Diphone 단위의 hidden Markov model을 이용한 한국어 단어인식

Korean Word Recognition Using Diphone-Level Hidden Markov Model

박 현 상*, 은 종 관*, 박 용 규*, 권 오 욱*
(Hyun Sang Park*, Chong Kwan Un*, Yong Kyu Park*, Oh Wook Kwon*)

ABSTRACT

In this paper, speech units appropriate for recognition of Korean language have been studied. For better speech recognition, co-articulatory effects within an utterance should be considered in the selection of a recognition unit. One way to model such effects is to use larger units of speech. It has been found that diphone is a good recognition unit because it can model transitional regions explicitly. When diphone is used, stationary phoneme models may be inserted between diphones.

Computer simulation for isolated word recognition was done with 7 word database spoken by seven male speakers. Best performance was obtained when transition regions between phonemes were modeled by two-state HMM's and stationary phoneme regions by one-state HMM's excluding /b/, /d/, and /g/. By merging rarely occurring diphone units, the recognition rate was increased from 93.98% to 96.29%. In addition, a local interpolation technique was used to smooth a poorly-modeled HMM with a well-trained HMM. With this technique we could get the recognition rate of 97.22% after merging some diphone units.

요 약

본 논문에서는 한국어 음성인식에 적합한 음성 인식 단위에 대해서 연구하였다. 좋은 음성 인식 시스템을 구현하기 위해서는 발음된 음성내의 조음화현상을 처리할 수 있는 인식단위를 선택해야만 한다. 따라서 음소보다 개념적으로 확대된 인식단위가 필요하게 되는데, diphone은 음소간의 전이영역을 modeling하기때문에 좋은 인식 단위가 될 수 있다. Diphone을 인식 단위로 할 경우에 안정적인 음소영역을 diphone사이에 삽입할 수도 있다.

7명의 남성화자가 발음한 74단어로 구성된 고립단어 인식 실험결과 diphone을 2-state HMM으로, 터짐소리 'ㅂ', 'ㄷ', 'ㄱ'와 묵음을 제외한 음소에 대해서 1-state HMM으로 나타냈을 때 가장 높은 인식률을 보였다. 이때 드물게 발생하는 diphone들을 하나의 단위로 merging했을 때 인식률이 93.98%에서 96.29%로 향상되었다. 또한 merging된 diphone과 제안한 국소보간법(local interpolation technique)을 사용함으로써 97.22%까지 인식률이 향상되었다.

*한국과학기술원 전기및 전자공학과
접수일자 : 1993년 6월 1일

Ⅰ. 서 론

인간의 가장 자연스러운 의사 표현 수단인 음성을 기계와 인간사이의 대화에 사용하려는 추세에 따라, 음성 인식을 위한 많은 연구가 지난 수십년동안 계속 진행되었고, 많은 음성 인식 시스템에 개발되었다. 그중 몇몇 시스템들은 높은 인식률을 보였음에도 불구하고, 화자종속, 고립어 인식, 소용량 어휘등으로 인식 환경에 가한 제한때문에 진정한 의미의 음성인식과는 아직 거리가 있는 실정이다[1].

화자 종속의 제한은 극복하기 어렵다고 알려져 있다. 이는 음성을 대표하는 대부분의 특징들이 화자에게 상당히 의존적이어서, 특정 화자에게서 잘 동작했던 시스템도 다른 화자에게서는 상당히 나쁜 인식 성능을 보일 수도 있기 때문이다.

연속어 음성은 크게 다음과 같은 두가지 성질로 인해서 고립어인식이나 연결단어 인식보다 훨씬 어렵게 된다. 연속어는 첫째 단어의 경계가 불명확하고, 둘째 조음화현상이 더욱 심화되어서, 같은 음소라도 문맥에 따라 다르게 발음되는 경우가 있다. 따라서 오인식률도 고립단어의 경우에 비해서 훨씬 높게 된다. 그러나 이런 문제에도 불구하고, 연속어 음성 인식의 중요성은 아주 명백하다. 그 이유는 인간과 기계사이의 통신수단으로 음성이 사용될 때, 자연스러움과 원하는 속도를 얻으려면 오직 연속어를 통해서만 가능하기 때문이다.

1000단어 이상의 인식 대상 어휘를 가진 시스템을 대용량 인식 시스템이라고 한다. 일반적으로 인식 대상 어휘가 증가할수록 비슷한 음운학적 성질을 가지는 단어들이 증가함에 따라 인식률이 낮아지게 된다. 또한 모든 단어 모델들을 학습시키고 각 단어 모델을 저장한다는 것은 매우 어려운 일이기 때문에 인식단위로써 단어를 사용하기보다는 그보다 작은 인식단위르 찾을 필요가 있게 된다. 이렇게 단어보다 작은 단위를 부인식단위(subword unit)라고 하는데, 부인식단위는 단어만큼 조음화현상을 잘 처리하지는 못하기 때문에 성능이 저하될 수 있는 단점을 가지고 있다. 대표적인 부인식단위에는 음소, 음절, diphone, triphone, allophone 등이 있다. 또한 음절이나 diphone, triphone등은 음소사이를 모델링할 수 있는 성질을 가지고 있기때문에 연속어 인식의 경우에서 많이 사용되기도 한다.

본 논문은 화자독립 대용량 연속어 음성인식 시스템을 구현하기 위한 준비 연구 단계로써, 대용량 한국어 음성 인식에 적합한 diphone 집합을 찾고 이를 음운학적으로 균형을 이룬 74개의 고립단어 인식실험을 통해서 평가하고자 한다.

Ⅱ. 입력 음성의 특징 추출

기본적으로 특징 추출법은 샘플링으로 얻어진 적은 양의 음성 샘플들을 보다 적은 양의 데이타로 변화시키는 일종의 압축 기술의 일종이다. 이러한 데이타는 음성의 중요한 특징을 충실히 보여주기 때문에 입력 음성에 대응될 수 있다. 음성 신호를 표현하기 위해서는 에너지나 zero-crossing rate (ZCR)과 같은 기본적인 특징으로부터 short-time spectrum, LPC 모델, 필터 뱅크 모델, homomorphic 모델과 같은 복잡한 표현 방식에 이르기까지 다양한 특징들이 제안되어 왔다[2]. 이중 가장 많이 쓰이는 방법인 LPC 모델과 필터 뱅크 모델을 비교해 보면 비록 협대역 음성에 대해 후자의 성능이 전자의 성능에 비해 다소 떨어진다고 알려져 있으나, 현재까지의 대부분의 상업용 인식 시스템은 LPC 방법보다는 필터 뱅크 모델을 사용하고 있다. 그 이유는 후자가 구현하기에 용이할 뿐아니라 다양하게 응용될 수 있기 때문이다. 3 kHz이상의 대역폭을 가진 음성에 대해서는 필터 뱅크 모델이 마찰음과 음성 전이에 관한 정보를 정확하고 신뢰성있게 제공할 수 있다. 필터 뱅크 모델을 선호하는 또 다른 이유는 잡음이나 다른 형태의 왜곡 현상에 강하기 때문이다. LPC 분석의 결과가 배경 잡음 수준이나, 위상 왜곡 및 반향과 같은 다른 왜곡에 크게 영향을 받는다는 것을 잘 알려진 사실이다[3].

본 연구에서는 필터 뱅크 모델을 사용했으며, 이것은 이 모델이 LPC모델에 비해 주위잡음에 강하고 LPC 모델에 버금가는 성능을 가지고 있기 때문이다. 이 방법에서는 N개의 음성 샘플들로 된 프레임단위의 처리를 통해서 특징 벡터를 구해낸다. 우선 입력 음성은 약 4.5 kHz의 차단 주파수를 갖는 저역 필터를 통과하고, 다음에 10 kHz로 양자화한다. 이 양자화된 음성 신호는 평탄한 스텍트럼 성질을 갖도록 하고, dynamic range를 줄이기 위해 다음식과 같은 전달 함수를 갖는 고정된 1차 preemphasis를 행한다.

$$H(z) = 1 - az^{-1} \tag{1}$$

여기서 a가 0.95일 때 이 시스템은 차단 주파수가 약 120 Hz인 고주파필터가 된다. 이 신호는 다음에

특징 추출을 위해 N(=300)개의 샘플들로 이루어진 프레임 단위로 나누어 진다. 각 프레임은 M(=100)개의 샘플들 만큼씩 떨어져 있다. M<N이면 인접한 프레임들 사이에는 중복된 부분이 있게 된다. 이 중복은 특징 계수들로 이루어진 벡터들을 smoothing하는 역할을 한다. 전처리 과정의 마지막 단계로 각 프레임의 양 끝 부분을 0으로 만들기 위해 각 데이타 블록에 smoothing window를 적용한다. 여기서는 다음 식으로 정의되는 Hamming window를 사용한다[2].

$$w(n)=0.54-0.46cos(\frac{2\pi n}{N-1}) \quad (2)$$

이런 전처리 과정을 통과한 음성신호에 대해서 512 point의 FFT를 취한다. 그리고 FFT한 결과에 대해서 음성특징 벡터로 사용될 필터 뱅크 출력치를 구하였다.

필터 뱅크 출력치를 구하기 위해서는 FFT를 통해 얻어진 음성의 주파수성분을 적당하게 필터 뱅크로 나누는 것이 필요하다. 필터 뱅크 모델에서 이러한 필터의 주파수 간격을 정하는 데는 여러가지 방법이 있다. 본 연구에서는 일찌기 Zwicker에 의해 제안된 자료를 사용해서, 200 Hz이하의 저주파 성분과 4500 Hz이상의 고주파 성분을 무시한 16개의 필터 뱅크를 사용한다. 다음에 고려할 사항은 신호 레벨의 변화에 대한 영향을 감소시키기 위해 신호 전력을 정규화하는 것이다. 여기서는 전 유효 대역내의 주파수 성분의 절대치를 그들의 평균치로 나눈, 소위 평균 정규화를 이용한다. 이렇게 구한 16개의 필터 뱅크 출력치와 음성신호의 에너지를 포함한 17차의 계수와 이들의 시간적인 변화들을 나타내는 Delta계수와 Delta Delta계수를 구함으로써 한 프레임에서 51개의 입력 특징 벡터를 추출했다.

Delta계수는 시간 t에서 계수 $c_t(m)$, $1\le m\le 17$에 대해서 현재의 시간 주변의 2K+1개의 계수를 이용하여

$$\Delta\hat{c}_t(m)=\sum_{k=-K}^{K} c_{t-k}(m), \qquad 1\le m\le 17 \quad (3)$$

와 같이 구해진다. 또한 (2.3)의 결과를 이용하여, Delta Delta계수를 다음과 같이 얻을 수 있다.

$$\Delta\Delta\hat{c}_t(m)=\sum_{k=-K}^{K} k\Delta\hat{c}_{t-k}(m), \qquad 1\le m\le 17 \quad (4)$$

(3)식과 (4)식에서의 K와 K'은 적당한 값을 실험적으로 구해서 사용했다.

Ⅲ. 연속분포 HMM에 의한 단어인식

A. 연속분포 HMM의 개요

HMM(Hidden Markov Model)은 이중적으로 결합된 stochastic process로 구성된 확률적 함수이다. HMM의 내부에 존재하는 것으로 가정되는 Markov chain은 유한한 갯수의 state와 각 state와 결부되어 있는 난수 함수들의 집합을 가지고 있다. 각각의 이산시간에서 프로세스는 어떤 state에 있고, 그 state를 결정하는 난수 함수로부터 한 출력벡터(또는 출력심벌)가 관측된다고 가정한다. 그리고나서 내부의 Markov chain은 전이확률행렬에 따라 다음 state로 전이하게 된다. 따라서 관측자의 입장에서는 오직 state로부터의 출력벡터만을 관측할 수 있을뿐이며, 내부의 state의 상태는 관측할 수 없는 상황이다.

HMM은 한 state에서 관측가능한 벡터들의 분포를 이산분포, 준연속분포, 연속분포등으로 결정할 수 있는데, 이에 따라 이산분포 HMM(Discrete HMM)[4][5], 준이산분포 HMM(semi-Continuous HMM)[8]과 연속분포 HMM(Continuous HMM)[6][7] 등으로 분류된다.

연속분포 HMM에서는 출력벡터분포가 $B=\{b_j(x)\}$, $1\le j\le N$의 형태를 가진다. 이때 $b_j(x)dx$는 관측벡터 O가 x와 $x+dx$ 사이에 있을 확률을 나타낸다. $b_j(x)$의 분포함수는 계산의 편이성때문에 대개 다음과 같은 Gaussian mixture density를 사용한다.

$$b_j(x)=\sum_{k=1}^{M} c_{jk}N(x,\mu_{jk},U_{jk}) \quad (5)$$

여기서 $N(x,\mu,U)$는 평균벡터 μ와 covariance 행렬 U를 가지는 D차원의 정규분포함수이며, Mixture 이득 c_{jk}는 다음과 같은 공리를 만족해야 한다.

$$\sum_{k=1}^{M} c_{jk}=1, \qquad 1\le j\le N \quad (6a)$$

$$c_{jk}\ge 0, \qquad 1\le j\le N,\ 1\le k\le M \quad (6b)$$

여기서 M은 mixture의 수이고 N은 state의 수이다. 따라서 분포함수는 다음과 같이 정규화 된다.

$$\int_{-\infty}^{\infty} b_j(x)dx = 1, \quad 1 \le j \le N \qquad (7)$$

따라서 식(5)의 분포함수를 통해 어떤 연속적인 분포의 근사화도 가능하게 된다.

B. 연속분포 HMM의 파라미터 추출

HMM의 파라미터를 추출하기 위해서는 이론적으로 수렴하는 것으로 알려진 Baum-Welch 알고리즘을 사용할 수 있지만, 기존의 결과에 의하면 평균벡터 μ의 maximun-likelihood estimate는 초기값에 매우 민감한 것으로 알려져 있다. 이런 문제를 해결하기 위해 신뢰성있는 연속분포 HMM의 초기 파라미터값을 구하기위해 segmental k-means 알고리즘이 처음에 제안되었는데[9], 이에 의해 결정된 파라미터의 초기값을 추정된 파라미터로 사용하는 방향으로 개념이 바뀌게 되었다[1]. 우선 segmental k-means 알고리즘을 제시하기 전에 어떤 음성 인식 단위를 HMM으로 나타낼 것인가에 대해 결정할 필요가 있다.

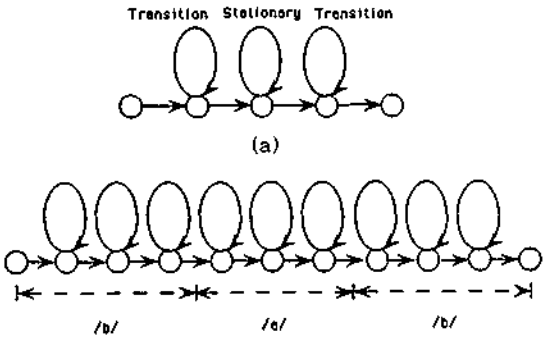


그림. 1. (a) 음소에 대한 HMM (b) 단어 '밥'을 나타내는 HMM

Fig 1. (a) HMM for a phone (b) HMM representing the Korean word 'bab'

하나의 단어는 음소의 열로 구성되어 있다. 음성을 HMM으로 모델링한 경우에는 주로 left-to-right HMM을 사용한다. 이러한 HMM형태의 특성상 단어를 나타내는 HMM을 구성된 음소를 나타내는 HMM들이 직렬연결된 것으로 볼 수 있다. '밥'이라는 단어모델에 대한 HMM이 그림 1에 도시되어 있다. 따라서 모든 음소를 학습함으로써 고립단어에 대한 HMM도 같이 학습하는 효과도 얻게 된다. 이런 식으로 음소단위로 학습을 할 경우의 장점은 인식 어휘의 수가 늘어나게 되었을때, 야기되는 계산량과 메모리의 문제를 해결할 수가 있다는 것이고, 단점은 인식률이 다소 떨어진다는 것이다.

음소단위로 격리단어 인식시스템을 학습하기 전에 모든 인식어휘에 대하여 각 단어를 발음나는 대로 음소를 통해 표기한 lexicon을 구성해야 한다. 그리고 나서 각 음소 HMM을 다음과 같이 segmental k-means 알고리즘에 의해 학습한다.

(1) 초기화 : 각 학습단어에 대한 관측벡터열을, 구성하고 있는 음소의 수에 따라서 균일하게 나눈다. 음소단위로 나눈다음에, 다시 음소모델에 할당한 state의 수만큼 다시 균일하게 분할한다. 따라서 한 단어가 7개의 음소로 구성되었고, 음소 HMM이 3개의 state를 가지고 있다면, 이 단어를 나타내는 관측벡터열은 21(7·3)개로 균일하게 분할된다.

(2) Clustering : 분할후, 각 음소의 각 state영역으로 분할된 학습 집합내의 관측벡터들 중에서 같은 음소와 같은 state를 나타내는 것들만을 모아서 VQ를 통해 M개의 cluster로 분할한다.

(3) 추정 : (2)의 과정을 통해서 다음과 같이 HMM 파라미터들을 갱신할 수 있다.

$\hat{c}_{jk}$ = j state의 k cluster로 분할된 벡터들의 수를 j state에 해당하는 벡터의 수로 나눈 값

$\hat{\mu}_{jkd}$ = j state의 k cluster로 할당된 벡터들에 대한 평균벡터의 d번째 값

$\hat{U}_{jkrs}$ = j state의 k cluster로 할당된 벡터들의 covariance 행렬의 (r, s)번째 값

(4) 재분할 : (2)와 (3)의 과정에서 얻어진 새로운 HMM 파라미터를 가지고 모든 학습 단어를, Viterbi decoding을 통해서 구한 state 열에 따라서 다시 음소와 각 음소의 state 단위로 분할한다.

(5) 반복 : likelihood값이 수렴할 때까지, 또는 주어진 반복횟수를 채울 때까지 (2)에서 (4)의 과정을 반복한다.

위의 알고리즘을 사용할 때, 초기에 파라미터 추정영역을 균일하게 설정하는 것에 대한 타당성에 대한 문제를 제기할 수 있다. 그러나 일반적으로 이런 방식의 초기화가 실제적으로는 수렴할 수 있는 안정적인 모델 파라미터를 만들어 낼 수 있는 초기화 방식들 중의 하나라는 것이 실험적으로 입증되어 있다.

음성인식 시스템은 학습과정과 인식과정으로 나누어 구성한다. 인식대상 어휘가 V개라면 앞에서 설명한 segmental k-means 알고리즘에 의해서 모든 어휘를 구성하는 부인식단위에 대한 HMM 파라미터

를 구할 수 있다. 각 단어의 HMM 모델은 구성된 부 인식단위의 HMM을 연결함으로써 나타낸다.

인식은 미지의 음성이 입력되면 음성 전처리 과정을 통하여 관측벡터열 O를 구한뒤, 각 단어에 대한 HMM λ를 이용하여 각 단어에서의 $P_i = P(O|\lambda_i)$, $i=1,2,...,V$를 계산한다. 그리고 이중에서 가장 큰 확률을 갖는 대상어휘를 인식단어로 결정한다.

I. 최적 음성 인식 단위의 선택

A. 한국어 diphone

Diphone에 대한 연구는 외국에서는 꾸준하게 진행되고 있지만[10][11], 국내에서는 이에 대한 연구가 거의 전무할 실정이다. 본 논문에서는 한국어 음성 인식을 위해 이전에 제안된 diphone 집합이 없기 때문에 다음과 같은 diphone 집합들을 정의하고 각각에 대한 인식 실험을 통해서 한국어에 적합한 diphone 집합을 찾고자 한다.

Diphone(1) : 음소사이의 전이 영역만을 모델링한다. N개의 음소에 대해서 N^2개의 diphone이 존재한다.

Diphone(2) : Diphone(1) 집합에 음소 모델을 추가한다. 단 묵음에 관련된 전이 영역 모델은 제거한다. $(N-1)^2+N$개의 단위가 존재한다.

Diphone(3) : Diphone(1) 집합에 음소 모델을 추가한다. 단 음소 모델중 묵음에 대한 모델은 제거한다. N^2+N-1개의 단위가 존재한다.

Diphone(4) : Diphone(3) 집합에서 반모음 /y/와 /w/에 대한 모델을 제거한다. N^2+N-3개의 단위가 존재한다.

Diphone(L) : 한 음소가 끝나는 부분을 모델링한다. 즉 특정한 좌측의 음소에서 불특정한 우측의 음소들과의 전이영역을 모델링한다. N개의 단위가 존재한다.

Diphone(R) : 한 음소가 시작하는 부분을 모델링한다. 즉 불특정한 좌측의 음소들에서 특정한 우측의 음소와의 전이영역을 모델링한다. N개의 단위가 존재한다.

Diphone(L)이나 diphone(R)은 한쪽의 문맥에만 의존하기 때문에 편이상 H-Diphone(Half-dependent Diphone)으로 명명하고, diphone(1), diphone(2), diphone(3), diphone(4)등은 양 쪽의 문맥에 의존하므로 B-Diphone(Both-dependent Diphone)으로 명명한다. Diphone(L)과 diphone(R)에 diphone(1), diphone(2), diphone(3), diphone(4)에서와 같은 변화를 줄 수 있다.

B. Diphone 모델의 신뢰도 개선

Diphone은 단위의 수가 많기 때문에, 각 단위를 충분하게 학습시키는 것이 어려워진다. 따라서 서로 유사한 특성을 가지는 diphone단위들을 하나의 단위로 merging함으로써, 부족한 학습 집합을 효과적으로 사용하는 알고리즘들이 많이 제안되었는데, 본 논문에서는 이미 Bell Lab에서 사용된 바있는 unit reduction rule[12]을 통해 diphone 단위들을 merging 시켜보았다. 이 방식은 아주 간단하기 때문에, 많은 계산량을 요구하지 않고, 효율적으로 merging을 수행한다.

Reduction Roule : if $c(P_L-P_R)<T$, $P_L-P_R \rightarrow \$-P_R$(or $P_L-\$$)

P_L	좌측음소
P_R	우측음소
P_L-P_R	Diphone단위
$c(P_L-P_R)$	Training set내에서의 diphone 단위의 빈도수
T	Count threshold
\$	불특정음소(don't care)

즉, diphone 단위의 빈도수가 T개이하가 되면, 이를 우측이나 좌측의 문맥에만 의존하는 diphone 단위로 바꾸어준다.

불충분한 학습에 대한 또 다른 접근 방식은 deleted interpolation[13]과 같은 방법을 통해서 유사한 음성구간을 모델링하는 diphone단위들을 적절히 보간(interpolation)함으로써 잡음에 강하게 만드는 것이다. B-diphone은 매우 특정한 구간을 적은 데이타를 통해 모델링하기 때문에 잡음에 매우 민감한 특징이 있는 반면, H-diphone은 충분히 학습되었기 때문에, 잡음에 상대적으로 강한 특성을 가지고 있다. 따라서 B-diphone을 유사한 음성구간을 다루는 H-diphone과 적절히 보간함으로써 잡음에 강하게 만들 수 있다. 이와 같은 목적으로 다음과 같은 국소보간법을 사용한다.

B-diphone과 H-diphone을 보간하려면 밀도함수가 같은 종류이어야 한다. 이때 B-diphone P_L-P_R의 state j에서의 출력 밀도 함수를 B_j^B라 하고, H-diphone $\$-P_R$(or $P_L-\$$)의 state j에서의 출력 밀도 함수를 B_j^H라 한다면, 다음과 같이 보간된 밀도 함수 $\bar{B}_j^B$를 만들 수 있다.

$$\bar{B}_i^H = \lambda(r_k) B_j^H + (1-\lambda(r_k)) B_j^H \quad (8)$$

전체 학습 집합내에서의 k라는 index를 가지는 B-diphone 단위의 빈도수를 c_k^B라 하고, 이 B-diphone unit에 대응하는 H-diphone의 학습 집합내에서의 빈도수를 c_k^H라 했을 때, unit rate r_k를 다음과 같이 정의한다.

$$r_k = \frac{c_k^B}{c_k^H}, \qquad 0 < r_k \leq 1 \quad (9)$$

H-diphone에 해당하는 B-diphone이 하나밖에 없을 때, r_k는 1이 된다. r_k가 1 되는 경우를 제외한 r_k의 평균을 μ_{eff}이라 정의한다. 이때 λ는 다음의 식에 의해서 결정된다.

$$\lambda(r_k) = \begin{cases} T_1, & r_k \leq \dfrac{\mu_{eff}}{2} \\ T_1 + (1-T_1)(1-e^{-a(r_k - 5\mu_{eff})}), & r_k > \dfrac{\mu_{eff}}{2} \end{cases} \quad (10)$$

여기서 T_1은 λ의 최소값이고, a는 r_k가가 μ_{eff}와 같을 때 λ의 값이 $T_1 + 0.99(1-T_1)$가 되도록 정했다. 국소 보간법에 의해 결정된 $\bar{B}_j^B$는 unit rate가 μ_{eff}보다 클 경우에는 거의 B_j^H와 같은 분포를 가지고, unit rate가 극히 적은 경우일지라도 일정한 가중치만큼 B_j^H와 보간한다. 위에서 알 수 있듯이 λ를 학습에 의해서 구하지는 않지만, B-diphone 집합과 H-diphone 집합을 사용해서 각각 학습하는 일은 여전히 필요하게 된다.

V. 실험 및 고찰

본 논문에서 사용한 음성 데이타는 음운학적으로 균형을 이룬 74개의 격리 단어로 구성되어 있다. 학습 집합은 서로 다른 5명의 화자로부터 1번씩 발음된 격리 단어들로 구성했으며, 인식 실험에는 학습에 참여하지 않은 서로 다른 3명의 화자로부터 1번씩 발음된 데이타가 사용되었다. 발음된 74개의 격리 단어는 다음과 같다.

아들	다리	동태	간판	귤	칼	농비	완수	
얘기	딸	의사	간식	하나	마음	옷	웬일	
밥	들개	가보	글	휴러	멥새	옷밥	원고	
바퀴	등쌀	값이	꿀	획기적	목	풀	약속	
뿔	된장	가구	고삐	자리	나	사람	양	용산
비행	돌다리	가족	곡식	잣송이	납기	셈	예	육성
보리	동백	갈치	구리	찌개	날뛰다	쌀	역사	
창	동이	감기	구웠다	줄기	남산	부구	연못	
달	동쪽	감자	괜찮다		늑대	왔다	욕	

4장에서 기본적인 한국어 diphone의 후보를 4개 선정해 보았는데, 각각의 diphone 집합을 segmental k-means 알고리즘에 의해서 학습한 후, 인식 실험을 했다. 같은 환경하에서 32개의 음소를 사용할 경우의 인식 결과도 알아 보았다. 실험 결과를 표1에 나타냈다.

표 1. 여러 diphone 집합에서의 인식률(I)
Table 1. Recognition rate for several diphone sets(I)

부인식단위	단위수	HMM state 수	학습횟수	인식률(%)
Phone(1)	32	3	2	84.25
			10	93.51
			20	93.98
Phone(2)	32	3	2	82.40
			6	91.66
Diphone(1)	178	3	2	84.72
			6	86.11
Diphone(2)	160	3	2	55.09
			6	64.81
		2	2	68.51
Diphone(3)	197	3	2	63.42
			6	66.74
		2	2	86.64
			6	89.35
Diphone(4)	196	2	2	87.96
			6	89.81

인식 결과를 살펴보면 학습 반복 횟수가 증가할수록 점차로 인식률도 증가하는 것을 알 수 있다. 그러나 반복 횟수가 10회이상이 되면 점차로 인식률이 수렴되는 양상을 보이거나 또는 overfitting에 의해서 오히려 인식률이 떨어지는 경우도 발생하게 된다. 특히 diphone 집합을 사용하게 되면, 위와 같은 현상을 낮은 반복 횟수에서도 관찰할 수가 있다. 따라서 이후의 학습에서는 6회이상 segmental k-means 알고리즘을 반복하지 않았다.

위에서 phone(1)은 인식 단어의 시점과 종점에 묵음 구간이 없다고 가정했을 경우이고, phone(2)는 묵음 구간을 가정했을 때이다. 이 두 경우를 비교해 보면 음소를 통해서 인식을 할 경우에는 묵음 구간을 삽입하는 것이 나쁜 결과를 보이는 것을 알 수 있다. 실험에서는 묵음에도 음소와 마찬가지로 3개의 HMM state를 할당했는데, 사실 묵음의 성질은 random 잡음에 가깝기는 하지만, 발생하는 구간이 짧고 에너지가 아주 작기 때문에 3개의 state를 묵음에 할당하는 것은 별로 바람직하지 못하다. 또한 실험에 쓰인 음성 데이타들은 묵음구간이 잘 제거되었기 때문에 학습 단어의 첫 음소와 마지막 음소에 해당하는 특징 벡터들의 일부분이 강제적으로 묵음 구간에 할당되는 경우도 생긴다(HMM의 한 state에서는 반드시 하나 이상의 특징 벡터가 관찰된다). 따라서 묵음 구간에 대해서는 1개 정도의 HMM state를 할당하는 것이 합당하다.

Diphone 집합에 대한 인식 실험의 결과를 보면, diphone(2) 집합은 매우 나쁜 성능을 보여주고 있다. 이는 묵음 구간과 음소사이의 전이 구간에는 매우 많은 정보가 숨겨져 있음을 알려 준다. 이 구간에는 음소를 발음하기 위해서 발성기관이 움직이는 동안에 발생되는 잡음과 발음후 발성기관이 제자리를 찾아가는 과정에서 생기는 잡음의 정보등도 포함되어 있다. 연속어 음성 인식을 위해서는 후보 단어들의 최적의 경로를 찾는 과정이 필요한데, Bellman의 optimality principle[14]에 따르면 입력 음성의 시작되는 부분이 매우 중요한 의미를 가지게 된다. 따라서 묵음과 연결된 diphone은 음성 단위로써의 중요성이 크게 되며, 특히 연속 음성의 경우에서는 '묵음-음소' 형태로 된 diphone 단위가 인식 결과를 결정하는 중요한 요소가 될 수도 있다.

Diphone(3) 집합의 결과를 보면, HMM state를 3개 할당할 경우와, 2개 할당할 경우에의 인식률이 큰 차이를 보이고 있다. Diphone 단위에 HMM state를 3개 할당하게 되면 segmental k-means 알고리즘을 사용할 때, HMM state당 할당되는 특징 벡터의 수가 제한되기 때문에 불충한 모델링이 야기된다. 이는 2개의 state를 할당한 경우와 비교해 보면 쉽게 증명이 된다.

Diphone(4) 집합과 diphone(4) 집합을 비교해보면 diphone(4)를 사용했을 때, 높은 인식률을 보여준다. 이는 반모음에 대한 모델을 생략하는 것에 대한 타당성을 제시하기는 하지만, 시간에 대한 반모음의 구간이 매우 짧기 때문에 묵음 구간에 대해서 논의한 것과 유사한 문제를 야기할 수 있다.

위에서의 실험은 순수 diphone 단위와 음소에 대해서 같은 수의 HMM state를 할당한 결과였다. 이 실험에서 diphone 집합은 89.81%의 인식률만을 보였다. 이 결과는 음소(93.98%)에 비해서 매우 낮은 것이다.

표2는 diphone 단위와 음소 단위에 대해서 서로 다른 HMM state를 할당했을 때의 인식 실험결과이다.

표 2. 여러 diphone 집합에서의 인식률(II)
Table 2. Recognition rate for several diphone sets(II)

부인식단위	단위수	Phone HMM state의 수	Diphone HMM state의 수	학습횟수	인식률(%)
Diphone(3)	197	1	3	2	83.33
			2	2	90.74
				6	93.05
Diphone(4)	195	1	3	2	79.62
				6	82.40
			2	2	89.92
				6	92.12

이 실험에서는 음소에 대한 HMM state는 1개로 고정시키고, diphone의 HMM state 수를 변화시켜 보았다. 결과를 보면 diphone의 state가 2개일 때가 훨씬 나은 성능을 보였다. 그러나 여기에서는 diphone(3) 집합을 사용했을 경우가 diphone(4) 집합을 사용한 경우보다 인식률이 조금 높아졌는데, 이는 반모음에 적당한 HMM state를 할당함에 따라 반모음 자체뿐만이 아니라 반모음에 결합된 diphone들까지도 학습이 잘 되었기 때문이다. 반모음도 묵음과 마찬가지로 state의 수에 민감하게 반응한다.

지금까지의 실험을 통해서 *diphone*(3) 집합을 기본적인 한국어 diphone 집합으로 결정하고 음소에 1개의 HMM state를, diphone에는 2개의 HMM state를 할당한다. 이후의 실험은 이와 같은 환경을 기본으로 한다. 그림2에 실험에 의해 결정된 단위에 대한 HMM이 나타나 있다.

앞 장에서 제안된 보간법을 적용하려면 H-diphone 집합을 설정해야 한다. 기존의 실험에 의하면 한 음소는 우측의 음소에 더 영향을 많이 받는다는 것이

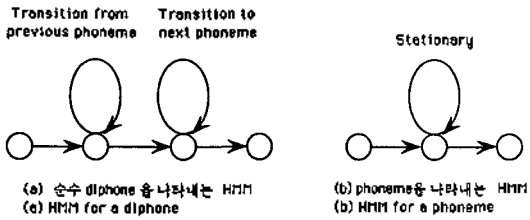


그림 2. 설정된 diphone set에 대한 HMM
Fig 2. HMM for the determined diphone set

알려져 있다. 실험에 의해서 이를 증명해 볼 수 있다. 이를 위해서 diphone(L) 집합과 diphone(R) 집합에 diphone(3) 집합의 조건을 가한다.

표 3. H-diphone의 인식률
Table 3. Recognitionn rate of H-diphone

부인식단위	단위수	Phone HMM state의 수	Diphone HMM state의 수	학습횟수	인식률(%)
Diphone(R)	100	1	2	2	83.33
Diphone(L)	100	1	2	2	73.61

표3을 보면 Diphone(R)을 사용하는 것이 합리적임을 알 수 있다. 따라서 보간의 대상이 되는 H-diphone 집합은 diphone(R) 집합으로 결정한다.

표4는 unit reduction rule을 사용할때, count threshold를 각각 5, 10, 15로 했을 때의 인식률을 보여준다.

표 4. Unit reduction rule에 대한 merging후의 인식률
Table 4. Recognition rate after merging by the unit reduction rule

Count threshold	단위수	학습횟수	인식률(%)
5	137	2	93.05
		6	93.98
10	110	2	93.51
		6	93.05
15	95	2	89.81
		6	92.59

지금까지의 실험에서 diphone을 인식 단위로 했을 때, 최고 93.98%의 인식률까지 얻을 수 있었다. 그러나 이런 인식률은 음소를 사용할 때에 비해서 그다지 높은 것이 아니어서 선정된 diphone 집합의 타당성에 의문을 품을 수가 있다.

앞에서 선정된 diphone 집합은 기본적으로 묵음을 제외한 모든 음소에 대한 HMM을 포함하고 있다. 묵음같은 경우는 주변의 diphone에 의해서 음성이 연속적으로 모델링되므로 인식 성능에서 큰 영향을 주지 않는다. 한국어의 터짐소리인 'ㅂ', 'ㄷ', 'ㄱ'는 지속 시간이 매우 짧아서 다른 음소에 비해 인식하기가 어렵다. 이들은 각 조음위치에서 혀, 입술로 공기의 흐름을 순간적으로 차단시켜 구강내의 압력을 높인 다음, 순간적으로 조음기관을 뗌으로써 공기가 파열되어 발성이 된다. 이러한 과정이 순간적이므로 이들 음소를 구별하기 위해서는 다음 모음으로의 천이 과정이 매우 중요하게 된다. 그러나 diphone을 사용할 경우에는 이런 천이과정이 모델링되므로 이들에 대한 음소 HMM을 따로 두는 것은 소모적이 될 수가 있다. 따라서 터짐소리 음소 모델을 제거한 diphone 집합을 생각할 수 있는데, 이런 집합을 가지고 인식 실험을 수행해 보았다. 그 결과는 표5에 나타나 있다.

표 5. 터짐소리를 제거한 diphone 집합에 의한 인식률
Table 5. Recognition rate of the diphone sets without plosives

Count threshold	단위수	학습횟수	인식률(%)
1	194	2	91.66
		6	93.98
5	134	2	95.37
		6	95.37
10	107	2	94.90
		6	96.29
15	92	2	93.98
		6	94.44

실험 결과, 터짐소리를 제거했을때 큰 폭으로 인식률이 향상되었다. 이로써 터짐소리들은 음소 모델이 lexicon의 구성에 참여하지 못하더라도, 주변의 diphone들은 학습이 더 잘되기 때문에 오히려 인식률이 증가한다는 것을 알 수 있다. 따라서 앞 절에서 결정한 한국어 diphone 집합에서 터짐소리에 대한 음소 모델을 제거한 집합을 최종적인 한국어 diphone 집합으

로 결정한다.

앞절에서 제안한 국소보간법을 적용하기 위해서는 학습된 H-diphone 집합이 필요하다. 이들에 대해서는 4번의 반복 학습을 했다. 결과는 표6에 나타나 있다. 표에서 인식률 I은 보간되지 않았을 때의 인식률이고 인식률 II는 보간되었을 때의 인식률이다. 인식률 I은 비교의 목적으로 같이 나타냈다. λ의 최소값은 실험적으로 0.9999로 결정했다.

표 6. 국소보간법을 사용했을 때의 인식률(%)
Table 6. Recognition rate after local interpolation

Count threshold	단위수	학습횟수	인식률I(%)	인식률II(%)
1	(H-diphone) 60	4	91.20	
	194	6	93.98	93.98
5	134	6	95.37	97.22
10	107	6	96.29	97.22
15	92	6	94.44	94.91

Ⅵ. 결 론

본 연구에서는 한국어 연속 음성 인식 시스템을 구현하기 위한 기초연구로써, 한국어의 인식에 적합한 인식 단위를 찾기 위해서 논의했다.

고립단어를 인식하는 경우에는 음소 사이의 조음화 현상이 단어를 구별하는데 큰 어려움을 주지 않지만, 연속어로 인식 범위가 확장될 경우에는 음소사이와 단어사이의 조음화현상때문에 단어의 경계가 대단히 애매모호해진다. 따라서 이런 조음화 현상을 수용할 수 있는 보다 강력한 인식단위가 필요하게 된다.

조음화현상을 고려하기 위해서는 음소보다 개념적으로 확대된 인식 단위가 필요하게 되는데, 현재까지 제안된 인식 단위중에는 triphone이나, diphone이 가장 합당한 것으로 알려져 있다. 따라서 본 논문에서는 diphone을 인식 단위로 설정하고, 한국어에 가장 적합한 diphone 집합을 찾기 위해서 기본적인 diphone 집합들을 제안한 다음, 이들로부터 가장 좋은 인식 성능을 보인 diphone set를 찾아 보았다.

실험결과 음소 사이의 전이구간(diphone)을 2-state HMM으로 모델링하고 , /b/, /d/, /g/와 묵음에는 HMM을 만들지 않고, 이들을 제외한 모든 음소를 1-state HMM으로 모델링한 것이 가장 적합하다는 것을 알 수 있었다. 이때의 인식률은 93.98%였는데, 이는 음소를 사용했을 경우와 동일했다. 그러나 unit reduction rule에 의한 merging을 했을 때, 93.98%의 인식률이 96.29%로 개선되었다.

성능을 개선하기 위해서 국소보간법을 제안했다. 실험결과, 인식률이 97.22%까지 향상되었다. 따라서 제안된 방법이 74개의 고립단어 인식실험에서는 타당성이 있음을 알 수 있었다. 그러나 이 방법들이 다른 경우에서도 똑같이 적용되리라는 보장이 없으므로 다른 인식 환경(다른 database, 연속어 인식등)에서도 검토해 볼 필요가 있다.

참 고 문 헌

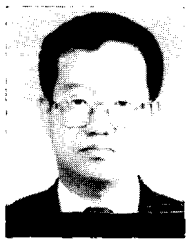
1. K. F. Lee, *Automatic Speech Recognition*, Kluwer Academic Publishers, 1989.
2. L. R. Rabiner and R. W. Schafer, *Digital Signal Processing of Speech Signal*, Englewood Cliffs, N.J., Prentice-Hall, 1978.
3. J. S. Lim, "Estimation of LPC Coefficients from Speech Waveforms Degraded by Additive Random Noise," *Proc. IEEE Int. Conf. ASSP*, pp. 599-601, 1978.
4. B. H. Juang and L. R. Rabiner, "Issues in Using Hidden Markov Models for Speech Recognition," *Advances in Speech Signal Processing*, Marcel Dekker, Inc., pp. 509-553, 1991.
5. L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, Jan. 1986.
6. L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Tech. Journal*, Vol. 64, No. 6, July-August 1985.
7. L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Some Properties of Continuous Hidden Markov Model Representation," *AT&T Tech. Journal*, Vol. 64, No. 6, July-August 1985.
8. X. D. Huang and M. A. Jack, "Semi-Continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language* (1989) 3, pp. 239-251.
9. L. R. Rabiner, J. G. Wilpon and B. H. Juang, "A Segmental K-means Training Procedure for Connected Word Recognition," *AT&T Tech. Journal*, Vol.

65, No. 3, 1986.
10. A. M. Collar and A. E. Rosenberg, "Unsupervised Bootstrapping of Diphone-like Templates for Connected Speech Recognition," *Proc. IEEE Int. Conf. ASSP*, 1987.
11. A. M. Colalr and A. E. Rosenberg, "A Connected Speech Recognition System Based on Spotting Diphone-like Segments-Preliminary Results," *Proc. IEEE Int. Conf. ASSP*, 1987.
12. C. H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, "Acoustic Modelling for-Large Vocabulary Speech Recognition," *Computer Speech and Language* (1990) 4, 127-165.
13. F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," *Pattern Recognition in Practice*, pp. 381-397, North-Holland Publishing Company, Amsterdam, the Netherlands, 1980.
14. R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.

▲박 현 상 1969년 2월 13일생

1991년 2월 : 한국과학기술대학 전기 및 전자공학과 졸업(공학사)
1993년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
1993년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정

▲박 용 규(회원) 1960년 12월 30일생



1984년 2월 : 한양대학교 전기공학과 졸업(공학사)
1987년 8월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
1987년 10월~현재 : 한국통신 연구개발단 연구원
1991년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정

▲권 오 욱(비회원) 1964년 3월 6일생

1986년 2월 : 서울대학교 전자공학과 졸업(공학사)
1988년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
1988년 3월~현재 : 한국전자통신연구소 연구원
1992년 3월~현재 : 한국과학기술원 전기 및 전자공학과 박사과정

▲은 종 관 : 10권 3호 참조