. 개발된 연속 음성인식 시스템은 N개의 최적 문장을 찾을 수 있으며 언어처리 알고리즘으로 bigram을 사용하였다. 인식의 기본 단위로 triphone을 사용하였으며 단어간, 단어 내의 조음현상도 모델링 하였다. 또한 연속음성에서 묵음현상을 표현해 주기 위하여 null transition을 음소 모델에 추가 하였다.

462 단어로 구성되고 언어 복잡도가 30인 호텔예약에 관한 연속음성에서 남성화자 3인에 대한 화자독립 인식률을 실험하였다. 한개의 최적문장만 고려하면 단어 인식률은 95.7%, 문장 인식률은 88.5%를 얻었으며 다섯개의 최적문장을 고려하면 단어 인식률은 98.6%, 문장 인식률은 96.5%로 향상되었다.

참 고 문 헌

[1] R.C. Rose and E.M. Hofstetter, "Task-independent wordspotting using decision tree based allophone clustering," *Proc. ICASSP-93*, pp.467-470, 1993.

[2] C.H. Lee et al., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and language*, vol.4 pp.127-165, 1990.

[3] L.R. Bahl et al., "A fast approximate acoustic match for large vocabulary continuous speech recognition," *IEEE Tr. Speech and Audio Processing*, vol.1, No.1, pp.59-67, 1993.

[4] F.K. Soong and E.F. Huang, "A tree-trells based fast search for finding the N-best sentence hypotheses in continuous speech recognition," *Proc. ICASSP-91*, pp.703-706, 1991.

[5] R. Schwartz and Y.L. chow, "The N-best algorithm : an efficient and exact procedure for finding the N most likely sentence hypotheses," *Proc. ICASSP-90*, pp.81-84, 1990.

[6] K.F. Lee, *Automatic speech recognition : the development of the SPHINX system*, Kluwer Academic publisher, Norwell, Mass., 1989.

[7] E.P. Giachin et al., "On the use of inter-word context-dependent units for word juncture modeling," *Computer Speech and language*, vol.6, pp.197-213, 1992.

[8] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, pp.1616-1624, 1985.

[9] R. Schwartz and S. Austin, "Efficient, high-performance algorithms for N-best search," *Proc. of the DARPA speech and natural language workshop*, pp.6-11, 1990.

표 2. 연속 음성인식 성능결과

화자	전 체 단어수	전 체 문장수	첨가	삭제	치환	단 어 인식률(%)		문 장 인식률(%)	
						Top 1	Top 5	Top 1	Top 5
A	581	151	4	0	10	97.8	99.3	93.4	98.0
B	584	154	7	0	24	94.7	98.3	85.7	95.5
C	571	148	11	0	18	94.9	98.3	86.5	96.0
합계	1736	453	22	0	52	95.7	98.6	88.5	96.5

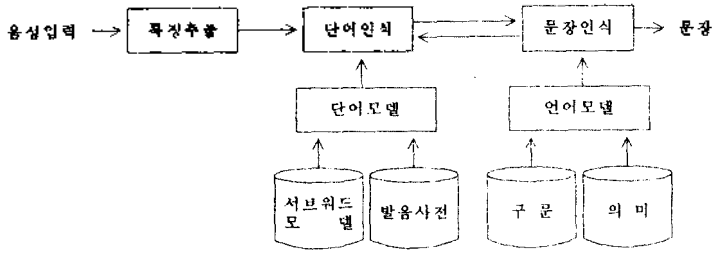

그림 1. 연속음성인식 시스템 개요

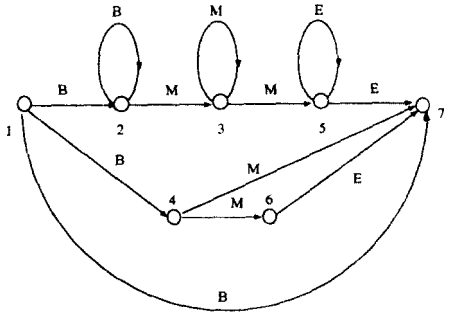

그림 2. 음소 모델 topology

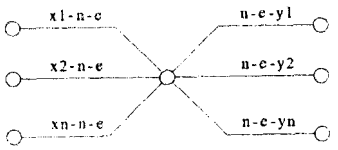

그림 3. 단어간 조음화 모델