FSN 기반의 대어휘 연속음성인식 시스템 개발

박 전 규, 이 윤 근
한국전자통신연구원 음성/언어정보연구센터

Development of FSN-based Large Vocabulary Continuous Speech Recognition System

Jeon Gue Park, Yunkeun Lee
Speech/Language Information Research Center, ETRI
E-mail : {jgp, yklee}@etri.re.kr

Abstract

This paper presents a FSN-based LVCSR system and it's application to the speech TV program guide. Unlike the most popular statistical language model-based system, we used FSN grammar based on the graph theory-based FSN optimization algorithm and knowledge-based advanced word boundary modeling. For the memory and latency efficiency, we implemented the dynamic pruning scheduling based on the histogram of active words and their likelihood distribution. We achieved a 10.7% word accuracy improvement with 57.3% speedup.

I. 서론

대어휘 연속음성인식의 전형인 받아쓰기(dictation), 대화 시스템 등을 위해서는 대용량의 말뭉치(corpus)로부터 통계적 언어모델을 계산하여 음성인식 탐색엔진에서 사용하는 것이 일반적이다. 그러나 복잡한 어휘구문론적 특성 및 전문용어를 포함하는 대상 응용영역이 주어졌을 때 언어모델 산출에 적절한 크기의 통계적 특성을 가진 말뭉치가 가용하지 않을 경우가 대부분이다. 이때 전문가에 의해 잘 정의된 문틀(sentence pattern)을 구축하고 문틀로부터 문맥자유문법 또는 FSN(Finite State Network)을 계산해서 탐색엔진에 적용하게 된다. 또는 이렇게 잘 정의된 FSN을 사용해서 문장을 생성한 다음 통계적 언어모델을 구하는 방법도 있다[1].

한편 바이그램 또는 트라이그램 수준의 통계적 언어모델을 채용하는 탐색기법과 마찬가지로 문틀로부터 추정된 FSN에 기반한 탐색기법도 자유도 및 복잡도(perplexity)가 높아질 경우 탐색공간의 폭발적 증가와 탐색시간의 지연으로 인해 범용 컴퓨터 대상의 실용적인 음성인식 시스템의 개발이 어렵게 된다.

본 논문에서는 이러한 배경에서 전문가에 의해 구축된 특정 응용영역 FSN에 기반한 대용량 어휘 음성인식 시스템의 구현에 대해 기술하고 있으며 특히 프루닝계획법(pruning scheduling)에 의한 인식 속도 및 성능 개선에 대해 기술하고 있다. 대상은 EPG(Electronic Program Guide)를 기반으로 하는 음성인식 TV가이드 시스템이다. II장에서는 음성인식 엔진의 구성, III장에서는 음성 데이터베이스, IV장에서는 실험 및 결과, 마지막 V장에서 결론을 기술한다.

II. 대어휘 음성인식 시스템

그림 1은 대어휘 연속음성인식을 위한 ETRI의 음성인식 엔진인 ESTk[4]의 시스템 구성을 도해하고 있다. 자연어처리 전문가에 의해 잘 정의된 문틀은 표 1의 예와 같이 클래스 기반의 BNF 형태로 기술되며, 이를 BNF 파서의 입력으로 사용하여 FSN 문법을 생성한다. 그래프 이론에 기반한 최적화 알고리즘[2]에 의해

FSN의 문법 노드와 링크를 최소화하여 음성인식 엔진에서 직접 사용되는 최종 FSN 문법이 생성된다. BNF 파서에 의해 추출된 인식용 어휘는 발음열 생성기에 따라 발성 사전을 생성하게 되는데 이때 운율정보에 기반하여 예외 발성 가능성이 있는 단어를 자동으로 검출하여 다중 발성 사전을 생성한다.[5]

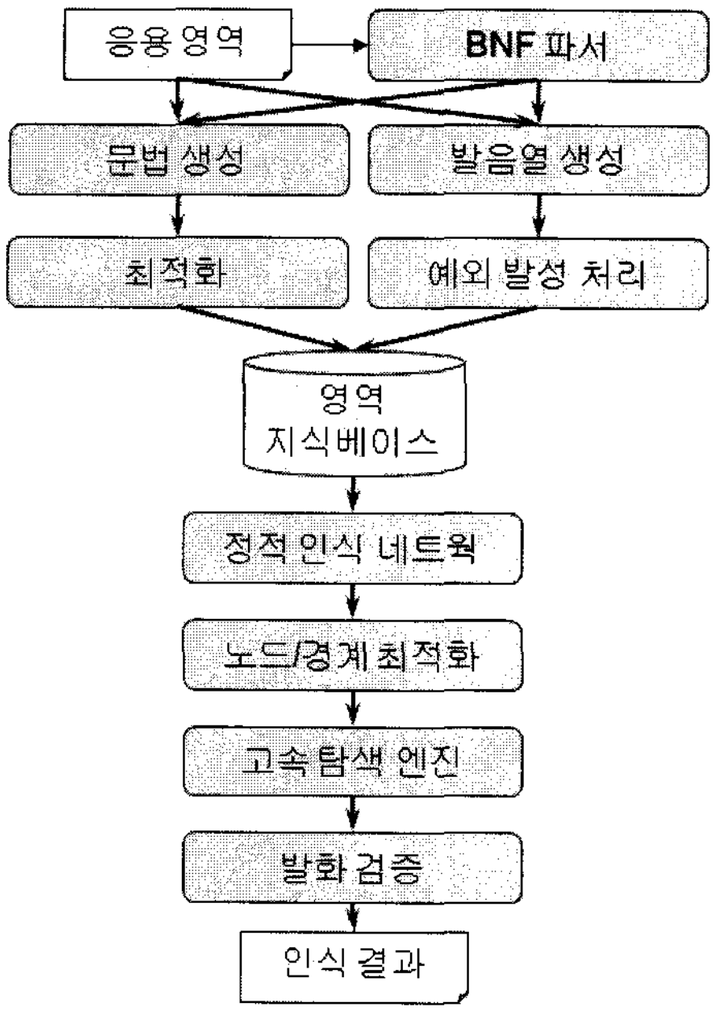

그림 1. FSN기반 음성인식 시스템의 구현

표 1. BNF 기술 예

```
<person> ::= 감우성 | 강남길 | 강동원;
<program> ::= 007 골든아이 | MBC 뉴스현장;
<among> ::= 이 중에서 | 이 중에 | 그 중에서;
<tv_find> ::= 검색해 | 검색해 줘 | 찾아봐;
<sentence> ::= <among> <person> <tv_find> |
<program> <tv_find>;
```

영역 지식베이스는 음향모델, 언어모델, 발성사전, 영역지식 등을 포함하는 데 특히 발성사전의 개별 엔트리는 인식 오류의 후처리 및 대화 모델에서 직접 사용될 수 있도록 표 2와 같이 품사 태그 및 클래스 정보 등이 부가되어진다. 첫 번째 열은 음소열로 변환될 발성형태, 두 번째 열은 클래스 정보가 부착된 인식용 어휘, 세 번째 열은 형태소 해석에 따라 ETRI 품사 태그가 부착된 결과이다.

표 2. 어휘 사전의 구성 예

발성형태	표제어/클래스	품사태그
황신혜	황신혜{<person_all>}	황신혜/nq
엠비씨	MBC{<channel>}	엠비씨/nq
엠비씨뉴스현장	MBC_뉴스현장{<program-뉴스-뉴스>}	엠비씨/nq@뉴스/ncn@현장/ncn

FSN과 발성사전에 따라 트리탐색에 기반한 래티스가 생성되는 데 초기화시 모든 문법 네트웍이 확장되는 정적네트웍을 사용한다. 다음 '노드/경계 최적화' 단계에서는 정적 네트웍에 대해 언어학적, 음향학적 지식 및 단어경계 트라이폰 최적화에 기반한 단어 경계 모델링(word boundary modeling)을 수행한다.

탐색 엔진은 GMM 선택법, KNN 추정, 프레임별 활성 단어의 수와 우도(likelihood)값의 관계에 대한 히스토그램 분석을 통해 가변적인 프루닝 값을 적용하는 프루닝 계획법[3], SIMD에 기반한 병렬 확률밀도함수 추정 등이 적용된 고속의 탐색 기법에 기반하여 N개의 인식 후보열을 생성한다.

III. 음성 DB 수집

20명의 남녀 화자를 대상으로 화자별로 다르게 설계된 5개의 시나리오에 따라 녹음을 수행하도록 했다. 배포된 음성 TV 가이드 시스템의 사용자 설명서에 따라 표 3의 예와 같이 하나의 시나리오는 5-8개의 질의 문장을 포함하도록 구성하였다. 이에 따라 총 1,810개의 단어로 구성되는 642 발화를 수집하였다. 조용한 사무실 환경에서 개인용 PC에 헤드셋을 사용하여 음성인식 엔진이 연동된 상태에서 하드디스크에 직접 녹음 및 저장을 수행하였다.

표 3. 예제 시나리오

1) 시나리오 1
   - 오늘 SBS 드라마 검색해
   - 사랑과 야망 녹화해
   - 오늘 MBC에서 뉴스 검색해
   - 3번 자세히
   - 이거 녹화해
   - 녹화 리스트

2) 시나리오 2
   - 탁재훈 나오는 프로 검색해
   - 1번 자세히
   - 31번으로 돌려
   - 06/07 프로농구 시범경기 녹화해
   - 내일 2시 이전 오락 프로 검색해
   - 무한도전 알람 설정해

IV. 실험 및 결과

음성 TV 가이드 시스템에 대한 성능 평가를 위해 표 1의 예와 같이 BNF로 정의된 290개 정도의 문틀이 수집되었다. 최적화를 통해 생성된 최종 FSN의 문법 복잡도는 약 160이다. 최종 FSN의 문법 노드의 수는 1,850개, 아크의 수는 24,349개, 사용된 어휘는 약 15,000개이다. 케이블 TV의 20개 채널만을 대상으로 검색용 어휘를 수집하여 구성하였는데 프로그램 및 출연자명은 각각 1,002개와 1,038명이다.

642개의 테스트 발성 중에서 약 20%정도의 문장을 OOG 및 OOV 발화에 해당하도록 문법을 구성하였는데 이중 OOG에 해당하는 발성은 13.5%(84개)이다. 표 4는 수집된 음성 데이터베이스에 대해서 실험한 결과를 요약하고 있다. 베이스라인은 프루닝 계획법을 적용하기 이전의 성능이다. 프루닝계획법 적용 시 81.8%의 단어인식성능을 얻었으며 오류개선율은 10.6%이다. OOV와 OOG를 배제하면 98.5%의 단어인식 성공률을 얻었다.

그림 2에서는 일반적인 고정 프루닝 값을 적용해서 프레임 동기 탐색을 수행할 경우(베이스라인)의 프레임별 활성 단어의 수와 프루닝 계획법(프루닝)에 의한 활성 단어의 수를 비교하고 있다. 응답시간의 관점에서 펜티엄4, 2.8GHz CPU를 적용하여 57.3% (1.57xRT)의 속도 개선 효과를 얻었다.

표 4. 실험 결과 (단위: %)

구분	베이스라인	프루닝계획법	오류개선율
단어인식율	79.7	81.8	10.6
문장인식율	67.4	69.8	7.4

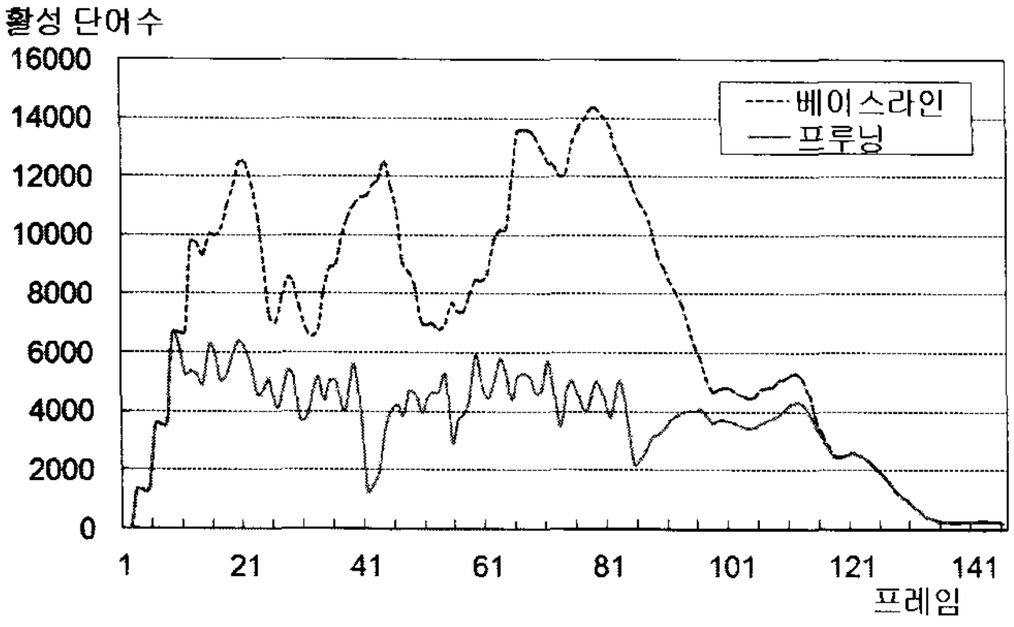

그림 2. 프루닝 계획법에 따른 프레임별 활성 단어

V. 결론

본 논문에서는 FSN에 기반한 대어휘 음성인식 시스템을 소개하고 그 응용 및 실험을 위해 TV가이드 영역을 대상으로 사용하였다. 잘 정의된 BNF 문법을 파싱하여 FSN을 생성하였으며 FSN 최적화 알고리즘을 적용하여 70% 정도의 문법노드와 링크를 소거하였다. 또한 언어학적, 음성학적 지식에 근거한 개선된 단어경계 모델링을 통해 80% 정도의 단어경계 트라이폰을 최적화하였다. 마지막으로 프레임별 활성단어와 그 우도의 히스토그램에 기반한 프루닝계획법을 적용하여 10.7%의 오류개선율을 얻었으며 실시간 성능을 57.3% 개선하였다.

참고문헌

[1] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here", Proceedings of the IEEE, 88(8), 2000

[2] M. Mohri, "Finite-state transducers in language and speech processing," Computational Linguistics, 23(2), pp. 269-311, 1997

[3] H. Van hamme and F. Van Aelten, "An adaptive-beam pruning technique for continuous speech recognition," ICSLP, pp.2083-2086, 1996

[4] 박전규, 이성주, 김상훈, "범주적 필러 모델과 음성 개선에 의한 원거리 핵심어 인식," 한국음향학회 제 22회 음성 통신 및 신호처리 학술대회, pp.73-78, 2005

[5] 김선희, 박전규, 나민수, 전재훈, 정민화, "운율 정보를 이용한 한국어 위치 정보 데이터의 발음 모델링," 정보과학회 논문지, 제 34권 제 2호, pp.95-102, 2007