연속분포 HMM을 이용한 음성인식 시스템에 관한 연구

o
김 상 덕*, 이 극*
*한남대학교 컴퓨터공학과 AI & HCI 실험실

A Study on Speech Recognition System Using Continuous HMM

o
Sang-duck Kim*, Geuk Lee*

*AI & HCI Lab, Department of Computer Engineering, Hannam University

요 약

본 논문에서는 연속분포(Continuous) HMM(hidden Markov model)을 기반으로 하여 한국어 고립단어 인식 시스템을 설계, 구현하였다. 시스템의 학습과 평가를 위해 자동차 항법용 음성 명령어 도메인에서 추출한 10개의 고립단어를 대상으로 음성 데이터 베이스를 구축하였다. 음성 특징 파라미터로는 MFCCs(Mel Frequency Cepstral Coefficients)와 차분(delta) MFCC 그리고 에너지(energy)를 사용하였다. 학습 데이터로부터 추출한 18개의 유사 음소(phoneme-like unit : PLU)를 인식단위로 HMM 모델을 만들었고 조음 결합 현상(co-articulation)을 모델링 하기 위해 트라이폰(triphone) 모델로 확장하였다. 인식기 평가는 학습에 참여한 음성 데이터와 학습에 참여하지 않은 화자가 발성한 음성 데이터를 이용해 수행하였으며 평균적으로 97.5%의 인식성능을 얻었다.

1. 서론

최근 들어 디지털 신호처리 기술과 통신기술 및 컴퓨터 처리능력의 급속한 발전과 더불어[1] 인간과 기계간의 의사소통을 보다 자연스럽고 정확하게 하는 man-machine interface 기술이 현실적 문제로 부각되고 있는 가운데, 음성인식 기술의 중요성이 대두되고[2] 실생활에 적용하기 위한 연구가 진행중이다.

가장 일반적으로 사용되는 음성인식의 방법으로 HMM(hidden Markov model)이 있다[3]. HMM은 음성인식을 위한 방법들 가운데에서 현재 가장 우수한 성능을 보이는 인식방법으로, 통계적인 특성으로 인해서 음성 인식결과의 이후처리에 수반될 언어 처리나 의미 처리 등의 통계적인 모델들과 잘 융합하는 장점을 가지고 있는 현재 가장 널리 사용되고 있는 방법이다[4].

HMM은 은닉(hidden)되어 있는 상태열(state sequence)의 통계적 특성을 벡터로 이루어진 관측열(observation sequence)의 통계적인 특성을 통하여 추정하는 이중의 통계적 프로세스이다.

HMM은 관측벡터의 출력확률분포를 모델링 하는 방법에 따라 이산분포(Discrete) HMM, 준연속분포(Semi-Continuous) HMM, 연속분포(Continuous) HMM으로 분류된다.

본 논문에서는 Gaussian mixture 연속확률밀도에 기반한 연속분포 HMM을 이용한 한국어 고립단어 인식 시스템의 설계 및 구현에 대해 소개한다.

2. 연속분포 HMM

연속분포 HMM에서는 음성 신호에서 추출한 특징 벡터를 그대로 사용한다. 모델의 파라미터를 추정하기 위해 관측된 신호들이 mixture Gaussian 프로세스에 의해 발성되었다는 가정 하에 maximum likelihood 방법을 사용한다. Gaussian 혼합 밀도 함

수가 어떠한 연속확률밀도도 근사화할 수 있다는 장점을 지니고 있어 남녀의 음성, 주파수 차이 등의 음성에서의 다양한 변이성을 모델링 하기에 적합하다[5].

연속분포 HMM의 경우 상태 j, 시간 t에서 입력 벡터 O_t를 관측할 확률은 다음 식(1)과 같이 표현된다.

$$b_j(O) = \sum_{k=1}^{M} c_{jk} N(O_t, \mu_{jk}, U_{jk}), \quad 1 \le j \le N \qquad (1)$$

여기서 M은 mixture의 수, c_{jk}은 상태 j에서 k번째 mixture에 대한 가중치(weight), $N(O_t, \mu_{jk}, U_{jk})$는 Gaussian 확률밀도 함수이다. μ_{jk}와 U_{jk}는 각각 상태 j에서 k번째 mixture의 평균벡터(mean vector)와 공분산 행렬(covariance matrix)이고 N은 전체 상태(state)수이다.

$b_j(O)$를 학습 데이터로부터 구하기 위한 Baum-Welch 재추정(re-estimation) 알고리즘은 다음과 같다.

상태수를 N, 심벌의 길이를 T, 전향확률을 $\alpha_i(t)$ $(i=1,2,\cdots,N;\ t=1,2,\cdots,T)$, 후향확률을 $\beta_i(t)$ $(i=1,2,\cdots,N;\ t=T,T-1,\cdots,0)$라하고, 시간 t에서 상태가 j일 확률 $\gamma_t(j,k)$를 다음 식(2)와 같이 정의한다.

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \right] \left[\frac{c_{jk} N(O_t, \mu_{jk}, U_{jk})}{\sum_{m=1}^{M} c_{jm} N(O_t, \mu_{jm}, U_{jm})} \right] \qquad (2)$$

위의 세 가지 확률변수 (α, β, γ)를 이용하여 모델 파라미터 $\overline{c_{jk}}$, $\overline{\mu_{jk}}$, $\overline{U_{jk}}$를 추정하는 식은 다음과 같다.

$$\overline{c_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \qquad (3)$$

$$\overline{\mu_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot O_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \qquad (4)$$

$$\overline{U_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (O_t - \mu_{jk})(O_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \qquad (5)$$

3. 음성 데이터 베이스

본 인식 시스템을 학습하고 평가하기 위해 사용된 음성 데이터 베이스는 자동차 항법용 음성 명령어 도메인에서 선택한 10개의 명령어를 대상으로 조용한 사무실 환경에서 PC sound-card를 이용해 녹음하였고, 16kHz waveform으로 sampling 하였으며, signed 16 bit로 저장하였다.

자동차 항법용 음성 명령어			
1. 전진	2. 후진	3. 저속	4. 중속
5. 고속	6. 가속	7. 감속	8. 정지
9. 좌회전	10. 우회전		

[표 1] 발음 목록

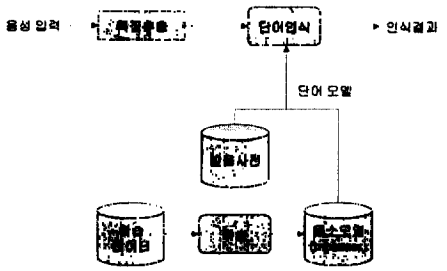
음성 데이터 베이스는 11명(학습용 5명, 평가용 6명)의 남성화자가 [표 1]에 나타난 발음 목록의 10개의 음성 명령어를 5번씩 발성한 550개의 독립된 단어들로 이루어 졌으며 구성은 [표 2]와 같다.

구분	학습용	평가용
화자수	5	6
발성수	250	300

[표 2] 음성 데이터 베이스의 구성

4. 시스템 구성

본 고립단어 인식 시스템은 화자 독립 시스템으로서 Entropic사의 HTK V2.1을 기반으로 구축하였다. 전체 시스템 구성은 크게 특징추출 과정, 학습 과정, 인식 과정으로 이루어 졌으며 시스템의 전체 구성도는 다음 [그림 1]과 같다.



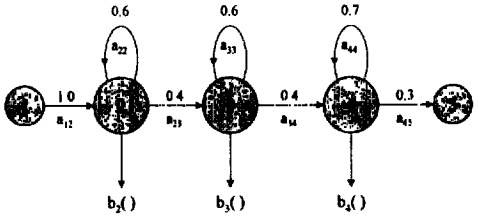
[그림 1] 시스템 전체 구성도

4.1. 특징추출

본 인식기에서는, 입력된 음성 신호는 전처리 과정을 통해 16kHz/16bit waveform으로 A/D 변환 과정을 거치며 음성 특징을 추출하기 위해 음성의 해석구간을 10ms씩 이동하면서 20ms길이의 프레임으로 분할한다. 각 프레임의 음성신호에 해밍 윈도우(Hamming window)를 씌워 음향적 특징추출에 사용하고 있다. 이를 기반으로 각 프레임별로 고속 푸리에 변환(FFT) 분석을 거쳐 12차의 MFCCs(Mel Frequency Cepstral Coefficients)를 추출하여, 전후 2개씩의 프레임을 참조하여 12차의 차분(delta) MFCC, 에너지(energy) 및 차분 에너지 등 총 26차의 특징 파라미터를 추출하여 인식에 사용하였다.

4.2. 학습과정

초기 HMM 모델 학습을 위한 단위로 유사 음소(phoneme-like unit : PLU) 18개를 설정하였다. 초기(protype) 모델의 위상(topology)은 [그림 2]와 같이 음소 기반(phone-based) HMM 음성 인식시스템에서 널리 사용되고 있는 자신 또는 다음 상태로만 천이 하는(skip이 없는) 3 state left-to-right model로 정의하였다.



[그림 2] Simple Left-Right HMM

모델의 각 상태에 대해 Gaussian 연속확률밀도 함수에 기반 하여 26개의 평균(mean)과 공분산(covariance)을 추출하여 파라미터로 사용하였다.

학습 데이터로부터 추출된 17개의 유사음소와 여기에 sil(silence)모델 1개를 합한 총 18개의 HMM 모델은 모든 평균과 공분산이 같도록 초기화된다.

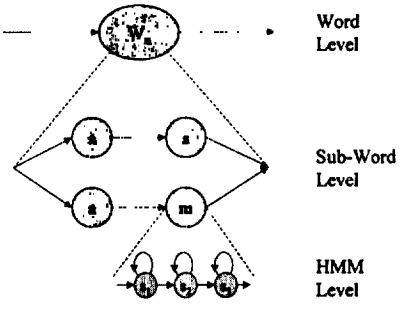
HMM 모델의 학습은 Baum-Welch 재추정 알고리즘에 의해 수행되어진다. 우선 학습 데이터를 상태수에 따라 균일 분할한 후 k-means 클러스터링(clustering) 알고리즘을 사용해서 각 상태에 해당하는 클러스터(cluster)의 평균 가중치(mixing weight) 등을 계산하여 파라미터를 초기화한다.

조음 결합 현상(co-articulation)을 모델링하기 위해 앞서 만들어진 유사음소 HMM모델을 문맥의존(context-dependent) 트라이폰(triphone) HMM 모델로 확장하고 다시 파라미터 재추정 과정을 거쳐 최종적인 HMM 모델을 생성한다.

4.3. 인식과정

인식 과정에서는 학습 과정에서 만들어진 HMM 모델을 가지고 평가 데이터를 이용해서 인식실험을 하게 된다. 이때, 고립단어 인식에 대한 문법 네트워크(grammar network) 정보가 필요하다. 본 시스템에서는 영역의존 의미문법을 컴파일 하여 FSN(Finite State Network)을 구성하였다.

인식 네트워크(recognition network)은 아크로 연결되어진 노드들의 집합으로 구성은 [그림 3]과 같다.

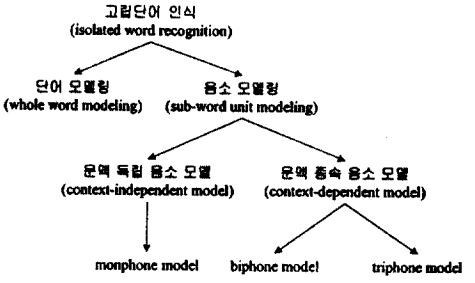


[그림 3] 3 계층 네트워크 구조

각 모델 노드는 그 자체가 아크에 의해 연결되어진 3-상태로 구성된 네트워크이다. 최상위 계층은 단어 레벨(word level)로 영역 의존 의미 문법을 파싱한 결과이고, 중간 계층은 각 단어가 어떤 음소로

이루어지는 가를 나타낸다. 최하위 계층은 3-상태로 구성되어진 각각의 음소 HMM을 나타낸다.

고립단어 인식에서 인식 단위는 [그림 4]와 같이 크게 단어 모델링(whole word modeling)과 음소 모델링(sub-word unit modeling)으로 나누어지는데 본 논문에서는 인식 단위로 음소 모델링 중 앞뒤에 오는 음소를 모두 고려한 문맥 종속 트라이폰 모델을 사용하였다.



[그림 4] 고립 단어 인식 단위

트라이폰 모델은 음소 모델보다 단어 내의 음운 현상을 효과적으로 반영하는 장점이 있지만, 음운 조합의 보다 다양한 표현으로 인해 그 수가 과도하게 많아지게 되므로 신빙성이 있는 모델 파라미터를 추정하기 위해서는 각 트라이폰 모델 당 어느 정도의 학습 데이터가 확보되어야 한다. 이러한 데이터 부족 현상에 대한 해결 방법으로 문맥이 같은 모든 트라이폰 모델은 전이확률(transition probability)과 상태 파라미터를 공유하게 하는 tied-state 트라이폰을 만들어 준다. 이때 파라미터의 특성을 공유하게 하는 것은 각 모델간의 변별력을 떨어지게 하는 요인이 되므로 모델간의 변별력에 크게 영향을 미치지 않도록 조정하는 것이 중요하다.

본 논문에서 구현한 고립단어 인식 시스템에서는 트라이폰의 수가 그다지 많지 않아 state tying 과정을 거치지 않고 바로 인식 실험을 수행하였다.

인식 시간의 단축을 위해 인식 과정에서 확률 값을 계산 할 때 최대의 확률을 가지는 경로(maximum likelihood state sequence)의 확률 값 보다 특정 임계값(threshold) 밑으로 떨어지는 경로에 대해서는 탐색을 중지하도록 하는 Viterbi beam search를 수행하였다. 본 논문에서는 임계값을 200으로 주었다.

입력된 미지의 단어를 인식하기 위해서는 이 단어가 각 모델에서 생성되는 likelihood 값을 계산한다. 그 결과 가장 높은 likelihood 값을 가지는 모델이 인식된 단어로 결정된다.

5. 실험 및 결과

인식기 성능 평가는 [표 3]과 같이 학습에 사용된 데이터와 학습에 참여하지 않은 화자가 발성한 데이터로 나누어 각각 수행하였다.

데이터 1	데이터 2
학습에 참여한 데이터	학습에 참여하지 않은 화자가 발성한 데이터

[표 3] 평가용 음성 데이터

평가용 음성 데이터에 대한 본 시스템의 인식 결과는 [표 4]와 같다.

구분	데이터 1	데이터 2	average
Correct(%)	100	95	97.5
Accurate(%)	100	94.3	97.15
Deletion	0	0	0
Substitution	0	15	7.5
Insertion	0	2	1
Number	250	300	total 550

[표 4] 인식 실험 결과

학습에 참여한 음성 데이터를 이용한 인식 실험 결과는 총 250개의 단어 중 250(100%)개 모두 정확하게 인식하였다.

학습에 참여하지 않은 화자가 발성한 음성 데이터를 이용해 인식 실험을 한 결과 총 300개의 단어 중 285개(95.00%)를 인식, 이중 283개(94.33%)를 정확하게 인식하였다.

[표 4]에서 Correct는 평가용 단어 중 정확히 인식된 단어의 인식률이며, Accurate는 Correct에서 불필요하게 추가된 단어의 수를 제외한 인식률이다.

Correct와 Accurate는 다음 수식을 통해 계산되어진다.

$$Correct = \frac{N-S-D}{N} \times 100\% \quad (6)$$

$$Accurate = \frac{N-S-D-I}{N} \times 100\% \quad (7)$$

위 [표 4]에서 나타나듯이 학습에 참여한 데이터의 인식률이 학습에 참여하지 않은 데이터의 인식률보다 약 5% 정도 우수함을 알 수 있다. 그 원인으로는 학습용 데이터와 평가용 데이터 사이에 마이크를 포함한 녹음 환경의 차이, 화자 변화에 따른 발성의 차이, 화자독립모델과 화자종속모델의 차이 등이 있는데 이는 인식률에 직접적인 영향을 미친다.

따라서 이에 대한 해결방법으로 소량의 학습 데이터를 사용하여 충분한 학습데이터로 학습된 화자 독립모델을 인식시스템을 사용하려는 특정화자에 적응시키는 화자적응(speaker adaptation)[6] 기술의 적용이 필요하다.

6. 결론 및 과제

본 논문에서는 연속분포 HMM을 기반으로 한국어 고립단어 인식 시스템을 설계, 구현하였다. 고립단어 인식 시스템 설계를 위하여 자동차 항법용 음성 명령어 도메인에서 10개의 단어를 선정하고 이를 대상으로 sil을 포함한 18개의 유사음소를 추출하였다. 이 서브 워드(sub-word)를 단위로 하여 유사음소 HMM 모델을 생성하였고, 단어 내의 조음 결합 현상을 모델링 하기 위해 트라이폰 모델로 확장하였다.

5명의 남성화자가 발성한 250개의 화자 독립 고립단어 음성을 이용해 유사음소 HMM 모델을 학습시켰다.

학습에 참여한 250개의 음성 데이터와 학습에 참여하지 않은 화자가 발성한 300개의 음성 데이터를 이용해 인식기 평가 실험을 수행하였는데 전자는 100%의 인식률을 보였고 후자는 95%의 인식률을 보였다.

본 논문에서 구현한 인식 시스템은 잡음 환경을 고려하지 않았다. 따라서 본 시스템을 자동차 환경과 같은 실생활에 적용하기 위해서는 주위의 가변적인 잡음환경에 강인한 시스템을 구축해야 한다.

나아가서는 자연스럽게 발성한 대화체 음성으로부터 미리 정해진 핵심어(keyword)들을 검출해 내는 핵심어 검출(keyword spotting) 기술을 이용하여 특정 도메인에 종속되지 않는 대화체 연속 음성을 인식할 수 있는 stand-alone의 음성인식 시스템에 대한 연구가 요구된다.

7. 참고 문헌

[1] Mariani, "Recent advances in speech processing," Proc. of Int. Conf. on Acoust Speech, and signal Processing, pp. 429-440, Glasgow, May. 1989

[2] 김광태, 서정일, 김현기, 홍재근,"ARHMM에서의 화자적응," 한국정보처리학회 추계 학술발표 논문집, 제 4 권, 제 2 호, pp. 1184-1188, 1997.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[4] 최환진, 상태 의존 모델링과 분별력 학습을 이용한 연속 음성 인식, 한국 과학 기술원, 박사학위 논문, 1997.

[5] 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 Technical Journal, Vol. 64, No. 6, pp. 1211-1234, July-August 1985.

[6] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," IEEE Trans. on Signal Processing, vol. 39, no. 4, pp. 806-814, Apr. 1991.

[7] 김형순, "핵심어검출 기술 및 그 응용," 음성통신 및 신호처리 워크샵, 제 13 권, 제 1 호, pp. 39-44. 1996.

[8] 위선희, Two-level 모델 기반의 음성 사전 생성기를 이용한 한국어 HMM 연속 음성 인식 시스템 개발, 서강대학교, 석사학위 논문, 1996.

[9] 이항섭, 박준, 권오욱, "한국어 대화체 인식 시스템의 구현", 음성통신 및 신호처리 워크샵, 제 13 권, 제 1 호, pp. 145-148. 1996.