제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

가변 어휘 인식 모델을 이용한 한국어 방송 뉴스 음성의 인식

유하진, 김훈, 최재승, 이종석
LG 종합기술원 정보기술연구소, MI group
E-mail: {hajin, hkim, seung111, ljs}@lgcit.com

Automatic Recognition of Korean Broadcast News Using Flexible Vocabulary Recognition Models

Ha-Jin Yu, Hoon Kim, Jae-Seung Choi, and Jong-Seok Lee
LG Corporate Institute of Technology
Information Technology Lab. MI group

요약

본 논문에서는 한국어 방송 뉴스 인식 시스템에 관하여 기술한다. 인식 실험 과정에서는 실제로 방송된 음성을 인식하였으나, 인식을 위한 음향 모델은 본 연구소에서 개발한 고립 단어 인식용 가변 어휘 인식 모델을 이용하였다. 가변 어휘 인식기는 방송 음성의 연속 문장을 이용하지 않고, 음향학적으로 고르게 분포된 고립 단어를 이용하여 학습되었다. 본 연구에서는 한국어의 특성상 문장이 영어권과 같이 단어 단위가 아닌 어절로 나누어 지는 점을 고려하여, 다양한 형태의 시간 표제어를 대상으로 실험하였다. 또한 탐색과정의 초기단계에 장거리(long-distance) 언어모델을 사용함으로써 인식 오류를 줄일 수 있었다.

1. 서론

현재 세계적인 음성 인식 시스템의 최첨단 수준이 어느 정도 인가를 알아볼 수 있는 방법 중의 하나는 DARPA에서 매년 실시하고 있는 벤치마크 테스트에 참여하는 여러 팀 등의 연구 실적을 살펴보는 일일 것이다. 최근에는 방송 뉴스 음성 인식을 대상으로 하고 있는데, 방송 뉴스 음성은 대어휘 연속 음성 인식 시스템을 개발하기 위하여 풀어나가야 할 많은 문제들을 충분히 갖추고 있다고 할 수 있다[1]. 방송 음성은 아나운서의 깨끗한 음성에서부터, 복잡한 거리에서의 인터뷰 등 다양한 특성을 가지고 있다. 거리의 소음이나 배경 음악 등이 포함되어 있고, 넓은 주파수 대의 음성과 전화 음성이 함께 포함되어 있을 수 있으며, 구어체의 문장, 어법에 맞지 않는 문장 등이 다양하게 포함되어 있다.

DARPA의 테스트에 참가하는 많은 팀 들은 벤치마크를 위하여 제공되는 충분한 양의 데이타를 이용하여 통계적인 학습 등을 통한 시스템의 개량을 할 수 있다. 그러나, 한국어를 대상으로 인식 실험을 하는 경우에는 사용할 수 있는 자료의 양이 극히 제한적이어서 현재 알려져 있는 우수한 방법을 시험해 보고 새로운 방법을 개발하기 어려운 상황이다. 따라서, 본 연구에서는 주어진 자원을 최대한 활용하여 방송 뉴스 음성 인식기를 구축하고자 한다. 그 첫번째 단계로 HMM 모델의 학습을 위하여 대량의 방송 녹음 음성을 이용하지 않고 이미 구성되어 있는 가변어 어휘 인식기를 사용하였다. 음향학적으로 고르게 분포된 고립 단어를 선정하여 다수의 화자가 발성한 음성 자료를 이용함으로써, 장기간에 걸친 방송 음성의 수집을 필요로 하지 않고 방송 음성 인식 실험을 할 수 있었다.

한국어는 영어권의 언어와 같이 문장이 단어 보다는 어절로 나누어 지므로, 동사나 형용사의 여러가지

활용형을 사용할 경우 영어에서의 단어와 같은 의미의 사전 표제어가 상당히 많아지게 된다. 따라서, 본 연구에서는 인식 단위를 명사, 조사, 동사 및 형용사의 어간 및 어미 등으로 세분화한 경우에서, 문장에서 사용된 어절의 형태 그대로의 경우까지 다양한 형태에 대하여 실험하였다.

또한 일반적인 탐색 과정의 초기 단계에서는 단순한 bigram 만을 사용하게 되는데, 이때 제거된 단어의 가설은 이후의 단계에서 복구하기 어렵게 된다. 따라서, 본 연구에서는 탐색의 초기 단계에서 추가로 사용할 수 있는 언어모델을 제안하였다.

본 논문의 구성은 다음과 같다. 먼저 다음 장에서는 가변어휘 인식기의 학습 및 인식에 사용된 음성 자료에 대하여 기술하고, 특징추출과 음향학적 모델 학습 및 인식 과정 등 시스템의 개요, 3 장에서는 제안된 언어모델을 설명하며, 4 장에서는 실험 결과를 정리하고, 5 장에서 결론을 맺는다.

2. 음성 자료 및 시스템 개요

가변 어휘 인식 모델의 학습을 위한 자료는 음향학적으로 고르게 분포된 6,700 개의 명사를 240 명의 남자, 160 명의 여자가 나누어 발성한 총 40,000 개의 고립 단어 음성을 사용하였다. 한국어의 동사와 형용사에는 '다' 등과 같이 동일한 글자로 끝나는 형태가 많으므로, 명사만을 대상으로 선정하였다.

인식 대상으로는 TV 수신기를 이용하여 수신된 9 일간의 뉴스 음성 3,516 문장과 65 일간의 일기예보 음성 857 문장을 사용하였다. 사용된 어휘 수는 17,671 개이다. 음성 자료는 최대 한 문장에 70 단어까지 포함하고 있으며, 평균 초당 약 21 음소의 속도로 발음되었다. 테스트를 위한 음성 자료는 문장 별로 나누어져 인식기의 입력이 되었다. 음성자료를 발성 화자, 발성 환경에 따라 분류하면 표 1 과 같다.

표 1. 자료의 분류

	남/녀	조용한곳	잡음이 있는곳	합계
앵커	남	430	36	466
	녀	352	47	399
리포터	남	479	1429	1908
	녀	48	207	255
인터뷰	남	59	366	425
	녀	3	60	63
합계		1371	2145	3516

음성 자료는 16kHz 16bit 로 표본화 되어 10ms 간격으로 20ms 의 해밍창을 이용하여 분석되었다. 특징 벡터로는 12 차 LPC 켑스트럼 계수와 24 차의 차분 켑스트럼, 12 차의 이차 차분 켑스트럼과 에너지, 일차, 이차 차분 에너지를 사용하였다. 24 차의 차분 켑스트럼은 40ms 와 80ms 차이로 얻어진 특징 벡터들로 구성된다. 이 4 개의 특징 벡터들로부터 각각 코드북을 생성하여 독립적으로 사용하였다. 인식 모델로는 파라미터를 공유하는[2] 환경 종속 (context dependent) 음소 SCHMM 을 사용하였다. 하나의 음소 모델은 3 개의 상태로 구성된 left-to-right 모델이다. 52 개의 기본 음소 모델의 학습 결과 얻어진 triphone 의 수는 12,261 개이다.

단어 내에서는 triphone 을 이용하고, 단어 경계에서 발생하는 음소의 경우에는, 단어의 시작 부분에는 우측 환경, 끝 부분에서는 좌측 환경에 각각 종속적인 음소 모델을 사용하였다. 단어에서 필요로 하는 triphone 음소 모델이 학습된 모델에 존재하지 않는 경우에는 monophone 을 사용하였다.

본 시스템의 인식 과정은 그림 1 과 같다. 인식 부분의 탐색 과정은 3 단계로 이루어진다[3][4].

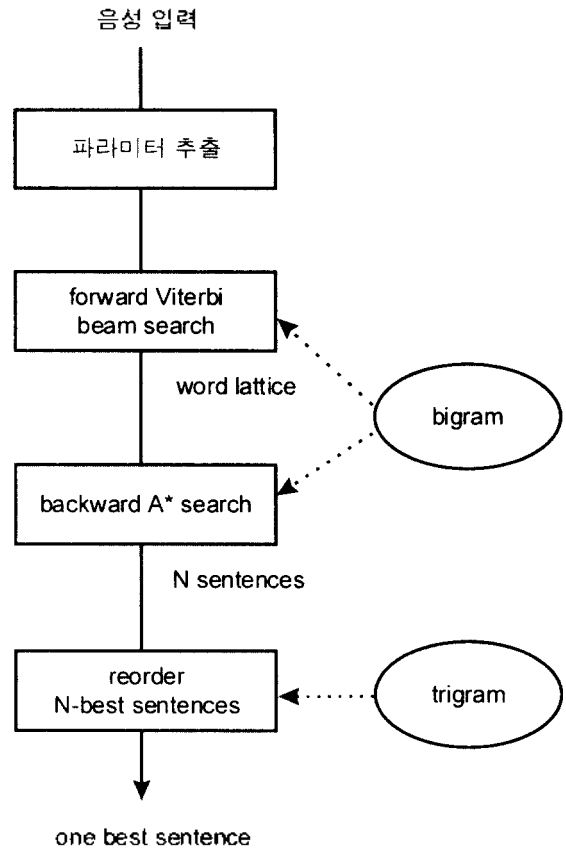

그림1. 탐색 과정

첫번째 과정에서는 bigram 과 다음 장에서 제안한 언어 모델을 이용한 시간 동기 (time synchronous) Viterbi 빔 탐색을 사용한다. 두번째 과정에서는 bigram 과 제안한 언어 모델을 이용한 시간 비동기 (time asynchronous) A* 역방향 탐색을 이용하여 N 개의 문장 후보를 구한다. 세번째 과정에서는 앞의 과정에서 얻어진 N 개의 문장 후보를 trigram 을 이용하여 재배열하여 최종적으로 하나의 문장을 얻어낸다. 본 연구에서는 10 개의 후보를 사용하였다.

3. 제안한 언어모델

Bigram 이나 trigram 은 인접해 있는 단어들 간의 관계는 잘 표현하여 탐색 과정에서 사용하기 용이 하지만, 이러한 언어 모델을 이용한 인식 결과는 문장 전체로 볼 때 옳지 않은 경우가 많다. 예를 들어, 인식한 문장 중에서 '내일은' 으로 시작하는 단문의 끝이 '맑았습니다' 로 끝나거나, '절대로'가 들어간 문장의 끝이 부정문으로 끝나지 않는다면 인식 결과에 오류가 있음을 예상할 수 있다. 그렇지만, 기존의 단거리 (short-term) n-gram 으로는 이와 같은 문제를 해결할 수 없다.

본 연구에서 제안하는 언어 모델은 한 문장 내에서 '내일은' 이라는 단어 이후에 '맑겠습니다'라는 단어가 나올 확률과 같이, 바로 인접한 단어간이 아닌 문장 내의 이후 또는 이전에 단어가 나타날 확률을 표현한다. bigram 에서 $P(w_j|w_i)$가 단어 w_i 의 바로 다음에 단어 w_j 가 나올 확률을 표현한다고 하면, $P_{forward}(w_j \mid w_i)$는 문장 내에서 단어 w_j 이전에 단어 w_i 가 나올 확률을 표현한다. 마찬가지로 $P_{backward}(w_j \mid w_i)$는 문장 내에서 단어 w_j 이후에 단어 w_i 가 나올 확률을 표현한다. 이 두 가지 확률을 각각 전향존재확률과 후향존재확률이라고 부르기로 한다. 전향존재확률 $P_{forward}$ 은 전향 탐색을 할 때 사용할 수 있으며, 후향존재확률 $P_{backward}$ 은 후향탐색을 할 때 사용할 수 있다.

이와 유사한 연구로는 SPHINX-II 에서 사용된 장거리 (long distance) bigrams [5] 이나 확장 (extended) bigram [7]이 있다. 일반적으로 장거리 모델 (long distance model)을 계산하기 위해서는 이전 정보를 가지고 있어야 하므로, 모든 영역을 탐색하는 초기 단계에서는 주로 사용되지 않고, 그 이후에 N 개의 후보 문장을 갖는 최우선 (best-first) 탐색 과정이다[5]. 마지막 단계에서 N 개의 후보 문장을 재정렬 (reordering) 하는데[6] 주로 사용된다. 그러나, 가장 첫번째 패스인 시간동기 Viterbi 빔 탐색과정에서 작은 빔 크기로 말미암아 올바른 단어가 제거 된다면 이후 단계에서 아무리 정밀한 탐색을 하여도 복구할 수 없게 된다. 따라서, 본 연구에서는 제안한 전향 존재 확률을 가장 첫번째 단계에서 사용하였다. 즉, 초기 전향 탐색의 새로운 단어로의 천이 과정에서, 그 시점의 이전에 찾아낸 모든 단어 가설을 대상으로 새로운 단어의 전향 존재 확률을 구한다. 이렇게 함으로써 가장 중요한 단계인 첫 단계에서 보다 많은 정보를 이용할 수 있게 된다. 두 번째 단계인 역방향 A* 탐색과정에서는 후향 존재확률을 이용한다.

4. 실험 및 결과

한국어에서는 동사나 명사가 여러 가지 활용형으로 변하므로, 인식기를 위한 사전을 구성할 때 같은 의미를 가지는 표제어가 상당히 많이 존재하게 된다. 가장 이상적인 방법으로는 사전에 기본형만을 저장해 놓고 문맥에 따른 여러 가지 활용형을 고려하며 탐색하는 방법이 있겠지만, 현재의 탐색 방식으로는 아직 해결 방법이 없다. 사전 표제어를 형태소 단위로 할 수도 있지만, 그렇게 하면 많은 표제어의 길이가 1~3 개의 음소로 구성되게 되어 인식률의 저하가 따르게 된다.

본 연구에서는 사전 표제어의 수에 따른 인식률 변화와 제안된 언어 모델의 성능을 평가하기 위하여, 65 일간의 뉴스 중 일기예보 부분만을 분리하여 857 문장을 사용하여 실험 하였다. 먼저 제안된 언어모델을 사용하지 않은 상태로, 단어의 크기를 3 단계로 세분화하여, 세 가지의 어휘 세트로 실험 하였다. 즉, 어절을 의미 또는 문장내의 기능을 가지는 가장 최소 단위로 나눈 경우에서부터 이절을 그대로 사용하는 경우 까지로 나누었다. 표 2 의 ① ~ ③ 에서 보이는 것과 같이 각각 972, 1496, 2288 개의 표제어를 등록하여 실험한 결과, 표제어의 단위를 세분화 하면 큰 인식률의 저하가 있음을 알 수 있었다. 이 결과는 많은 단어가 너무 적은 수의 음소로 이루어져 있기 때문으로, 단어간 (cross-word) 음소 모델을 사용하면 향상될 것으로 보인다.

한편, 일기예보 방송 가운데에는 "내일은 맑은 날씨가 ..." 등과 같이 일기와 관련된 어휘를 사용하는 경우와, 지두에서 날씨에 따른 건강 관리에 관련된 내용을 말하는 경우 등과 같이 일기와 관련이 없는 어휘를 사용하는 경우가 있다. 본 실험에서는 일기예보에 주로 쓰이는 어휘만을 사용한 부분을 따로 묶어 인식 실험에 사용해 보았다. 그 결과, 표 2 의 ④에서 보이는 것과 같이 비교적 높은 인식률을 얻을 수 있었다.

또한, 화자 종속 인식 실험을 위해 857 문장을 이용하여 모델 학습을 하고, 학습에 사용하지 않은 195 문장을 인식 실험에 사용한 결과, 표 3 의 ④과 같이 높은 인식률을 얻을 수 있었다. 표 3 의 ⑤는 앞서와 같이 358 개의 일기에 관련된 어휘만을 사용한 문장에 대

한 실험 결과이다.

제안한 전향 및 후향 존재 확률 언어 모델의 성능을 평가하기 위하여 표 2 의 ③ 의 경우에 대하여 실험하였다. 이때, 모든 경우에 대하여 매 프레임당 활성 상태 (active state) 수를 일정 수로 유지하였다. 그 결과, 단어 인식 오류가 16.3%에서 14.8% 로 8.8% 감소하였다. 인식 시간은 울트라스팍 워크스테이션에서 문장당 평균 11.1 분에서 10.5 분으로 10.6% 감소하였다.

17,671 단어의 사전을 이용한 전체 뉴스 3,516 문장의 인식 실험 결과 72.0%의 단어 인식률, 56.8%의 문장 인식률을 얻을 수 있었고, 문장 당 평균 인식 시간은 15.9 분 이었다.

5. 결론

본 연구는 앞으로의 한국어 방송 뉴스 음성 인식을 위한 예비 실험이다. 음향학적 모델은 방송 뉴스 음성을 사용하지 않고 가변 어휘 인식기를 위한 모델을 사용하였다. 문장에 포함된 단어를 여러 단계로 세분화 하여, 세 가지의 단어 사전을 이용한 결과, 사전의 표제어의 수가 1000 단어에서 2300 단어까지 변화함에 따라서 단어 인식률이 50%에서 90%까지 변화 하였으며, 어절을 세분화하지 않고 사용한 경우가 가장 높은 인식률을 나타내었다. 또한 제안한 언어모델을 탐색 과정의 초기 단계에 사용함으로써, 인식 오류를 8.8% 감소시키고, 인식 시간을 10.6% 절감하는 효과를 얻을 수 있었다.

본 연구의 결과에 따라, 음향학적으로 고르게 분포된 고립 단어를 이용하여 학습한 가변 어휘 인식기에 언어 모델이 주어지면, 임의의 응용 분야의 연속 음성 인식이 가능하다는 것을 확인할 수 있었다.

6. 참고문헌

[1] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker & S.J. Young, "The 1997 HTK broadcast news transcription," Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Virginia.

[2] M. Y. Hwang, X. D. Huang, F. A. Alleva, "Predicting Unseen Triphones with Senones," IEEE Trans. Speech and Audio Processing, vol. 4, no. 6, pp. 412-419, Nov. 1996

[3] 이승배, 이종석, "N-Best 문장탐색기법을 이용한 연속음성 인식시스템," 제 13 회 음성통신 및 신호처리 워크샵, pp. 151-154

[4] R. Schwartz, 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

[5] Rosenfeld, R. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, 1994

[6] Niesler, T. Category-based statistical language models, Ph.D. thesis, St. John's College, 1997

[7] J.H.Wright, G.J.F.Jones and H.Lloyd-Thomas, A consolidated language model for speech recognition, Proc. of Eurospeech 93, Berlin, vol. 2, pp. 977-980, 1993

표 2. 어휘 세트에 따른 인식률 변화

구분	문장수	단어수	단어 인식율	문장 인식율
①	857	972	37.0 %	11.1 %
②	857	1496	49.9 %	15.9 %
③	857	2288	84.2 %	59.2 %
4	213	358	89.3 %	64.3 %

표 3. 화자 종속 모델 인식 결과

구분	문장수	단어수	단어 인식율	문장 인식율
1	195	757	97.1 %	91.8 %
2	88	213	98.1 %	96.6 %