HMM의 교정 학습과 후처리를 이용한 연결 숫자음 인식에 관한 연구

○우인봉,이강성, 김순협
광운대학교 컴퓨터공학과

A Study on the Recognition of the Connected Digits Using Corrective Training with HMM and Post Processing

Woo In-Bong, Lee Kang-Sung, Kim Soon-Hyeob
Kwang-Woon Univ. Dep't of Computer Engineering

요 약

HMM (Hidden Markov Model) 은 좋은 결과를 보이면서 현재 음성 인식 분야에서 널리 사용되는 알고리즘이다. 그러나, 이 HMM 의 학습 방법인 maximum like-ihood estimation 은 인식률을 극대화하는 모델의 파라메터 값을 생성하지 못하는 단점이 있다. 이러한 문제점을 보완하기 위하여 연결어 인식 알고리즘인 Segmental K-means 의 학습 과정에 교정 학습법 (Corrective Training) 을 도입하여 모델 파라메터 값을 재조정해 준다.

한국어 연속 숫자음은 영어 연속 숫자음과 달리 연음 현상의 영향을 많이 받는다. Level Building 과정에서 연음에 의한 오류를 감소시키기 위해 연음에 의해 발생할 수 있는 단어를 별도의 모델로 추가했다. 이렇게 추가된 단어 모델들에 대한 몇 가지 규칙을 인식 결과에 적용하여 출력을 다시 조정한다.

본 시스템은 TMS320C30 프로세서 내장한 DSP (Digital Signal Process) 보드와 IBM PC 상에서 구현되었고, 표준 패턴은 실험실 잡음 환경에서 남성 화자 3 명을 대상으로 작성하였다. 인식 결과 21 종 전화 번호 252 개 데이타에 대하여 화자 종속으로 92.1% 인식률을 나타내었다.

1. 서 론

HMM 알고리즘은 모델을 작성할 때 같은 category 에 속하는 데이타들만을 사용하고 다른 category 의 모델과의 상대성은 전혀 고려하지 않기 때문에 비슷한 확률값을 출력하는 모델을 생성할 수가 있다. 본 논문에서는 이러한 단점을 보완하기 위하여, 선형 분류의 학습 절차와 유사한 교정 학습법 (corrective training) 을 도입하였다. 이것은 이미 생성된 모델을 선정된 데이타들의 상대적인 값이 고려되도록 다시 학습시키는 과정이다. 이 방법을 연결어 인식 알고리즘인 Segmental K-means 과정에서 새로이 발생하는 단어 토큰과 모델에 적용하여 각 모델들을 재조정하였다. 학습 대상은 HMM 파라메터중 인식률에 가장 큰 영향을 미치는 출력 확률값이다. 학습시 출력 확률값은 선정된 학습 데이타를 고려하여 재조정되는 것이므로 학습데이타의 선정에 따라 인식률이 오히려 저하되는 경우가 있다. 따라서 인식률을 증가 시키는 데이타를 찾기 위한 반복적 실험이 요구 된다.

한국어 숫자음은 대부분 단음절이고 초성이 'ㅇ' 인 단어가 많아 연속 발음시 영어 숫자음에 비해 연음 현상을 많이 받는다. 연결어 인식 알고리즘인 Segmental K-means 과정이 어느 정도의 조음 현상은 극복할 수 있으나 한국어 연결 숫자음의 경우와 같은 연음 현상을 극복하지는 못한다. 이러한 점을 개선하기 위하여 숫자음 조합 시에 발생할 수 있는 새로운 단어들에 대한 모델을 별도로 만들고, 이들의 결합 규칙을 적용하는 후처리 과정을 첨가시켜 인식된 결과를 규칙에 맞게 바꾸어 인식률을 향상시켰다.

2. 인식 알고리즘

2.1 HMM 을 이용한 음성인식

HMM 알고리즘에는 인식과 관련된 전향 (forward) 알고리즘, 후향 (backward) 알고리즘, Viterbi 알고리즘, 학습과 관련된 Baum-Welch 재추정 (reestimation) 알고리즘이 있다.[5]

(1) 전향 알고리즘 ⇒ $P(O \mid \lambda)$

(2) 후향 알고리즘 ⇒ $P(O \mid \lambda)$

(3) Viterbi 알고리즘 : 최적 상태열 (state sequence)

(4) Baum-Welch 재추정 알고리즘 : A, B, π 학습

HMM 을 이용한 음성 인식 시스템에서 각 대상 어휘들에 대한 HMM 들이 블랙 박스와 같은 역할을 하여 미지의 데이타를 각 모델에 적용했을 때 가장 큰 값의 확률을 나타내는 것을 인식된 것으로 한다. 본 논문에서 상태 전이는 상태 자체에 시간적 순서를 부여할 수 있는 left-to-right 모델을 사용하였고, 상태수와 코드 워드 수는 어휘수와 계산량을 고려하여 선정하였다. HMM 을 이용한 기본적인 인식 과정은 다음과 같다.

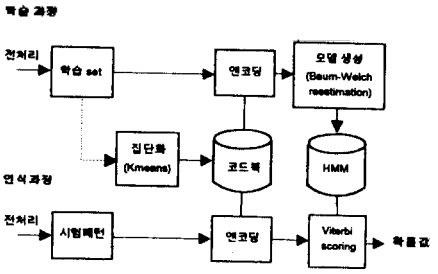


그림 1. HMM 을 이용한 음성 인식 흐름도

2.2 연결어 인식 알고리즘

2.2.1 Level Building 알고리즘 [7]

연결어 인식에 사용되는 대표적인 알고리즘으로 단독어 표준 패턴들을 미지의 연결어와 비교하여 일치하는 최적 단어열을 결정한다. 즉, 단독어 표준 패턴들의 최적열을 결정한다. 단독어 표준 패턴으로 HMM 을 사용했고 시험 패턴과 표준 패턴의 정합은 Viterbi 알고리즘을 이용했다.

2.2.2 Segmental K-means 알고리즘 [8]

초기 연결어 인식은 단독어를 학습시켜 수행되어 왔다. 이것은 느린 발성음과 명확하게 발음된 정상 속도의 발성음의 경우 비교적 잘 동작하였으나 조음 현상이 나타나는 일반적인 발성음에는 적합하지 못한 방법이다. 이러한 문제점을 해결하기 위해 단독어 표준 패턴을 학습시키는 것이 아니라 단독어 표준 패턴을 이용하여 연속 단어열로 부터 추출해 낸 단어들 학습시키는 방법이 고안되었다. 이 학습 방법에서는 표준 패턴을 개선하기 위해 추출된 패턴과 단독어 표준 패턴을 조합시킨다.

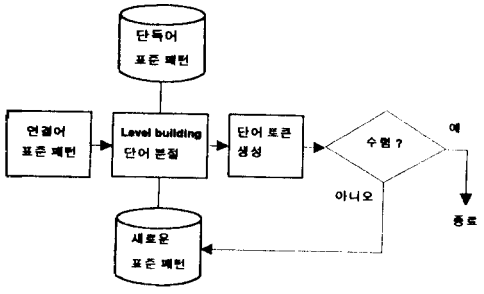


그림 2. Segmental K-means 알고리즘 흐름도

3. 교정 학습의 도입

HMM 학습 과정에서 일반적으로 사용되는 Baum-Welch 재추정 알고리즘은 자신의 category 에 해당하는 데이타들만을 고려할 뿐 다른 category 의 데이타들은 전혀 고려하지 않기 때문에 서로 비슷한 확률값을 출력시키는 모델들이 형성될 수 있다. 특히, 연결어 인식의 경우, Segmental K-means 에 의해 자동으로 분절된 모델들의 에러율은 더욱 증가 된다. 이러한 문제점을 보완하기 위하여 본 논문에서는 L.R. Bahl, P.V.de Souza 등에 의해 개발된 교정 학습법[2][3] 을 연결어 인식 알고리즘에 도입하였다. 교정 학습법은 선형 분류에 사용되는 error-correcting 학습법과 유사한 방법이다. 이 방법은 Baum-Welch 재추정 알고리즘으로 만들어진 모델에 다른 category 와의 상대적인 특성이 부여되도록 다시 학습하는 과정이다.

3.1 교정 학습

3.1.1 빈도수 (frequency count)

관측 심볼의 출력 확률값을 조정하기 위해서는 먼저 각 심볼인 코드 워드 (code word) 의 상대적인 발생 횟수를 구해야 한다. 이것을 빈도수[2] 라하며 아래와 같은 식으로 나타낼 수 있고, 실제로는 Baum-Welch 재추정 알고리즘을 이용하여 구한다. 이것은 결과적으로 한 음성 데이타에서 사용된 전체 코드 워드의 발생 횟수에 대한 각 코드 워드 발생 횟수의 상대적 비율이 된다. 빈도수 (count) 는 다음과 같은 식으로 나타낼 수 있다.

$$\eta(b_{lm} \mid O,\ \lambda_j) = b'_{lm} \frac{\partial L(O \mid \lambda_j)}{\partial b_{lm}} \Big|_{b=b'}$$

$$= \frac{\sum_{I=1}^{S} P(O,\ I \mid \lambda_j) \times \eta_{lm}(I)}{\sum_{I=1}^{S} P(O,\ I \mid \lambda_j)} \Big|_{b=b'} \qquad [4]$$

l : 상태 번호 (state index : $1 \le l \le L$)

m : 관측 심볼 번호 (observation symbol index : $1 \le m \le M$)

j : 모델 번호 (model index : $1 \le j \le p$, p : 어휘수)

S : 모든 상태열의 수 (total state sequences)

$I = (i_1, i_2, \ldots, i_T)$: 상태열 (state sequence)

O : 관측열 (observation sequence)

b'_{lm} : 심볼의 출력 확률 (output probability)

3.1.2 가중치

한 관측열 (observation sequence) 을 자신의 category 에 해당하는 모델에 적용한 값과 다른 category 의 모델에 적용한 값의 차이를 가중치로 설정하는데, 서로 독립적인 확률 분포를 결합할 때 어떤 데이타가 정확히 어떤 영향을 미치는지 알 수 없으므로 두 개의 상수값 β_W (within-class learning rate), β_B (between-class learning rate) 를 설정하여 가중치의 크기를 실험적으로 조절한다. β_W 는 학습하고자 하는 category 의 데이타 중에서 다른 category 로 틀리게 인식되는 데이타들을 얼마만큼 고려할 것인가를 결정하는 상수이고, β_B 는 다른 category

외 데이타중에서 학습하고자 하는 category 로 틀리게 인식되는 데이타들을 얼마만큼 고려할 것인가를 결정하는 상수이다.

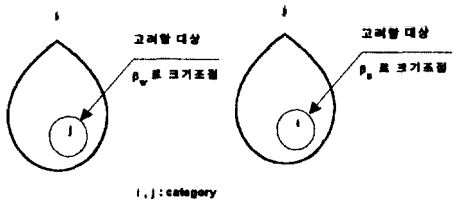


그림 3. β_W, β_B 학습률 (학습 대상 : 모델 i)

인식이 정확히 되었더라도 근소한 확률값의 차이로 인식에 성공했을 경우, 이와 같은 category 외 어떤 데이타가 들어왔을 경우 틀리게 인식될 확률이 높다. 이러한 문제점을 고려하여 확률값의 차이가 어느 상수값 (near-miss criterion : δ) 이하를 나타내면 이 모델도 다시 학습시킨다.[2][3]

$$R = \log\left[\frac{P(O \mid \lambda_i)}{P(O \mid \lambda_j)}\right]$$

$$= \log P(O \mid \lambda_i) - \log P(O \mid \lambda_j), \quad 0 \le R < \delta$$

3.1.3 최종 알고리즘

모든 모델에 대하여 다음과 같은 방법[4] 으로 심볼의 출력 확률값을 갱신한다.

$$\eta^*_{im} = \sum_{\forall C_i} \eta(b_{im} \mid O, \lambda_i) + \beta_W \gamma^* \eta(b_{im} \mid O, \lambda_i) - \beta_B \gamma^0 \eta(b_{im} \mid O, \lambda_i)]$$

⇒ 이 값을 이용하여 HMM i 의 상태 l 의 코드 워드 m 의 새로운 출력 확률값을 갱신한다.

① 항 : 학습 모델이 λ_i 일 때 λ_i 의 category C_i 에 속하는 음성 데이터의 관측열 O 에 대한 상태 l 의 관측 심볼 m 의 빈도수

② 항 : $\gamma^* \eta(b_{im} \mid O, \lambda_i)$

$$\gamma^* = \begin{cases} 0 & \text{if } R > \delta \\ 1 - R/\delta & \text{if } 0 < R \le \delta \\ 1 & \text{if } R \le 0 \end{cases}$$

β_W : within class 학습률

δ : near-miss constant, score 차이가 작은 데이타의 영향 고려

R = logP(O | λ_i) - logP(O | λ_j)

③ 항 : $\gamma^0 \eta(b_{im} \mid O, \lambda_i)$

$$\gamma^0 = \begin{cases} 0 & \text{if } R > \delta \\ 1 - R/\delta & \text{if } 0 < R \le \delta \\ 1 & \text{if } R \le 0 \end{cases}$$

β_B : between class 학습률

R = logP(O | λ_j) - logP(O | λ_i)

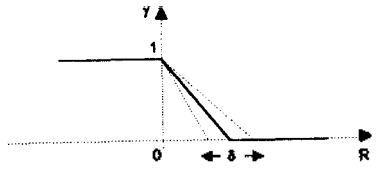


그림 4. δ 값에 따른 가중치 γ 의 변화율

3.2 연결어 인식에 도입

Segmental K-means 학습 과정에서 새로이 발생하는 단어 토큰과 모델들에 대하여 그림 5 와 같이 교정 학습법을 도입한다. 여기서, β_W, β_B, δ 는 생성된 단어 토큰을 새로운 모델에 적용했을 때의 인식률을 이용하여 설정한다.

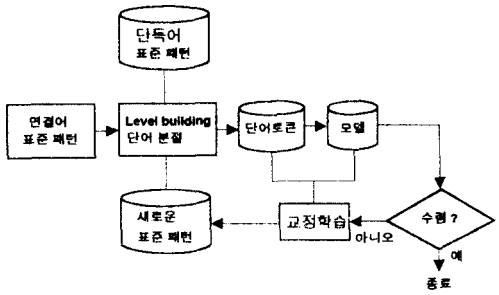


그림 5. 교정 학습이 도입된 Segmental K-means 알고리즘

4. 후처리

4.1 연음 현상

한국어 숫자음은 연속되었을 때, "35" (/사오/) 등과 같이 발음 특성상 많은 연음 현상이 발생하여 자동 다이얼링 시스템과 같은 연결 숫자음 인식에 큰 문제로 작용한다. 영어 숫자음과 달리 한국어 숫자음은 모두 단음절이고, 10 개의 숫자음 중에서 6 개의 숫자음 (/공/, /일/, /삼/, /육/, /칠/, /팔/) 이 종성 발침을 갖고 있고, 4 개의 숫자음 (/일/, /이/, /오/, /육/) 의 초

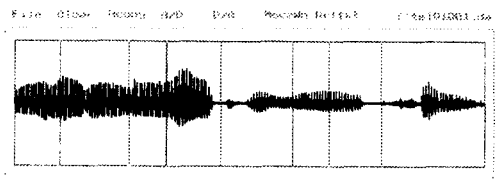


그림 6. 연결 숫자음에서의 연음 및 조음 현상 /512-0257/

성이 'ㅇ'으로 선행 음절 종성 받침의 영향을 받는다.

본 논문에서는 전화번호 발성시 사용되는 /에/ 발음을 추가하여 연음 현상을 고려하였다. 한국어 연결 숫자음에서 나타날 수 있는 연음 현상은 표 1 과 같다.

표 1. 연결 숫자음에서 나타날 수 있는 연음 현상

선행숫자	후행 숫자			
	1	2	5	에
0 /공/	공일 고-ㅇ-일	공이 고-ㅇ-이	공오 고-ㅇ-오	공에 고-ㅇ-에
1	일일 일릴 일-ㄹ-일	일이 일리 일-ㄹ-이	일오 이로 일-ㄹ-오	일에 이레 일-ㄹ-에
3	삼일 사밀 사-ㅁ-일	삼이 사미 사-ㅁ-이	삼오 사모 사-ㅁ-오	삼에 사메 사-ㅁ-에
6	육일 유길 유-ㄱ-일	육이 유기 유-ㄱ-이	육오 유고 유-ㄱ-오	육에 유게 유-ㄱ-에
7	칠일 치릴 치-ㄹ-일	칠이 치리 치-ㄹ-이	칠오 치로 치-ㄹ-오	칠에 치레 치-ㄹ-에
8	팔일 파릴 파-ㄹ-일	팔이 파리 파-ㄹ-이	팔오 파로 파-ㄹ-오	팔에 파레 파-ㄹ-에

※ /공/ 은 전화 번호 명명시 일반적으로 /영/ 대신 사용.
※ '-' : 과도 부분

4.2 연음 처리

위와 같은 연음 현상을 고려하기 위하여 추가되어야 할 단어 모델은 다음과 같다.

/일-ㄹ/, /유-ㄱ/, /치-ㄹ/, /파-ㄹ/, /고-ㅇ/, /모/, /미/, /밀/, /게/, /레/, /메/, /용/, /늑/

여기서, 일-ㄹ, 유-ㄱ 과 같은 발음은 /12/, /62/ 등의 연속 발음으로 부터 수동 분절한 것을 이용한다.

추가된 모델들에 대해 교정 학습과 Segmental K-means 학습을 그대로 적용하여 연결어 표준 패턴을 작성하고, 인식 과정에서 결과에 대해 다음과 같은 규칙을 적용한다.

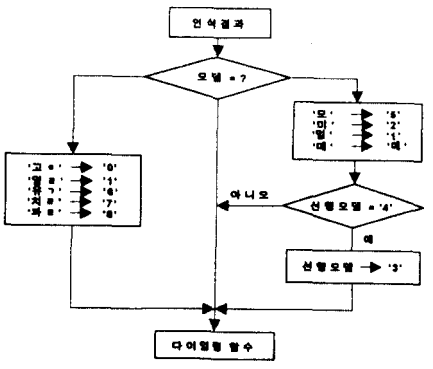


그림 7. 후처리 흐름도

5. 실험 및 고찰

표준 패턴

- 아날로그 필터 : 70 Hz - 4.5 KHz
- A/D : 10 KHz 샘플링, 16 bit 양자화
- Pre-emphasis : $H(z) = 1 - 0.95z^{-1}$
- Window : Hamming window
- LPC cepstrum 계수 : 10 차

단독 숫자음 표준 패턴은 연음 현상을 고려하여 숫자음 10 개와 연음으로 나타날 수 있는 8 가지 단어로 선정하였다. 표준 패턴은 서울 지역 20 대 남성 화자 3 명이 각각 3 번씩 발음한 것으로 실험실 잡음 환경하에서 작성 하였다. HMM 의 상태수는 4, 코드북의 크기는 64 로 하였다.

표 2. 연음 현상을 고려하여 선정된 단어 모델

모델	1	2	3	4	5	6
발음	일 [il]	이 [i]	삼 [sam]	사 [sa]	오 [o:]	육 [yuk]
모델	7	8	9	10	11	12
발음	칠 [c^{h}il]	팔 [p^{h}al]	구 [ku]	공 [kon]	일-ㄹ	유-ㄱ
모델	13	14	15	16	17	18
발음	치-ㄹ	파-ㄹ	모 [mo]	미 [mi]	밀 [mil]	에 [e]

단독 숫자음 표준 패턴 = 3 화자 * 3 회 * 18 모델 = 162 개

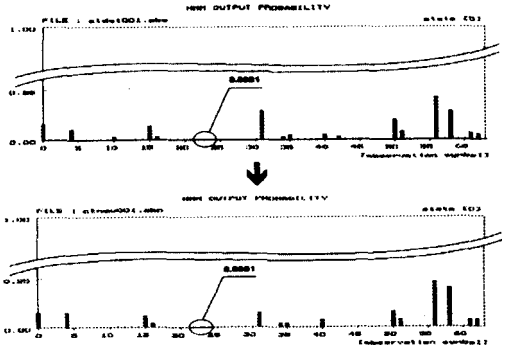
본 논문에서는 막대한 계산량 때문에 제외되었으나 추가로 고려되어야 할 발음은 /01/ 에서의 /고-ㅇ/, /3에/ 에서의 /메/, /66/ 에서의 /육/ 과 /늑/, /60/ 에서의 /공/, /69/ 에서의 /구/, /63, 73, 83/ 등에서의 /쌍/, /64, 74, 84/ 등에서의 /싸/ 등이다.

연음 현상을 처리하기 위해 추가한 모델들은 /12/, /62/ 등의 발음으로부터 수동 분절 (hand segment) 하여 작성하였다.

연속 숫자음 표준 패턴은 연속해서 일어날 수 있는 모든 경우의 수를 조합하여 구성한 21 종의 연속 숫자음이다.

5.1 교정 학습 실험

그림 8 은 숫자음 /1/ 에 대한 모델의 상태 0 의 출력 확률값이 교정학습에 의해 변화된 모습을 나타낸다. 이 그림은 64 개의 코드 워드중에서 숫자음 데이타 /1/ 의 학습시에 관측된 심볼 (observation symbol = code word index) 들의 출력 확률을 나타낸다.



(모델 = '1', 상태수 : 4, 코드 워드수 : 64)

그림 8. 교정 학습으로 변화된 출력 확률값

다음은 초기 표준 패턴으로 사용된 단독 숫자음의 인식률과 Segmental K-means 학습 과정에서 발생하는 단어 토큰들을 새로 생성된 모델에 적용했을 때의 인식률이다. 교정 학습은 생성된 토큰중의 일부를 사용하였고 Segmantal K-means 학습의 최종 단계에서 적용 하였다.

표 3. 단독 숫자음 인식 실험 결과

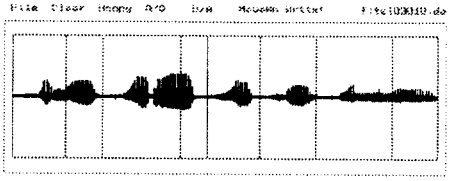
β_W	β_θ	인식률(%)
1.0	1.0	81.5
0.0	0.01	92.3
0.02	0.01	94.3
0.001	0.001	91.3

표 4. 단어 토큰 인식 실험 결과

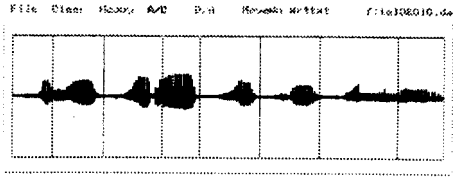
β_W	β_θ	인식률(%)
1.0	1.0	76.1
0.01	0.01	91.2
0.001	0.001	88.1

5.2 후처리를 적용한 인식 실험

실험 결과 252 개 전화 번호에 대해 화자 종속으로 후처리 없는 인식 실험에서 90.2%, 후처리한 인식 실험에서 92.1% 의 단어 단위 인식률을 얻었다. 그림 9 는 후처리, 즉, 연음 모델을 적용 했을 때와 그렇지 않았을 때, 연결 숫자음 /358-8736/ 이 level building 에 의해 자동 분절된 것을 나타낸다. 그림 9-(a) 는 연음에 의해 /458-8736/ 으로 인식되었을 때 분절된 결과이고, 그림 9-(b) 는 모델 /4/, /15/, /8/, /18/, /8/, /7/, /3/, /6/ 이 출력되었을 때의 분절된 결과이다. 그림 9-(b) 의 모델들은 후처리 과정에서 /15/ 에 의해 /4/ 가 /3/ 으로 바뀌어 최종적으로 /358-8736/ 을 출력하였다.



(a) 인식 결과 /458-8736/



(b) 인식 결과 /358-8736/

그림 9. (a) 후처리하지 않았을 때 level building 분절
(b) 후처리했을 때 level building 분절

5.3 고 찰

실험 결과 교정 학습은 약간의 인식률을 개선 시켰으나 모델의 수, 상태 수, 코드 워드 수가 증가함에 따라 계산량의 증가가 극심하고, 세 파라메터 δ, β_W, β_θ 를 일일이 실험을 통해 구해야 하는 문제점이 있으므로 숫자음 인식이나, 음소 단위 인식과 같은 소규모 어휘에 적용하는 것이 알맞을 것이다.

후처리에서 보다 나은 결과를 위해서는 개개인의 연음의 정도 차이를 고려하여 모델을 학습시켜야 하고 연음으로 발생할수 있는 발음외에도 /쌈/, /꾸/ 와 이 경음화로 발생할 수 있는 모델도 추가되어야 할 것이다.

결과적으로 인식률이 증가되는 최대 화자수 설정, 연음 정도에 따른 정확한 모델의 선택, 발생 가능한 모든 연음 모델의 추가, 모델 수에 따른 최적 상태수와 코드 워드수의 설정 등이 이루어 졌을 때 최대 인식률이 얻어질 것이다.

6. 결 론

HMM 의 교정 학습은 각 모델들이 출력하는 확률값의 차이를 더욱 크게 하여 인식률을 증가 시키기도 하지만, 학습 대상의 선정에 따라서는 오히려 인식률을 저하시킨다. 이 학습법이 모델을 만들기 위한 과정이라는 점에서, 가능한 한 많은 실험을 거쳐 학습 데이타를 선정하는것이 바람직 하겠다. 또한, 세 가지 파라메터를 일일이 실험으로 구해야 하는 불편함을 해결하기 위해 이들 파라메터가 각 모델에 미치는 영향에 대해 연구와 그 영향에 따라 자동적으로 파라메터를 구해 내는 연구가 이루어져야 할 것이다.

후처리 실험 결과, 특히, /35/ 등의 발음에서 좋은 결과가 나타나 연음 처리한 Level building 이 타당성을 보여 주었다. 개인별 연음 현상의 정도 차이를 고려한다면 보다 나은 결과를 얻을 수 있을 것이다.

참 고 문 헌

[1] Bahl L.R., P.F.Brown, P.V.deSouza and R.L.Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," ICASSP, p.S11.2, 1988

[2] T.H.Applebaum, B.A.Hanson, "Enhancing The Discrimination of Speaker Independent Hidden Markov Model with Corrective Training," ICASSP, p.S6.13, 1989

[3] L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP magazine, pp4-17, January 1986.

[4] 김 남수 "음성인식을 위한 HIDDEN MARKOV MODEL PARAMETER TRAINING 에 관한 연구," 한국과학기술원

[5] L.R. Rabiner, J.G. Wilpon, and B.H. Juang, "A segmental k-means training procedure for connected word recognition based on whole word reference patterns," AT&T Tech. J., vol. 65, no.3 pp.21-31, May/June 1986.