2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

음성인식의 고속화를 위한 프레임 단위 적응 프루닝 알고리즘

°황 철 준*, 오 세 진, 김 범 국*, 정 호 열, 정 현 열
영남대학교 전자정보공학부
*대구과학대학 정보전자통신계열

A Frame Unit Based Adaptive Pruning Algorithm for the Fast Speech Recognition

°Cheol-Jun Hwang*, Se-Jin Oh, Bum-Koog Kim*, Ho-Youl Jung, Hyun-Yeol Chung
School of Electrical Eng. & Computer Science, Yeungnam University
*Informational Electronics & Communication Div., Taegu Science College

요 약

본 논문에서는 인식이 진행되는 동안 탐색 공간을 효과적으로 줄임으로써 음성인식의 고속화를 달성할 수 있는 새로운 프레임 단위 적응 프루닝 알고리즘을 제안하고 실험을 통하여 그 유효성을 확인하였다. 이것은 앞 프레임과 뒤 프레임 사이의 최대 확률은 높은 상관성을 가지므로 프루닝 문턱치를 앞 프레임의 최대 확률로부터 효과적으로 구할 수 있다는 사실에 근거를 두고있다. 이 방법에서는 앞 프레임의 최대 우도 확률과 후보 확률들의 조합으로 현재 프레임의 프루닝 문턱치를 갱신함으로써 현재 프레임의 문턱치를 인식 과정 중에 얻을 수 있기 때문에, 인식 태스크가 바뀌어도 문턱치를 구하기 위한 사전 실험을 수행할 필요가 없게 된다. 또한, 프레임 단위로 적응적으로 얻어진 문턱치는 다른 환경 하에서도 인식 속도의 향상을 가져올 수 있게 된다. 제안된 알고리즘의 유효성을 확인하여 위하여 한국어 주소 인식 시스템에 적용하였다. 본 시스템은 48개의 유사음소단위(PLUs)를 인식의 기본단위로 하고, 적응알고리즘으로는 최대사후확률추정법(MAP; Maximum A Posteriori Probability Estimation)을, 인식 알고리즘으로는 OPDP(One Pass Dynamic Programming)법을 이용하였다. 남성화자 3인이 25개의 연결 주소명을 대상으로 인식 실험을 수행한 결과, 제안된 프레임단위 적응프루닝 문턱치를 적용한 경우를 기존의 고정 프루닝 문턱치와 가변 프루닝 문턱치를 적용한 경우와 비교하였을 때 인식률의 변화없이 탐색공간이 상대적으로 각각 14.4%와 9.14%가 감소되어 제안된 프레임 단위 적응 프루닝 알고리즘의 유효성을 확인할 수 있었다.

1. 서 론

최근 음성인식 기술의 발전과 컴퓨터의 보급이 가속화됨에 따라 음성인식 기술의 실용화에 많은 관심이 증대되고 있다. 일반적으로 키보드, 마우스 그리고 기타 입력 장치를 이용한 기계나 컴퓨터와의 통신은 이들 입력장치의 사용상 불편함이 크고 성가신 면이 있다. 이와 같은 불편함을 해소하기 위한 유용한 수단의 하나는 음성인식을 입력장치로 이용하는 방법이다[1].

음성인식을 입력장치로 이용하기 위해서는 실시간 인식이 가능하여야 함과 동시에 인식률도 높아야 한다. 하지만 인식 속도와 인식률 사이의 관계는 서로 상반되는 점을 가진다. 즉, 복잡한 음향모델을 이용하고 인식대상 어휘를 선택하기 위한 넓은 탐색공간을 제공하면 비교적 높은 인식률을 달성할 수는 있으나 인식속도가 늦어져 실용화의 필수조건인 실시간 인식이 어려워진다. 반대로 프루닝방법과 같은 고속화 기법을 사용하여 탐색 공간을 제한하고 간단한 음향모델을 이용하면 인식속도는 개선되나 인식률이 낮아질 가능성이 크다. 이와 같이 인식률을 향상시키거나 유지하면서 인식 속도를 줄이는 것은 쉬운 일이 아니다.

이와 같은 점을 고려하여 본 논문에서는 인식시스템의 인식률 저하없이 인식 속도를 줄이는 방법에 대하여 검토하고자 한다. 음성인식 속도를 향상시키기 위해 많이 이용되는 대부분의 프루닝 알고리즘은 인식 과정에서 탐색 공간을 감소시키기 위해 정해진 문턱치를 적용한다. 그러나 이 경우 고정 문턱치를 사용하기 때문에 다양한 환경에서 발성된 음성의 인식에 적용하기 위해서는 인식 태스크가

바뀔 때마다 프루닝 문턱치를 얻기 위한 많은 사전 실험을 필요로 하게 된다.

저자들에 의한 사전 연구[2-3]에서는 전체 프레임으로부터 고정 프루닝 문턱치를 얻은 후 탐색의 진행에 따라 값이 변하는 가변 프루닝 문턱치를 적용하였다. 이 방법들은 각 프레임에서 후보 단어들을 어느 정도 효과적으로 제한할 수 있었지만, 여전히 탐색할 필요가 없는 많은 공간들을 탐색하는 문제가 있었다. 이러한 문제를 해결하기 위하여, 본 논문에서는 인식 과정 중에 탐색 공간을 효과적으로 줄이기 위한 새로운 프레임 단위 적응 프루닝 알고리즘을 제안한다.

2장에서는 빔 탐색과 프루닝에 대하여 설명하고, 3장에서 본 논문에서 제안된 프레임 단위 프루닝 문턱치 알고리즘을 설명한다. 4장에서는 본 논문에서 이용하는 한국어 주소 입력 시스템의 개요를 설명한 후, 5장에서 제안된 알고리즘을 적용한 실험 결과를 설명하고, 마지막으로 6장에서 결론을 맺는다.

2. 빔 탐색과 프루닝 기법

일반적으로 음성인식은 예측된 전체 후보와 입력 음성을 정합시키는 방법을 이용하는데, 대상 어휘수가 증가하고 인식 알고리즘이 복잡해짐에 따라 대규모 탐색 공간이 필요하며 이에 따라 많은 처리시간이 요구된다. 따라서 실시간 음성인식을 위해서는 전체의 후보와의 정합을 수행하지 않고서도 고정도의 인식 성능을 얻을 수 있는 효과적인 탐색 수법이 필요하다.

주어진 입력 음성에 대하여 우도가 가장 큰 후보를 탐색하는 확률적 음성인식법에서는 모든 가능한 후보를 고려할 경우 어휘수의 증가에 따라 탐색 공간이 지수 함수적으로 증가하게 되어 많은 계산량과 메모리가 필요하게 된다. 이를 해결하기 위해 제안된 빔 탐색법[4]은 각 프레임에서 몇 개의 부분 경로만을 추정하기 때문에 인식 대상 어휘수의 증가와 관계없이 일정한 수준 이하로 탐색 공간을 줄일 수 있다. 이하 이 방법에 대해 간략한다.

만약 i프레임에서 최대 우도를 가진 경로의 현재 시점이 (i, j^*)라면, 어떤 (i, j)에 대해서 그 경로는 다음 조건을 만족하면 프레임 $i+1$에서 후속하는 정합을 고려하게 된다.

$$P_{\max}(i,j) \le P_{\max}(i,j^*) + \lambda \qquad (1)$$

이때 중요한 것은 음향모델과 각 후보와의 우도의 정도이다. 빔 폭과 프루닝 문턱치를 엄격하게 적용하면 최종 프레임에서 정해가 될 수 있는 경로가 중간 과정의 경로 선택에서 프루닝 될 가능성이 있다. 따라서 최적 경로를 보증하기 위해서는 더 큰 값의 문턱치 조건을 주어야 한다. 그러나, 이렇게 하면 탐색공간이 넓어져서 더 많은 탐색시간이 필요하게 된다. 이 문제를 해결하기 위하여 탐색 공간을 더욱 유동적으로 제한하기 위하여 프레임 동기형 가변 프루닝 문턱치가 제안되었다. 이 방법에서 프레임 $i+1$에서 경로는 식 (2)에 의하여 프루닝 된다.

$$P_{\max}(i,j) \le P_{\max}(i,j^*) + \lambda(k) \qquad (2)$$

여기서, $\lambda(k)$는 프레임 동기형 가변 프루닝 문턱치이고, 탐색이 진행됨에 따라서 변화하게 된다. 즉, 일정 프레임이 지나면 일정 문턱치를 감소시킨 후 다음 일정 프레임에서는 실제 탐색 공간에 따라서 문턱치를 변화시켜 주면서 프루닝 시킨다. 이때 적용되는 문턱치는 시작 프레임에서는 상대적으로 느슨하고, 끝 프레임으로 갈수록 더 엄격해진다.

3. 프레임 단위 적응 프루닝 문턱치

위 방법들이 각 프레임에서 후보 단어들을 이전에 제안된 방법들에 비해 보다 효과적으로 제한할 수 있었지만, 여전히 탐색 할 필요가 없는 공간을 탐색한다. 따라서 본 논문에서는 인식 과정 중에 탐색 공간을 효과적이고 자동으로 줄이기 위하여 프레임 단위 적응 프루닝 알고리즘을 제안한다.

이 알고리즘은 이웃 프레임사이의 최대 우도 확률들의 상관성이 크므로 앞 프레임의 최대 우도 확률로부터 효과적인 프루닝 문턱치를 얻을 수 있다는 점에 착안하여, 앞 프레임의 최대 우도 확률과 후보 우도 확률들의 조합으로 현재 프레임에서의 프루닝 문턱치를 프레임 단위로 갱신하는 방법이다.

현재 프레임의 프루닝 문턱치는 식 (3)을 이용하여 계산되어진다.

$$\lambda(k) = \frac{1}{N}\sum_{s=1}^{N}\{P_{\max}(i-1,j^*) - P_{hyp}(i-1,s)\} \qquad (3)$$

여기서, $P_{max}(i-1,j^*)$는 프레임 $i-1$에서 최대 우도 확률이고, $P_{hyp}(i-1,s)$는 프레임 $i-1$에서 여러 후보들의 우도 확률이고, 그리고 N은 프레임 $i-1$에서 후보의 수이다.

식(3)으로부터 알 수 있는 바와 같이 제안된 알고리즘은 현재의 문턱치가 인식 과정 중에 얻어질 수

있기 때문에, 인식 태스크가 바뀌더라도 문턱치를 구하기 위하여 여러 번의 사전 실험을 필요로 하지 않는다. 또한, 문턱치가 적응적으로 얻어지기 때문에 다른 환경 하에서도 인식 속도를 향상시킬 수 있다. 그림 1에 제안된 프레임단위 적응프루닝 문턱치를 고정 프루닝과 가변 프루닝 문턱치와 비교하여 나타내었다.

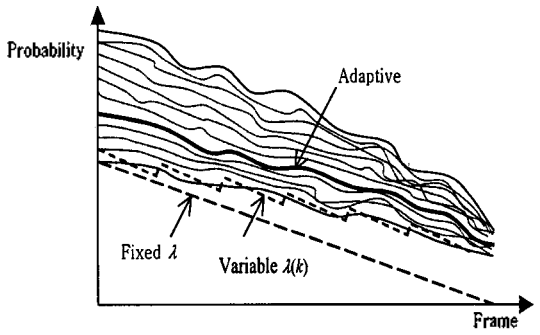

그림 1. 적응 프루닝 문턱치
(고정 프루닝과 가변 프루닝 문턱치와 비교)

4. 한국어 주소 인식 시스템과 음성 데이터

제안된 방법의 유효성을 확인하기 위하여, 본 연구실에서 개발한 한국어 주소 인식 시스템에 적용하였다. 본 시스템은 사운드 카드와 마우스를 가진 개인용 컴퓨터에서 동작한다. 사용자는 키보드나 마우스 뿐 아니라 음성을 이용하여 주소를 입력할 수 있다. 본 시스템은 48개의 4상태 3출력 확률 분포를 가진 연속 HMM 유사음소단위(PLUs)를 인식의 기본 단위로 하고, 사용환경 변화에 의한 인식 성능 저하를 최소화하기 위해 최대사후확률추정법(MAP)을 사용하고, 인식 알고리즘으로는 OPDP법을 이용한다 [5].

음성 데이터로는 화자독립 기본모델 (Speaker Independent HMM; SI-HMM)의 작성을 위하여 한국전자통신연구원(ETRI)에서 작성한 PBW(Phoneme Balanced Words) 445단어 음성 데이터베이스 중 14인의 1회 발성을 이용한다. 적응화 단계에 있어서는 사무실 환경에서 3인의 남성화자가 데스크탑 마이크를 이용하여 발성한 100개의 연결주소단어 중 25개의 연결주소단어를 이용하여 SI-HMM을 적응화한다. 인식단계에서는 나머지 75개의 연결주소단어를 사용한다. 상위 클래스에서 하위 클래스로 단계적으로 인식이 진행되는 한국어 주소의 특징을 고려한 유한상태 오토마타를 구성하였고, 전국의 모든 행정 단위를 포함하고 있다.

모든 음성 데이터는 16 kHz, 16 bits로 양자화 되었고, 각 프레임은 16 msec(256 samples)의 해밍 윈도우를 곱하여 5 msec씩 이동하면서 분석하였다. 이렇게 얻은 음성 샘플로부터 auto-correlation 방법으로 14차 LPC 계수를 추출하였으며, 추출한 LPC 계수로부터 10차의 MFCC를 구하여 특징 파라미터로 사용하였다. 그리고 10차의 회귀 계수를 동적 특징으로 사용하였다[6]. 표 1은 음성 데이터와 시스템의 환경을 나타낸다.

표 1. 학습, 적응화, 인식용 음성데이터

화자(수)	남성(14)	남성(3)	남성(3)
발성형태(수)	PBW(445)	연결단어(25)	연결단어(75)
발성회수	1	1	1
사용단계	학습	적응화	인식
환경	방음부스	사무실	사무실
녹음장치	DAT	PC	PC
마이크	헤드셋	데스크탑	데스크탑

5. 인식 실험 및 결과

인식실험은 연결단어로 이루어진 한국어 주소 인식을 위하여 적은 메모리 공간과 탐색 시간을 필요로 하는 OPDP 알고리즘[7]을 이용하여 수행되었다. 초기 음향학적 모델은 14인의 남성 화자가 발성한 445 단어를 이용하여 학습한 연속 혼합 HMM을 사용한다. 환경 변화에 의한 성능 저하에 강건하고 실용화에서 만족할 만한 성능을 얻기 위하여, 초기 HMM을 MAP 추정법으로 재학습하였다. 적응화는 각 화자에 대하여 25개의 연결 단어를 사용하였다. 10차의 MFCC와 10차의 RGC를 모든 실험에서 특징 파라미터로 사용하였다.

제안된 프루닝 문턱치 알고리즘의 유효성을 확인하기 위하여 3인의 화자가 발성한 75개의 연결단어에 대하여 탐색 공간을 측정하였다. 실험은 워크스테이션(167MHz)상에서 off-line으로 수행되었다. 주소에 대하여 인식률(CWRR; Connected Word Recognition Rate)과 탐색 공간(SS; Search Space)을 측정하였고, 주소를 이루는 각 단어에 대하여 단어 인식률(WRR; Word Recognition Rate)을 3인의 화자에 대하여 구하였다. 탐색 공간을 측정한 이유는 일반적으로 컴퓨터 성능에 따라 같은 인식 태스크라 하더라도 인식 시간이 달라질 수 있다. 따라서 신뢰성 있는 결과를 얻기 위하여 탐색되는 전체 단어와 후보 단어를 고려하여 식 (4)를 이용하여 탐색 공간을 측정하였다.

$$SS(\%) = \frac{N_{match}}{N_{match} + N_{skip}} \times 100.0 \quad (4)$$

여기서, N_{match}는 각 프레임에서 탐색되는 단어의 평균수이고, N_{skip}은 각 프레임에서 탐색되지 않는 단어의 평균수를 나타낸다.

표 2는 빔 폭이 10인 경우 고정 프루닝 문턱치, 가변 프루닝 문턱치, 그리고 제안된 프레임 단위 적응 프루닝 문턱치를 적용한 경우의 인식실험 결과를 나타낸다.

표 2. 인식 실험 결과
(고정, 가변, 적응 프루닝 문턱치)

고정 프루닝 문턱치			
Pruning threshold	CWRR(%)	WRR(%)	SS(%)
-500	96.0	98.7	30.96
-400	96.0	98.7	30.63
-300	95.7	98.5	30.46
가변 프루닝 문턱치			
Pruning threshold	CWRR(%)	WRR(%)	SS(%)
FSV1	96.0	98.7	29.17
FSV2	96.0	98.7	29.11
FSV3	96.0	98.7	29.09
FSV4	96.0	98.7	28.99
FSV5	96.0	98.7	28.87
FSV6	95.7	98.5	28.80
적응 프루닝 문턱치			
Pruning threshold	CWRR(%)	WRR(%)	SS(%)
Adaptive	96.0	98.7	26.23

단, FSV1: -400, -370, -350, -320, -290, -260
FSV2: -400, -380, -360, -330, -300, -230
FSV3: -400, -380, -360, -300, -300, -190
FSV4: -360, -340, -320, -300, -300, -10
FSV5: -310, -310, -300, -290, -270, -10
FSV6: -310, -310, -300, -290, -270, -10

3인의 화자를 대상으로 한 인식실험 결과, 고정 프루닝과 가변 프루닝 문턱치를 이용한 경우, 최고의 인식률은 연결단어에 대해서는 96.0%, 개별단어에 대해서는 98.7%로 나타났다. 이 인식률을 탐색공간을 추정하는 기준으로 정하였다. 탐색공간의 경우, 고정 프루닝 문턱치에 대해서는 최저 30.46%, 가변 프루닝 문턱치에 대해서는 최저 28.80%를 보여 가변 프루닝 문턱치를 이용한 경우 고정 프루닝 문턱치보다 1.66%의 탐색공간이 줄어드는 것을 확인하였다. 본 논문에서 제안한 프레임 단위 적응 프루닝 문턱치를 적용한 경우, 3인의 화자 평균 26.23%를 보여 가변 프루닝 문턱치보다 2.64%의 탐색공간이 줄어든 것을 알 수 있었다.

6. 결 론

본 논문에서는 인식이 수행되는 동안 탐색공간을 줄이기 위해 프레임 단위 적응 프루닝 문턱치 알고리즘을 제안하였다. 이 알고리즘은 이웃 프레임사이의 최대 확률의 상관성이 큰 점에 착안하여, 앞 프레임의 최대 확률로부터 효과적으로 프루닝 문턱치를 얻는 방법으로 현재 프레임에서 적응 프루닝 문턱치는 앞 프레임의 최대 확률과 후보 확률의 조합으로 결정할 수 있다.

제안된 방법의 유효성을 확인하기 위하여 한국어 주소 인식 시스템에 적용한 후 인식실험을 수행한 결과, 제안된 알고리즘이 고정 프루닝과 가변 프루닝 문턱치에 비하여 인식률의 저하없이 14.4%와 9.14%의 탐색 공간을 상대적으로 줄일 수 있음을 확인할 수 있어 제안된 방법의 유효성을 확인할 수 있었다.

향후 프레임 단위 적응 프루닝 알고리즘을 다양한 테스크로 확장하여 제안한 알고리즘의 유효성을 확인하고, 실용화 시스템에 적용하고자 한다.

참 고 문 헌

[1] M. K. Ravishankar, "Efficient Algorithms for Speech Recognition," Ph.D Thesis, Carnegie Mellon University, 1996.

[2] H. Y. Chung, C. J. Hwang, and S. W. Lee, "A Bimodal Korean Address Entry/Retrieval System," Proceedings fo ICSLP'98, Sydney, Australia, 1998.

[3] 황철준, 오세진, 김범국, 정호열, 정현열, "실시간 주소 음성인식을 위한 인식 시스템의 인식속도 개선," 1999년도 한국음향학회 하계학술 발표대회 논문집, 1999. 7.

[4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan Publishing Company, 1993.

[5] J. C. Junqua, and J. P. Haton, "Robustness in Automatic Speech Recognition," Kluwer Academic Publishers, 1996.

[6] X. D. Huang, Y. Ariki, and M. A. Jack, "Hidden Markov Models for Speech Recognition," Edinburgh University, 1990.

[7] S. Nakagawa, "Speech Recognition Based on Stochastic Model," IEICE Press, Japan, 1998.