실시간 화자독립 음성인식을 위한 고속 확률계산

정회원 박 동 철*, 안 주 원**

Fast computation of Observation Probability for Speaker-Independent Real-Time Speech Recognition

Dong-Chul Park*, Ju-Won Ahn** ***Regular Members***

요 약

H/W에 구현되는 음성인식 시스템에서 인식속도의 향상을 위한 새로운 알고리즘이 본 논문에서 제안되었다. 제안된 고속 관측확률 계산(Fast Computation of Observation Probability : FCOP) 알고리즘은 관측확률식을 근사화시키는 방법으로, CDHMM에서 상태(state)로 주어지는 확률분포함수들 중에서 일부를 효과적으로 제거하여 계산량을 최소화시키는 방법이다. 실제 H/W 환경의 음성인식에 응용한 실험 결과, 기존의 방법에 비해 인식률의 저하를 최소로 유지하며, 명령어 사이클을 20%~32% 감소시킬 수 있었으며, 인식속도를 약 30%향상시킬 수 있었다. 제안된 알고리즘을 제한된 자원을 가지는 실제의 휴대폰에 탑재하여, 인식속도 및 인식률을 측정한 결과 인식률의 저하를 0.2% 이하로 유지하면서, 인식속도를 30% 이상 증가시킬 수 있었다.

Key Words : Speech Recognition,CDHMM, Observation Probability, PDF

ABSTRACT

An efficient method for calculation of observation probability in CDHMM(Continous Density Hidden Markov Model) is proposed in this paper. the proposed algorithm, called FCOP(Fast Computation of Observation Probability), approximate observation probabilities in CDHMM by eliminating insignificant PDFs(Probability Density Functions) and reduces the computational load. When applied to a speech recognition system, the propsed FCOP algorithm can reduce the instruction cycles by 20%-30% and can also increase the recognition speed about 30% while minimizing the loss in its recognition rate. When implemented on a practical cellular phone, the FCOP algorithm can increase its recognition speed about 30% while suffering 0.2% loss in recognition rate.

Ⅰ. 서 론

가전제품, 통신용 단말기, 자동차, 완구 등에 쉽게 음성인식 기술을 적용할 수 있을 만큼 임베디드 음성인식 기술이 점차 개발되면서[1-3], 활발한 움직임을 보이고 있는 음성인식 시장에서는 성능에 대한 영향을 최소화하며, 메모리 사용량을 최소화시키는 기술에 대한 연구가 진행되어 왔다. 특히, 휴대 전화기에서 음성인식의 가장 중요한 요소는 프로그램 메모리 크기와 인식속도이다. 하지만 휴대 전화기 H/W 구성상 대부분의 메모리와 연산에 관련된 자원 중 음성인식에 사용될 수 있는 부분은 매우 제한적으로, 보통 대부분의 휴대 전화기 H/W에서 메모리 자원 중 다이얼링 등 무선 통신기능 등의 주요기능에 사용되고 남는 약 100Kbyte 미만만이 음성인식기능에 배정되는 매우 열악한 상태이며, CPU 구동속도 또한 매우 제한적이다. 이러한 이유로 많은 메모리를 필요로 하고, 인식에 요구되지만 연산

*, ** 명지대학교 정보공학과 지능컴퓨팅 연구실
논문번호 : KICS2005-03-119, 접수일자 : 2005년 3월 24일

량이 많은 일반적인 음성인식 알고리즘은 휴대폰용 음성 인식기에의 적용은 제한적일 수밖에 없다.

최초의 휴대폰용 음성인식기는 H/W상의 제약을 만족시키기 위하여 DTW(Dynamic Time Warping)를 이용한 화자종속형태에 의한 명령어 인식이나 제한적 이름인식을 수행하였다[4]. 이 방법은 소수 인식에서 높은 인식률을 보이나 인식 대상이 늘어날 때마다 메모리를 추가로 요구하기 때문에, 인식 대상의 수를 소수로 제안할 수밖에 없는 단점을 가지고 있고, 화자가 추가 될 때마다 매 번 학습 과정을 거쳐야 한다는 번거로움을 가지고 있다. 한편, 화자독립 음소모델은 새롭게 추가되는 인식대상 또는 화자에 따라 메모리의 추가나 이에 따른 또 다른 모델이 별도로 필요 없으며, 추가의 학습이 필요하지 않기 때문에 인식 대상의 변화가 많은 가변어 인식에 널리 사용된다[5]. 그러나, 연속밀도 은닉 Markov 모델(Continous Density HMM: CDHMM)을 사용하는 화자독립 음성인식은 DTW 나 Discrete HMM (DHMM) 보다 인식률이 월등히 좋지만 인식속도 면에서는 DTW나 DHMM과 비해 더 느리다[4]. 이 문제를 해결하기 위해서 접근 방법은 크게 두 가지로 나뉜다. 하나는 탐색 공간(Search Space)을 줄이는 방법이고 두 번째는 CDHMM의 관측확률(Observation Probability)의 연산에 소요되는 연산량을 줄이는 방법이다.

본 논문에서는 위에서 언급한 화자독립 음소모델을 사용하고, CDHMM을 이용한 음성 인식기를 실험 환경의 Baseline으로 사용 하였다. 탐색공간을 줄이기 위해 Tree-Structured VQ(TSVQ)와 Pruning Beam Search를 동시에 사용하여 향상된 결과를 얻을 수 있음을 먼저 보이고[6][7], CDHMM의 observation 확률을 계산하는데 소요되는 연산량을 줄이기 위해 사용된 Gaussian Mixture Maximum Approximation 방법과 함께 사용할 수 있는 새로운 방법을 제안하였다[8].

본 논문의 구성은 2장에서 고속탐색을 위한 알고리즘인 TSVQ와 Pruning Beam Search를 간단하게 요약하고, 3장에서는 CDHMM의 관측 확률을 계산하는데 필요한 연산량을 감소시키기 위해 사용한 알고리즘인 Gaussian Mixture Maximum Approximation 방법과 이를 개선하는 방법을 제안한다. 한편, 4장과 5장에서는 본 논문에서 제안된 알고리즘을 실험적으로 증명하기위한 실제적 실험 환경과 실험 결과를 보이며, 본 논문의 결론이 5장에서 주어진다.

Ⅱ. 탐색 알고리즘

음성인식에서 사용되는 TSVQ와 Pruning Beam Search 방법은 탐색공간을 줄이기 위해 사용되는데, TSVQ는 인식 계산을 위한 추가 모델에 대한 메모리가 증가되는 단점이 있지만, 인식률의 낮은 감소율과 속도 향상에 탁월한 성능을 보인다[6][7].

2.1 Tree-Structured VQ

이 알고리즘은 혼합체(Mixture)를 구성하는 모든 확률분포함수들에 대한 확률 계산량을 줄이기 위해 사용되는데, 확률분포함수들은 그림 1의 예와 같이 구성된 구조를 따라 계산된다. 그림 1에서 k(1,1)부터 $k(1,J_1)$ 은 전 음소모델의 모든 상태(state)를 구성하는 혼합체(Mixture)들을 J_1개의 군(set)으로 재분류 한 것이며, 이들을 본 논문에선 부모모델(Parents Model)이라 부른다.

음성 특징(feature)은 먼저 윗 단의 모든 부모모델들에 대한 Gaussian 확률분포를 구한다. 그 중에 큰 순서대로 정해놓은 수만큼의 확률분포함수 군(set)만을 선택하고 그 군들에 속해있는 확률분포함수의 혼합체에 대해서만 연산을 수행하는 알고리즘이다. 나머지 확률분포함수에 대해서는 자신을 포함하고 있는 부모모델의 확률값으로 대치하기 때문에, 제외된 확률분포함수의 수만큼의 연산 이득을 볼 수 있게 한다. 본 논문에서는 그림 1의 2층 구조를 사용한다.

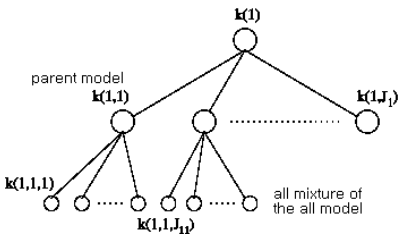

그림 1. 나무 구조의 확률분포함수

2.2 Pruning Beam Search

가지치기(pruning) 알고리즘의 주요 개념은 미리 정해진 임계값 이하의 확률을 가지는 음소모델들은 탐색대상에서 제외시키고, 임계값 이상의 확률을 가지는 음소 모델들만 탐색대상에 포함시켜 활성화시켜, 연산량을 줄이는 알고리즘이다. 그 임계값은 아래의 식을 따른다.

$$P_t = P_b - P_p \quad (1)$$

여기서 P_t는 임계값 확률이고, P_b는 활성화 되어있는 모든 어휘목록 중 최적상태(best state) 확률이고, P_p는 가지치기에 쓰이는 beam의 폭이다.

그림 2는 시간(frame)에 따른 문턱값을 넘어서는 어휘 목록의 숫자를 보이고 있다.

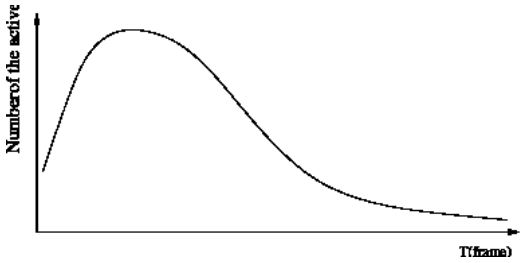

그림 2. 시간(frame)에 따른 활성화된 음소모델의 수

Ⅲ. 관측확률 계산량 절감방법

음성인식 시스템에서 가장 많은 연산량을 필요로 하는 것 중의 하나가 CDHMM에서의 관측 (observation) 확률을 계산하는 부분이다. 이 부분의 연산량을 줄이기 위해 새로운 근사화 알고리즘이 본 논문에서 제안된다. 기존의 Gaussian 혼합밀도 최대 근사화 (Mixture Density Maximum Approximation: MDMA) 방법은 상태 내에 있는 모든 Gaussian 혼합밀도들을 합하는 방법 대신, 각 상태의 최대확률를 갖는 혼합체의 값으로 상태내의 모든 혼합체의 확률을 근사화하는 것이다. 주어진 상태 λ에서 관측 O에 관한 확률분포는 다음으로 주어진다[4].

$$P(O|\lambda)=\sum_{}^{M}\left(\frac{\omega}{\sqrt{2\pi\sigma^2}}\exp\left(-\sum_{}^{F}\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \quad (2)$$

여기서 F는 특징벡터의 차원, x는 F 차원을 가진 관측특징벡터, w는 Gaussian 혼합체 가중치(mixture weight), μ는 평균, 그리고 σ는 표준편차이다. 한편, 식 (2)는 식(3)으로 변형될 수 있다.

$$P(O|\lambda)=\max_{M}\left(\frac{\omega}{\sqrt{2\pi\sigma^2}}\exp\left(-\sum_{}^{F}\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \quad (3)$$

그런데, 이것은 최대값을 찾기 위해서 각 혼합체에 대한 연산을 모두 수행해야 하기 때문에 계산상의 이득을 기대할 수 없다. 물론 log 확률을 합산하기 위해서 log와 exp가 들어가지만, 이것도 table로 처리하게 되면 없어지므로, 단순히 합산을 하지 않는다는 의미가 있을 뿐이다. 즉, 식 (3)에서의 연산에서는 하나의 혼합밀도를 구하기 위해, 모든 특징차원의 확률밀도함수들에 관한 연산을 수행한 후 합을 구해야 한다.

이러한 $P(O|\lambda)$ 연산상의 복잡성을 최소화하기위해, 본 논문에서는 이 확률밀도함수들 중 선택된 몇 개의 확률밀도함수들에 대해서만 연산하고, 그들로 혼합밀도를 대신하는 것이다. 이를 위해 식 (3)을 log 확률분포함수를 사용하여 다음으로 표현할 수 있다.

$$P(O|\lambda)=\max_{M}\left(\log\omega-\frac{1}{2}\left(\log 2\pi\sigma^2+\sum_{}^{F}\frac{(x-\mu)^2}{\sigma^2}\right)\right) \quad (4)$$

식(4)에서 $\frac{(x-\mu)^2}{\sigma^2}$ 부분은 변수 x에 대한 2차 함수이고 이것은 그림 3처럼 표현될 수 있다. 한편, 그림 3에서 정해진 임계값 (threshold:TH)의 아래 값들은 x가 변함에 따라 값의 변화가 상대적으로 작아 상수로 근사화 할 수 있으며, 또한 그 값은 매우 작기 때문에 합산에 영향을 거의 미치지 않는다. 따라서 최대 혼합밀도를 선택함에 있어서, 임계값 아래 값을 가지는 확률밀도를 제외한 나머지 것들의 합에 의한 계산으로도 값만으로도 무리가 없다.

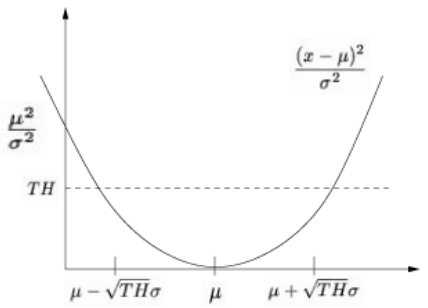

그림 3. log 확률분포함수

즉, 식 (4)에서의 변수는 오직 관측 특징벡터 x 뿐이며, 나머지 것들은 각 확률밀도함수의 고정된 상수 값이다. 결과적으로 식(4)는 식(5)처럼 표현될 수 있다.

$$P(O|\lambda)=\max_{M}\left(W-\frac{1}{2}\left(Const+\sum_{}^{F}\frac{(x-\mu)^2}{\sigma^2}\right)\right) \quad (5)$$

여기서 F'는 식 (2)~(4)에서 쓰인 특징벡터의 차원수 중에서 임계값 th보다 작은 확률밀도함수에 해당되는 것을 제외한 특징벡터의 차원수이다. 즉, $F' \le F$이다.

그림 3의 임계값은 아래처럼 유도된다.

$$\frac{(x-\mu)^2}{\delta^2} < TH$$

따라서,

$$\mu - \sqrt{TH}\delta < x < \mu + \sqrt{TH}\delta \qquad (6)$$

여기에서 μ와 σ는 미리 알려진 고정된 상수값이므로, 수식(6)의 양 끝값을 table화 시킬 수 있고. 그 조건은 결국 다음으로 정리될 수 있다.

$$LB < x < UB \qquad (7)$$

여기서, LB= $\mu - \sqrt{TH}\sigma$이고, UB= $\mu + \sqrt{TH}\sigma$이다.

위에서 구해진 임계값에 의한 근사화를 이용할 때, 이 조건 속에 속하는 입력특징은 연산에서 제외하고 물론 합산도 하지 않는다. 그리고, 그 나머지만이 Gaussian 혼합밀도를 구하기 위해 사용된다. 이 경우, 이 조건 안에 드는 확률밀도함수 만큼에 해당하는 연산을 생략할 수 있으므로, 연산량의 절감을 가져올 수 있다. 임계값을 높이면 연산량의 절감을 통한 속도 향상을 이룰 수 있으나, 또한 그만큼의 인식률 저하를 초래할 수 있다. 임계값은 인식률을 해치지 않으면서, 최대의 인식속도를 보장하기 위한 방향으로 결정된다.

Ⅳ. 실험 환경

실험에 사용된 데이터는 한국남성 300명에 의해 발음된 이름 60,000개가 사용되었다. 발음된 각 이름 데이터는 승용차량 내부에서 휴대용 단말기를 통해 녹음된 음성을 8K 16Bit Mono PCM으로 변환시켰다. 또한, 녹음된 음성의 신호대 잡음비(SNR)는 10-20db 이다. 특징추출을 위해 Pre-Emphasis, Rectangular Window가 사용되었으며, Pre-Emphasis의 필터계수는 0.937이며, 프레임 크기는 32msec, 16msec overlap을 사용하였다. 각 프레임당 특징벡터는 12차 LPC Cepstral 계수와 정규화된 에너지, 그리고 그들 각 각의 변화값(Delta)이 사용되었다. 41개의 음소 모델링을 위해 각 음소당 상태수는 2-3, 혼합체의 수는 6-10개가 사용되었다. 학습에 사용된 데이터는 한국남성 250명이 이름 200개를 1번씩 발음한 50,000개의 데이터가 사용되었으며, 이를 제외한 나머지 10,000개의 데이터는 성능 검증에 사용되었다. 검증은 현대 ARM Board(Model name GMS320C27)에서 수행되었다.

Ⅴ. 실험 결과

인식률과 인식속도에 미치는 각 알고리즘의 영향을 조사하기 위해 제시된 데이터에 대해 PC에서 인식률을 측정했고, 인식속도는 인식을 위해 사용되는 명령 사이클 (instruction cycles)의 수를 사용하였으며, 인식속도를 표현하는데 있어서 계산에 필요한 연산자의 수도 나타내었다.

표 1은 실험에 사용된 알고리즘을 요약한 것이며, 그림 4에서 막대그래프는 각 알고리즘의 인식률을, 선 그래프는 계산에 필요한 명령 사이클의 수를 표현한다. 알고리즘을 추가함에 따라 감소하는 명령 사이클의 수를 볼 수 있다. 명령 사이클의 수는 ARM debugger v2.5를 사용하여 측정되었다. 기존의 TSVQ, Pruning Beam Search, Gaussian Mixture Maximum Approximation (GMMA) 등의 방법을 조합해서 사용한 디코딩 방법인 알고리즘 D는 기존의 알고리즘 A에 비해 사이클은 약 23.2%가량 감소됐다. 또한, 제안된 FCOP 알고리즘을 추가함으로써, 사이클은 12.9%가량 더 감소됐다. 물론, 속도 향상은 약간의 인식률 감소를 초래한다는 사실도 또한 알 수 있다. 그러나, 인식률의 차이는 적용된 모든 알고리즘에서 0.2%이하이다. 한편, 표 2는 각

표 1. 실험에 사용된 알고리즘

알고리즘	알고리즘 내용
A	Baseline + Table(logadd)
B	Baseline + TSVQ
C	Baseline + Table + TSVQ
D	Baseline + TSVQ + GMMA
E	D + Reduced PDF Computation

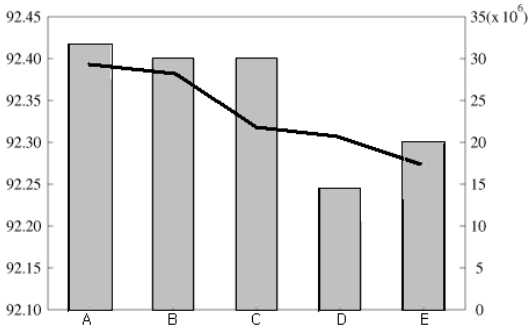

그림 4. 각 알고리즘의 인식률과 instruction cycle 비교

표 2. 각 알고리즘의 인식률과 산술적 계산 수

알고리즘	인식률 (%)	곱하기	나누기	더하기	빼기
a	92.42	1136139	557351	630113	600224
b	92.41	779978	378420	455718	423147
c	92.41	779978	378420	455718	423147
d	92.23	777757	377346	398863	400410
e	92.31	412050	194768	215734	217281

알고리즘의 인식률과 구체적인 산술적 계산량을 보인다.

위의 알고리즘 중 기존의 알고리즘인 a와 c에 대해 제안된 알고리즘 e가 실제 휴대전화기 상에서는 어떠한 성능을 보이는 가를 알아보기 위한 실험이 수행되었는데, 그 결과는 표 3과 같다. 표 3에서 알 수 있듯이, 제안된 알고리즘은 최소한의 인식률 감소를 통해 인식시간을 기존의 알고리즘에 비해 최대 약 40% 감소시킬 수 있음을 알 수 있다. 이는 인식속도 면에서는 1.68배의 속도 향상을 가져오는 것으로, 실시간 화자독립 음성인식기의 구현에 매우 유용한 도구로 제안된 알고리즘이 사용될 수 있다는 것을 보인다.

표 3. 실제 휴대 전화기 상의 인식률과 속도

알고리즘	인식률	인식시간(sec)
a	92.58%	0.32
c	92.54%	0.22
e	92.21%	0.19

Ⅵ. 결 론

음성인식 시스템을 H/W에 구현할 때 필연적으로 요구되는 인식속도의 향상을 위해, 고속 관측확률 계산(Fast Computation of Observation Probability : FCOP) 알고리즘이 제안되었다. 제안된 FCOP 알고리즘은 관측확률식을 근사화 시키는 방법으로, CDHMM의 상태(state)로 주어지는 확률분포함수들 중에서 전체 계산 결과에 제한된 영향을 미치는 것들을 효과적으로 제거하여 계산량을 최소화시키는 방법이다. 제한된 자원을 가지는 임베디드용 음성인식에서 인식성능 유지와 더불어 CDHMM모델의 계산량을 최소화시키는 것에 초점을 두는 FCOP 알고리즘은 한국어 이름 인식 문제에 적용되어, 인식성능과 명영 사이클 면에서 기존의 알고리즘들과 비교되었다. TSVQ와 가지치기 Beam 탐색의 두 방법을 함께 사용하여 탐색공간을 줄임으로써 인식률에 거의 영향을 미치지 않고 인식에 드는 명령어 사이클을 20%이상 절감시킬 수 있었으며, 제안된 FCOP를 적용한 결과, 명령어 사이클을 다시 약 12.9% 더 감소시킬 수 있었다. 물론, 이 경우 약 0.2%의 인식률 저하를 가져오지만, 휴대폰과 같은 제한된 H/W 환경에서 괄목할만한 인식속도의 향상을 위해서는 무시할 수 있는 수준이다. 한편, 제안된 알고리즘을 실제의 휴대폰에 탑재하여, 인식속도 및 인식률을 측정한 결과 인식률의 저하를 최소화하면서, 인식속도를 68% 이상 증가시킬 수 있었다. 따라서, 제안된 FCOP는 임베디드용 음성인식에서 필요한 연산량을 대폭 감소시키며, 그로 인한 인식시간을 기존의 알고리즘에 비해 약 40% 줄일 수 있어, 실시간 음성인식기의 구현에 매우 유용한 도구로 사용될 수 있다.

참 고 문 헌

[1] S. Phadke *et. al.* “On design and implementation of an embedded automatic speech recognition system,” *Proc. of Int. Conf. on VLSI Design 2003*, pp. 127-132, 2004.

[2] F. Elmisery *et. al.* “A FPGA-based Viterbi algorithm implementation for speech recognition system,” *Proc. of ICASSP-01,* pp. 1217-1200, 2001.

[3] S. Melnikoff and S. Quigley, “Implementing log-add algorithm in hardware,” *Electronics Letters,* V. 39, No. 12, pp. 939-940, 2003.

[4] L. R. Rabiner, B. H. Juang. *Fundamentals of speech recognition*. Prentice-Hall Inc., 1993

[5] K. Shinoda. and K. Iso, “Efficient reduction of gaussian components using MDL criterion for HMM-based speech recognition,” *Proc. ICASSP-02,,* pp 869-872, 2002.

[6] T. Watanabe *et. al.* “High speed speech recognition using tree structured probability density function,” *Proc. ICASSP-95,* vol.1, pp 556-559, 1995.

[7] S. Renals, “Phone deactivation pruning in large vocabulary continuous speech recognition,” *IEEE Signal Processing Letters*, vol.

3, no. 1, 1996.

[8] S. Ortmanns *et. al.* "An efficient decoding method for real time speech recognition," *Proc. of ESCA, Eurospeech99*, pp.499-502, 1999

박 동 철 (Dong-Chul Park) 정회원

1980년 2월 서강대학교 전자공학과(공학사)

1982년 2월 한국과학기술원 전기 및 전자공학과(공학석사)

1990년 6월 Univ. of Wahsington(Seattle), Electrical Engineering (Ph.D.)

1990년 8월~1994년 2월 조교수, Florida Int'l Univ. Dept. of Eelct. and Comp. Eng.

1994년 3월~현재 명지대학교 정보공학과 교수

1997년~2000년 IEEE Tr. on Neural Networks, Associate Editor

1999년~현재 IEEE Senior Member

<관심분야> 신경망 알고리즘 개발, 음성인식, 신경망의 금융공학에의 응용

안 주 원 (Ju-Won Ahn) 정회원

1999년 2월 명지대학교 전기전자공학부(학사)

2002년 2월 명지대학교 전기전자공학부(석사)

현재 LG전자 주임연구원

<관심분야> 음성인식, 통신 시스템