1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

연속음성인식 시스템의 성능 향상을 위한 반복학습법을 이용한 언어모델

o오 세 진*, 황 철 준*, 김 범 국**, 정 호 열*, 정 현 열*

영남대학교 정보통신공학과*

대구과학대학 전자과**

Language Models Using Iterative Learning Method for the Improvement of Performance of CSR System

Se-Jin Oh*, Cheol-Jun Hwang*, Bum-Koog Kim*, Ho-Youl Jung*, Hyun-Yeol Chung*

Dept. of Information & Communication Eng., Yeungnam Univ.*

Dept. of Electronics, Taegu Science College**

요 약

본 연구에서는 연속음성인식 시스템의 성능 향상을 위하여 음성의 채록환경 및 데이터량 등을 고려한 효과적인 언어모델 작성방법을 제안하고, 이를 항공편 예약 시스템에 적용하여 성능 평가 실험을 실시한 결과 91.6%의 인식률을 얻어 제안한 방법의 유효성을 확인하였다.

이를 위하여 소량의 200 문장의 항공편 예약 텍스트 데이터를 이용하여 좀더 강건한 단어발생 확률을 가지도록 하기 위해 일반적으로 대어휘 연속음성인식에서 많이 이용되고 있는 단어 N-gram 언어모델을 도입하고 이를 다양한 발성환경을 고려하여 1,154 문장으로 확장한 후 동일 문장을 반복 학습하여 언어모델을 작성하였다. 인식에 있어서는 오인식과 문법적 오류를 최소화하기 위하여 forward - backward pass 방법의 stack decoding 알고리즘을 이용하였다.

인식실험 결과, 평가용 3 인의 200 문장을 각 반복학습 회수에 따라 학습한 각 언어모델에 대해 평가한 결과, forward pass 의 경우 평균 84.1%, backward pass 의 경우 평균 91.6%의 문장 인식률을 얻었다. 또한, 반복학습 회수가 증가함에 따라 backward pass 의 인식률의 변화는 없었으나, forward pass 의 경우, 인식률이 반복회수에 따라 증가하다가 일정값에 수렴함을 알 수 있었고, 언어모델의 복잡도에서도 반복회수가 증가함에 따라 서서히 줄어들며 수렴함을 알 수 있었다.

이상의 결과로부터 소량의 텍스트 데이터를 이용한 제한된 태스크에서 언어모델을 작성할 때 반복학습 방법이 유효함을 확인할 수 있다.

1. 서 론

현재까지의 음성인식에 관한 연구는 주로 잡음이 없는 환경에서 성능이 좋은 인식기를 개발하는 것이 주요 관점이었으나 최근에는 점점 실생활과 관련된 응용분야에 많은 관심을 가지게 되었다. 예로서, TV 와 라디오 방송언어를 이용한 음성인식을 들 수 있다[3].

현재, 외국의 경우 Wall Street Journal 과 같은 신문에 대해 여러 가지 언어에 대한 대어휘 연속음성인식에 관한 연구가 활발하게 수행되고 있다[2,5]. 특히, 방송언어를 이용한 음성인식의 경우 1995 년부터 DARPA 에서 지금까지 계속 연구를 수행하고 있다[3].

어휘수가 증가할수록 적절한 문법을 작성하는데 어려움이 있는데 이를 해결하는 방법의 하나로 N-gram 언어모델을 이용하고 있다. N-gram 언어모델의 확률 값을 잘 추정하기 위해서는 많은 량의 텍스트 데이터가 필요하며, 대부분의 대어휘 인식에 관한 연구에서 대규모의 텍스트 데이터로부터 언어모델을 효율적으로 작성하는 연구가 활발하게 수행되고 있다. 그러나, 훈련용 데이터가 소량인 경우 이를 효과적으로 구성하기 어렵다.

따라서, 본 논문에서는 제한된 태스크에서 소량의 텍스트 데이터를 이용하여 언어모델을 효율적으로 작성하는 방법을 제안하고 작성한 언어모델의 유효성을 인식실험을 통해 확인하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 시스템의 전체 구성과 사용된 음성 데이터 및 분석조건에 대해서 설명하고, 3 장에서는 소량의 텍스트 데이터를 이용한 효율적인 언어모델 작성방법, 4 장에서 forward – backward pass 의 단계적 인식방법, 그리고 5 장에서는 인식실험 및 결과에 대해서 설명한다. 마지막으로 6 장에서 본 논문에 대한 결론을 맺는다.

2. 시스템 구조

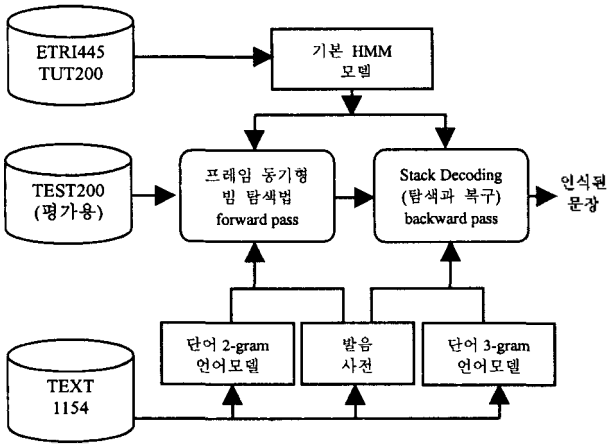

그림 1. 시스템의 전체 구성도

2.1 음성 데이터

그림 1 에 시스템의 전체 구성을 나타내었다. 시스템에서 사용한 음향모델은 left-to-right 의 5 상태 4 천이로서 초기상태와 최종상태에서는 확률분포가 없는 연속분포 HMM 으로 구성하였다. 또한 묵음을 포함한 51 개의 유사음소단위로서 ETRI 445 데이터 베이스 중 19 명의 단어발성과 TUT200 데이터 베이스 중 8 명의 200 문장발성으로 HTK[7]를 이용하여 작성하였다. 단어와 연속음성 사이의 혼동을 피하기 위해 단어와 연속음성을 같이 학습하였다. 그리고 TUT(Toyohashi University of Technology)의 다른 3 명의 200 문장을 평가용으로 설정하였다.

모든 음성 데이터는 16kHz, 16bits 로 샘플링 하고 $1-0.97z^{-1}$의 필터로 프리엠퍼시스하였다. 입력음성의 각 프레임에 25msec 의 Hamming windows 를 곱하여 10msec 마다 분석하였으며, 이를 통해 추출한 12 차의 MFCC 와 power 성분 그리고 1 차, 2 차의 미분 MFCC 와 각각의 미분 power 로서 모두 39 차의 특징 파라미터를 구성하였다. 또한, 모든 음성 채널을 고려하여 켑스트럼 평균 정규화(CMN)를 수행하였다. 표 1 에 음성 데이터의 분석조건을 나타내었다.

표 1. 음성 데이터의 분석조건

Sampling frequency	16kHz
Resolution	16bits
Pre-emphasis	$1-0.97z^{-1}$
Window	Hamming windows(25msec)
Frame period	10msec
Feature parameters (39 order)	12 ord. MFCC + Pow(1) + 12 ord. ΔMFCC + ΔPow(1) + 12 ord. ΔΔMFCC + ΔΔPow(1)
Cepstral Mean Normalization (CMN)	

2.2 발음사전

발음사전에는 화자들의 여러 가지 발성을 고려하여 한국어 음운규칙을 적용하였다. 사전에 포함된 목록은 음향모델과 언어모델 모두에 존재한다. 그리고 어휘는 언어모델을 작성하기 위한 텍스트 데이터에서 1 번 이상 출현하는 656 단어들로 구성하였다.

3. N-gram 언어모델

3.1 언어모델의 정의

일반적으로 언어모델은 하나의 문장에서 하나의 단어 w가 얼마나 빈번하게 발생하는가를 반영하는 것으로 문자열 w에 대한 확률 $P(w)$로서 공식화 할 수 있다. 언어모델은 이전의 확률 $P(w)$ 즉, 문장 W (혹은 단어열 $W=w_1,w_2,\Lambda,w_n$)의 발생확률을 구하는데 사용된다. 여기서 $P(w)$는 식(1)과 같이 나타낸다.

$$P(W)=P(w_1)P(w_2\mid w_1)P(w_3\mid w_1w_2)\,\Lambda\,P(w_n\mid w_1w_2,\Lambda,w_{n-1}) = \prod_{i=1}^{n}P(w_i\mid w_1,\Lambda,w_{i-1}) \quad (1)$$

그러나, 다양한 종류의 단어의 조합에 있어서 조건부 확률 $P(w_i\mid w_1,\Lambda,w_{i-1})$을 전부 구하는 것이 불가능하다. 단어수 L의 경우 $P(w_i\mid w_1,\Lambda,w_{i-1})$을 완전히 구해내기 위해서는 L^i 개의 단어조합이 필요하다. 또, 실제로 효과적인 모델이 생성되려면, 단어의 이전 정보 w_1,Λ,w_{i-1}을 같은 종류의 항으로 분할하고 모델의 파라미터의 수를 줄여줄 필요가 있다. N-gram 모델에서 어느 시점에서의 단어발생은 직전의 $N-1$ 단어에 의존하는 것으로 생각할 수 있으므로 N-gram 언어모델은 식(2)와 같이 나타낼 수 있다.

$$P(w_n\mid w_1\Lambda\,w_{n-1})=P(w_n\mid w_{n-N+1}\Lambda\,w_{n-1}) \quad (2)$$

N-gram 확률은 코퍼스 중에 출현하는 단어 N와 $N-1$개의 출현빈도수로부터 식(3)과 같이 추정될 수 있다.

$$P(w_n\mid w_{n-N+1}\Lambda\,w_{n-1})=\frac{C(w_{n-N+1}\Lambda\,w_n)}{C(w_{n-N+1}\Lambda\,w_{n-1})} \quad (3)$$

여기서, C는 단어열이 코퍼스 중에서 출현하는 회수를 나타낸다.

N-gram 언어모델의 학습에 있어서, 학습 데이터 중에 나타나지 않는 단어에 대해서는 확률 값을 0 으로 추정하여 학습을 중단하는 문제가 발생한다. 또, 만약 학습 데이터 중에 출현하여도, 출현빈도가 적은 단어에 대해서는 통계적으로 신뢰성있는 확률 값을 추정하는 것이 어렵다. 이러한 문제를 sparseness 라고 부르며, 이를 해결하기 위한 방법으로 discounting 과 back-off 의 조합을 이용한다[6]. Discounting 은 출현빈도가 많은 trigram 에

대해 그 빈도수를 줄여서 재추정하는 방법이고, back-off 는 출현빈도가 너무 적어서 추정하기 어려운 trigram 에 대해 scaled bigram 확률로 대체하여 추정하는 방법으로 다음과 같이 나타내어진다.

$$P'(w_k \mid w_{k-1}, w_{k-2}) = B(w_{k-1}, w_{k-2})P(w_k \mid w_{k-1}) \qquad (4)$$

여기서, B는 $P'(w_k \mid w_{k-1}, w_{k-2})$을 적당하게 정규화 하는 back-off 함수이다. 다른 높은 차수의 back-off N-gram 의 경우에도 위와 같은 형태로 적용한다.

3.2 반복학습법을 이용한 언어모델

일반적으로 언어모델의 확률 값을 추정하기 위해서는 많은 량의 텍스트 데이터가 필요하지만, 이런 텍스트 데이터를 수집하는 데에는 많은 시간과 경비가 소요된다. 이를 해결하기 위하여 소량의 텍스트 데이터로서 동일 문장을 반복 학습을 통하여 각 단어의 출현빈도수를 높여서 더 강건한 단어발생 확률을 가지도록 하는 방법을 제안한다. 이때 언어모델로서는 bigram 과 trigram 언어모델을 이용한다.

이 방법으로 단어 N-gram 언어모델을 작성하기 위해서 연속문장의 채록환경과 발성형태, 단어발생 확률 등을 고려하여 제한된 항공편 예약 200 문장을 1,154 문장으로 확장한 후 10 회에서 100 회까지 동일 문장을 반복하여 학습하였다. 학습 텍스트 데이터는 전처리 과정을 통해 문장의 시작(silS)와 끝(silE)를 제외한 쉼표와 여러가지 특수기호는 제외하였다.

또한 언어모델이 통계적으로 신뢰성을 확보할 수 있는 추정을 위해 출현빈도가 적은 단어에 대해 back-off 방법을 이용하여 학습하였다. back-off 계수의 계산을 위해서는 Witten-Bell discounting 방법[6]을 적용하였다. 기본 N-gram 의 입력에 대한 cutt-off 문턱차 값으로 bigram 의 경우 1, trigram 의 경우 2 를 각각 설정하였고 각 언어모델은 CMU-Cambridge toolkit[6]을 이용하여 작성하였다.

또한, 언어모델 작성의 경우, 단계적 인식 알고리즘을 고려하여 forward pass, backward pass 를 위해 텍스트 문장을 각각 전향 및 후향으로 나열하였다.

4. 단계적 인식 방법

인식방법으로서 디코딩 알고리즘은 관측된 음성에 대해서 가장 적절한 단어열을 찾는 것이다. 하지만 어휘수가 증가하고 태스크가 복잡해 짐에 따라 원하는 단어열을 구한다는 것은 쉬운 일이 아니다.

따라서 본 연구에서는 이를 효과적으로 해결하기 위해, forward pass 에는 단어 bigram 을, backward pass 에는 단어 trigram 을 이용하는 two pass (forward – backward) 인식 알고리즘[5]을 이용하였다.

Forward pass 에서는 프레임 동기형 빔 탐색 알고리즘을 이용하여 동적으로 목구조 형태의 사전에 bigram 확률을 지정하여 사전에 있는 모든 단어에 대해서 탐색을 수행한다. 이 결과 인식된 모든 후보 단어가 들어 있는 단어 격자 테이블이 만들어진다. 이 격자 테이블에는 단어의 구간과 경로 정보가 들어있다.

Backward pass 에서는 forward pass 에서의 많은 계산과 추정에 의한 음향적, 언어적 오류를 목구조 격자 탐색을 통해서 어느 정도 복구해 준다. 이 방법은 forward pass 에서 후보로 선정된 단어에 대해서만 탐색을 수행하므로 속도가 forward pass 에 비해 빠르다. 또한 forward pass 와 반대방향의 탐색을 통해 forward pass 에 만들어진 단어격자 테이블을 이용하여 각 단어의 시작점을 결정하고 여러 경로 중에서 하나의 경로를 선택한 뒤 그 후보 단어를 예측하고 스코어를 효율적으로 계산한 뒤 필요한 정보를 다른 단어 격자 테이블에 저장해 둔다. 이렇게 만들어진 단어격자 테이블은 best-first stack decoding (A* 알고리즘)을 수행하여 오류를 보정한다.

5. 인식실험 및 결과

제한된 텍스트 및 소량의 음성 데이터를 이용하여 효과적인 언어모델을 작성하기 위한 반복학습법의 유효성을 확인하기 위해, 한국과 일본 사이의 항공편 예약 태스크를 대상으로한 연속음성인식 실험을 수행하였다.

인식실험에 사용한 연속음성 및 텍스트 데이터는 13 인의 사용자가 가상의 조건에서 항공편 예약에 관련된 질문과 대답에 관한 내용으로 구성하고 있다. 13 인의 음성 데이터 중 평가용 데이터는 3 인의 200 문장으로 설정하고 나머지 10 명의 음성 데이터 중 발성 상태가 양호한 8 명의 연속음성을 단어와 함께 음향모델의 학습에 사용하였다.

그림 2, 3 에 단어 N-gram 언어모델을 이용한 인식실험 결과와 반복학습에 따라 작성한 언어모델의 복잡도 (perplexity) 변화를 각각 나타내었다.

기존의 문맥자유문법(CFG)과 1-pass Viterbi 빔 탐색법을 이용한 인식실험에서, 평가용 100 문장에서 하나의 문장에 포함된 평균 단어수는 약 5.1 개이고 복잡도는 43.2 이다. 화자적응화된 음향모델에 대해 3 인의 100 문장을 이용하여 기존의 방법[8]을 이용한 평가결과 평균 68.3%의 문장 인식률을 얻었다.

인식률 개선을 위해 단어 N-gram 언어모델을 이용하고, 제한된 소량의 텍스트 데이터를 이용한 반복학습 방법의 유효성을 검토하기 위해, 각각 10 회에서 100 회까지 반복 학습한 언어모델에 대해 언어모델의 복잡도와 forward pass 와 backward pass 에 의한 인식실험을 수행하였다.

평가용 3 인의 200 문장을 각 반복 회수에 따른 각 언어모델에 대해 평가한 결과, forward pass 의 경우 평균 84.1%, backward pass 의 경우 평균 91.6%의 문장 인식률을 얻어 backward pass 의 경우에서 forward pass 의 경우에 비해 약 7.5% 향상된 인식률을 얻었다. 이는 bigram 을 이용한 forward pass 에서 발생한 음향적 언어적 오류를 trigram 을 이용한 backward 에서 많이 복구해 주기 때문이라 생각된다. 또한, 그림 2 에서 나타낸 바와 같이 반복 회수가 증가함에 따라 backward pass 의 인식률의 변화는 많지 않지만, forward pass 의 인식률은 회수에 따라 증가하다가 일정하게 수렴하는 것을 볼 수 있다. 또

한, 그림 3으로부터 언어모델의 복잡도 변화도 반복 회수가 증가함에 따라 서서히 줄어들며 수렴함을 알 수 있다.

이상의 결과로부터 소량의 텍스트 데이터를 이용한 제한된 태스크에서 언어모델을 작성할 때 반복학습 방법이 유효함을 확인할 수 있었다. 그리고 기존의 사용한 CFG (unigram)보다 N-gram 언어모델을 이용하는 것이 더 효율적임을 알 수 있었다.

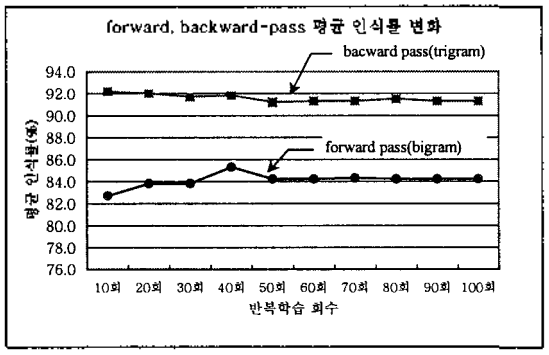

그림 2. 반복학습에 의해 작성한 언어모델을 이용한 연속음성인식 결과

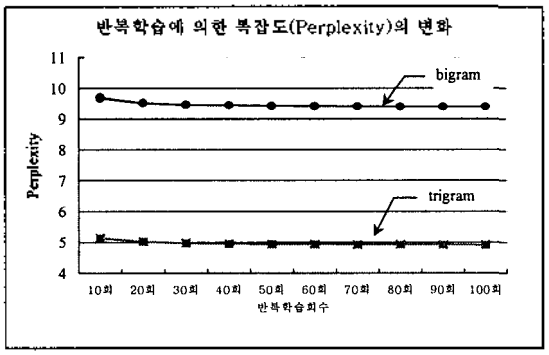

그림 3. 반복학습에 의한 언어모델의 복잡도

6. 결 론

본 연구에서는 연속음성인식 시스템의 성능 향상을 위하여 언어모델 작성에 있어서 제한된 항공편 예약 200 문장으로 연속음성에 적합한 언어모델을 작성하는 효율적인 방법으로 반복학습법을 제안하고, 이를 이용하여 성능 평가 실험한 결과 평균 91.6%의 인식률을 얻어 제안한 방법의 유효성을 확인하였다.

반복학습법을 이용한 언어모델 작성에 있어서는 제한된 항공편 예약 200 문장의 소량의 텍스트 데이터를 이용하여 보다 더 강건한 단어발생 확률을 나타내도록 하기 위해 다양한 상황을 고려한 1,154 문장으로 확장한 후 동일 문장을 반복 학습하여 작성하였다.

이렇게 작성한 언어모델을 이용하여 평가용 3 인의 200 문장을 각 반복 회수에 따라서 forward - backward pass 의 단계적 인식방법으로 인식실험을 수행한 결과, forward pass 의 경우 평균 84.1%, backward pass 의 경우 평균 91.6%의 문장 인식률을 얻었다. 또한, 반복회수가 증가함에 따라 backward pass 의 인식률의 변화는 없지만, forward pass 의 인식률은 회수에 따라 증가하다 일정값으로 수렴함을 알 수 있었고, 언어모델의 복잡도에서도 반복 회수가 증가함에 따라 복잡도가 서서히 줄어들며 수렴함을 확인할 수 있었다.

이상의 결과로부터 소량의 텍스트 데이터를 이용한 제한된 태스크에서 언어모델을 작성할 때 반복학습 방법이 유효함을 확인할 수 있었고 소량의 텍스트 데이터뿐만 아니라 대규모의 텍스트 데이터를 이용한 언어모델의 작성에도 효율적일 것으로 기대된다.

참 고 문 헌

[1] L.R. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition," Prentice Hall, 1993.

[2] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Japanese Large Vocabulary Continuous Speech Recognition Using a Business Newspaper Corpus," Proc. ICSLP 96, 22-25, 1996.

[3] K. Takagi, R. Oguro, K. Hashimoto, and K. Ozeki, "Performance Evaluation of Word Phrase and Noun Category Language Models for Broadcast News Speech Recognition," Proc. ICSLP 98, pp. 2507-2510, Dec. 1998.

[4] Jong Ryong Choi, Bum Koog Kim, Hyun Yeol Chung, and S. Nakagawa, " A Korean Flight Reservation System using Continuous Speech Recognition," The Journal of the Acoustical Society of Korea, Vol. 15, No. 3E, Sept. 1996.

[5] Long Nguyen, Richard Schwartz, "Efficient 2-Pass N-best Decoder," Proc. Of the DARPA Broadcast News Transcription and Understanding Workshop, 1997.

[6] P. Clarkson, R.Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proc. Eurospeech 97, pp. 2707-2710, Sept. 1997.

[7] S. Young, J. Jansen, J. Odell, D.Ollason, and P. Woodland, "The HTK Book," 1995.

[8] 오세진, 김범국, 정현열, "연속음성인식 시스템의 성능개선," 한국음향학회 하계 학술발표대회 논문집, 1997. 11.