연속 HMM에 따른 우리말 음성인식 조사

임창욱°, 신좌철, 김석동
호서대학교 컴퓨터공학부

The Study of Korean Speech Recognition for Various Continue HMM

Changwug Lim°, Chwacheul Shin, Sukdong Kim
carrotez@freechal.com, scc9977@korea.com, sdkim@office.hoseo.ac.kr

요약

본 논문은 연속 밀도 함수를 갖는 HMM별 한국어 연속 음성 인식에 관한 연구이다. 여기서 우리는 밀도 함수가 2개에서 44개까지 갖는 연속 HMM모델에서 가장 효율적인 연속 음성 인식을 위한 방법을 제시한다. 음성 모델은 36개로 구성한 기본음소를 사용한 CI-Model과 3,000개로 구성한 확장음소를 사용한 CD-Model을 사용하였고, 언어 모델은 N-gram을 이용하여 처리하였다. 이 방법을 사용하여 500개의 문장과 6,486개의 단어에 대하여 화자 독립으로 CI Model에서 최고 94.4%의 단어 인식률과 64.6%의 문장 인식률을 얻었고, CD Model에서는98.2%의 단어 인식률과 73.6%의 문장인식률을 안정적으로 얻었다.

1. 서론

음성 처리 기술과 컴퓨터 과학의 발달로 컴퓨터에 의한 음성인식을 실생활에 응용하려는 시도가 현재 다양하게 이루어지고 있다.90년대 중반부터 본격적으로 연구가 이루어지고 있는 대용량 연속 음성 인식 분야는 선진 여러 나라에서는 현재 활발히 연구되고 있다. 영어권에서는 Wall Street Journal을 활용하여 실용화 단계까지 이르고 있으며 프랑스에서는 Le Monde, 독일에서는 Frunkfurter Rundschau, 이탈리아는 Sole 24 Ore, 일본에서는 Nihon Keizai 신문을 활용하여 이 분야에 대한 연구가 상당히 진척되고 있다.[1][2][3][4][5] 그러나 우리말의 본격적인 연구가 국내에서는 아직 이루어지고 있지 않는 실정이다. 그 이유는 우리말은 영어와는 다르게 단어사이에 공백이 없어 문장 내에서 단어를 자동적으로 찾기가 어렵기 때문이다. 그러므로 대용량 연속 음성 인식에 있어서 중요한 역할을 하는 통계적 언어 모델의 사용을 위해서는 우리말을 다양한 방법으로 처리할 필요가 있다.

대규모 어휘를 가진 연속음성의 인식은 상당히 어렵다. 그 이유는 첫째, 단어 구간이 불확실하다. 그 결과 많은 잘못 가정된 단어가 종종 만들어 진다.그래서 의미 정보나 단어 문맥을 제공하는 복잡한 언어 모델을 사용하여 여러 개의 가정 중에서 가장 그럴듯한 것을 선택하는 것이 필요하다. 둘째, 상호 조음 효과가 매우 강해서 어느 순간의 음성은 앞뒤의 음성에 영향을 많이 받는다. 이것들을 다루기 위해서는 Di-phone, Tri-phone과 같이 문맥 정보가 고려된 보다 정교한 음성 모델이 필요하다. 언어모델과 음성 모델이 복잡해질수록 작업의 난이도가 커지게 되어 컴퓨터의 계산능력이나 기억능력을 초과하기도 한다. 인식속도, 필요한 기억 장치의 자원과 인식률과 같은 세 가지 종류의 성능은 서로 상충된다. 예를 들면 탐색 공간을 줄여 인식속도를 늘리고 간단한 음성과 언어 모델을 이용해 기억용량을 줄이면 인식률이 떨어진다. 즉 동시에 높은 인식률을 유지하면서 인식속도를 증가시키고 기억 용량을 줄이는 문제는 상당히 힘들다.

본 논문에서는 Viterbi Beam탐색과 Stack decoding을 결합한 다단계의 탐색 방법으로 인식을 하였고, 음성 모델은 Uni-phone단위와 Tri-phone단위로 Continuous HMM방법을 사용하

였으며, 언어 모델은 통계적인 방법인 N-grams을 사용하였다.[6][7][8]

2. 사용한 음성 모델

각각의 HMM 상태에서의 emitting은 출력확률 분포와 관계가 있다. 이 분포는 그 상태에서 생성되는 관찰의 정도를 결정한다. 이러한 분포는 서로 다른 종류의 음성을 구분해야하고 자연스러운 음성 고유의 편차를 내포해야한다. 초기에는 이산 분포를 사용했으나 근래에 와서 실시간 인식을 위해 시간과 기억장소의 절약을 하기 위해서를 제외하고는 사용되고 있지 않다. Continuous나 Semi-continuous분포를 주로 사용하는데 Gaussian혹은 Gaussian 확률 밀도 함수의 혼합으로 표현한다.

$$b_{jm}(o_t) = N(o_t; \mu_{jm}, \Sigma_{jm})$$

$$= \frac{1}{\sqrt{(2\pi)^n |\Sigma_{jm}|}} e^{-\frac{1}{2}(o_t-\mu_{jm})'\Sigma_{jm}^{-1}(o_t-\mu_{jm})}$$

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} b_{jm}(o_t). \quad (1)$$

여기서 c_{jm}, μ_{jm}과 Σ_{jm}은 각각 상태 j에서의 혼합 Gaussian 분포의 m번째 성분에 대한 weight, mean과 covariance를 말한다. 여기서 $b_{jm}(o_t)$ 함수를 고정된 개수의 표준함수를 사용하게 되면 Semi-continuous분포를 갖는 모델이 되며, 개수를 제한하지 않는 비표준함수를 이용하는 것이 Continuous 모델이 된다. 하나의 상태에 대하여 각각의 Gaussian함수와 weight 벡터가 대응된다. 상태의 개수를 결정하는 또 다른 요소는 HMM모델에 대응하는 음소의 개수이다. 음소는 기본음소(uni-phone)와 확장음소(tri-phone)로 구성된다. 기본음소의 경우 영어는 보통 50개, 일본어는 42개를 사용한다. 본 연구에서는 우리말 기본음소를 36개로 하였다. 예를 들어 종성의 'ㅇ' 의 경우 본인이 연구한 결과, 기본 음소로만 구성되는 음성모델로 인식 실험을 한 결과가 중성모음에 따른 'ㅇ' 을 구분한 것이 구분하지 않는 것보다 인식률의 높음을 확인할 수 있었다.

3. 실험 및 결과

음성을 16 KHz, 16-bit로 sampling하였다. 첫 번째로 프레임 크기를 10 msec로 정하였다. 각 프레임 마다 12개의 mel-scale 주파수 캡스트럼 벡터와 하나의 power 계수를 구했다. 또한 캡스트럼 계수와 Power계수를 각각 1, 2차 미분을 하여 각 프레임마다 4가지 종류의 특징 벡터들을 사용했다.

continuous HMM은 각각의 음소에 대하여 [표 1]과 같이 여러 종류의 확률밀도를 가지고 있다 실험에 사용한 밀도 함수 개수는 12종류를 사용하였다. 한 밀도 함수에 대하여 uni-phone model로 구성한 Context Independent Model과 Tri-phone Model로 구성한 Context-Depedent Model을 적용하여 총 24개 모델을 학습시키고 인식을 하였다. CI(Context Indepedent)는 36개의 기본 음소를 사용하였고 CD(Context Depedent)는 기본 음소를 포함하여 총 3,000개의 음소 모델을 사용하였다.

Model Type	확률 밀도 함수 개수
model(12)	2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44

표 1. 확률 밀도 함수 개수

3.1 학습 및 인식 음성 자료

음소 모델 학습에 사용한 음성 데이터는 162명이(남자:92명, 여자:70명) 각기 30분에서 2시간 정도로 발음한 것으로 읽은 자료는 우리말의 음소가 모두 들어있는 음성이 되도록 국내의 신문사 Web site에서 무작위로 발췌하였다. 우리말 음성 모델을 만들기 위한 학습 시간은 평균 20시간/Iteration 정도였고 밀도 함수가 작은 모델은 시간이 단축되고 밀도함수가 큰 모델은 시간이 더 많이 걸렸다. 총 12개 모델에 대하여 한 모델당 5번의 반복을 수행하였다.

인식 데이터로는 학습에 참여하지 않은 음성으로 ETRI에서 제공했던 연속 음성인 PBS 데이터베이스를 사용하였다. 500개 문장 총 6,486개 단어를 대상으로 인식을 행하였다.

3.2 언어 모델 자료

언어 모델을 만들기 위한 기초 자료 역시 인터넷을 통해 한국의 신문사 및 방송사의 자료를 수집하였다. 우선 기본 언어 모델을 만들기 위

해 일반적인 단어가 들어 있는 신문기사와 방송사 뉴스대본을 포함하여 62,824단어로 구성된 총10,778문장을 수집하였다. 또한 6,486개 단어로 구성된 인식 대상 문장 500문장을 포함하여 총 11,278문장으로 3-gram을 이용하여 언어모델을 구성하였다. 인식대상 단어에 대한 언어모델의 Perplexity는 15.2였다.

3.3 인식 결과

단어 인식률은 전체 6,486개를 대상으로 오인식 단어수를 계산하였고, 문장 인식률은 한 문장에서 오인식 단어가 1개가 있어도 오인식으로 처리하였다. 오인식은 3가지 종류로 나누었다. 다른 단어로 인식된 경우, 단어가 삭제된 경우, 추가된 경우 등 3가지 오인식으로 나누었으며 다른 단어로 인식된 경우가 전체 오류의 대부분을 차지하였고 삭제, 추가 순으로 나타나고 있다.

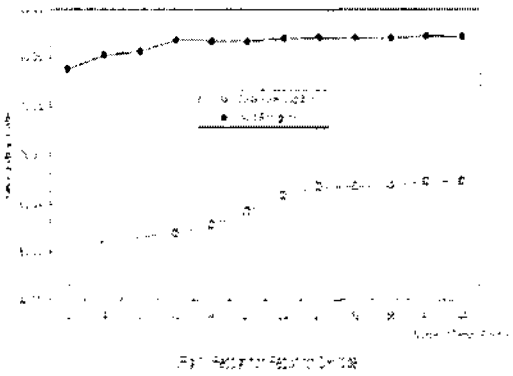

그림 1. CI Model의 인식 결과

CI Model인 경우 [그림1]에서 보는 바와 같이 단어 인식률은 확률 밀도 함수의 개수가 2개 일 때 87.7%이며 밀도 함수의 개수가 증가 할수록 인식률이 증가되나 24개 이상은 증가율이 거의 없다. 그러나 문장 인식률에서는 40개 까지 증가하는 것을 알 수 있다. 따라서 단 음소(uni phone)로 인식을 할 경우 확률 밀도 함수의 개수가 40개가 가장 좋은 인식을 얻을 수가 있었다. CD Model인 경우 [그림 2]에서 보는 바와 같이 단어 인식률은 확률 밀도 함수의 개수가 2개 일 때 97.4%이며 밀도 함수의 개수에 크게 영향을 받지 않아 8D 이상은 증가율이 거의 없다. 문장 인식률에서도 인식률 편차가 CI에 비해 크지 않으며 28D이상 거의 인식률 차이가 없었다.

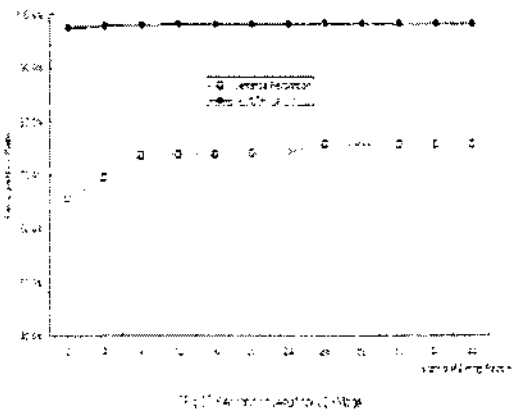

그림 2. CD Model의 인식 결과

실험에 참여한 CI 전체 모델에 대하여 평균 단어인식률은 92.9%이며 평균 문장인식률은 58.2%이다. CD 전체 모델에 대하여 평균 단어인식률은 98.1%이며 평균 문장인식률은 73.6%이다. CD Model의 인식률이 CI에 비해 단어 인식률은 평균 10%, 문장 인식률은 평균 15%정도 인식률이 높았다.

Density F.	CI Model(MB)	CD Model (MB)
2d	0.1	10
4d	0.3	19
8d	0.5	38
12d	0.8	58
16d	1.0	77
20d	1.3	96
24d	1.6	115
28d	1.8	134
32d	2.1	154
36d	2.3	173
40d	2.6	192
44d	2.9	211

표 2. 자원 크기

표2 에 나타난 바와 같이 인식효율이 높은 CI가 40D인경우 2.5M Byte인 반면 CD 28D 인 경우 134M로 70배 정도 큰것을 알 수 있었다. 표에있는 Resource는 mean, variance, transition matrix, mix weight를 모두 합한 값이다.

4. 결론

효율적인 우리말 연속 음성 인식을 위해 연속 HMM 모델을 적용하여 연속 밀도 함수별 인식률을 살펴보았다. CI-Model에 대하여 단어 인식률은 밀도함수 개수가 24개일때 가장 효율적인 모델이며, 문장 인식률은 40개일때 효율적인 모델임을 알 수가 있었다. CD-Model에 대하여 단어 인식률은 밀도함수 개수가 24개일때 가장 효율적인 모델이다. CD Model의 인식률이 CI에 비해 단어 인식률은 평균 10%, 문장 인식률은 평균 15%정도 인식률이 높았다. 반면 자원의 크기는 CD Model이 CI에 비해 70배 정도 크다. 보다 효율적인 모델을 발견하기 위해서 다양한 연구가 앞으로 필요하다.

참고문헌

1. D. B. Paul and J. M. Baker, 'The Design for the Wall Street Journal-based CSR corpus," Proc. ICSLP-92, pp. 899-902, 1992

2. J. Gauvain, L. F. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large French read-speech corpus," Proc. ICSLP-90, pp. 1097-1100, 1990

3. H. J. M. Steeneken and D. A. van Leenwen, "MultiLingual Assessment of speaker independent large vocabulary speech-recognition systems: SQUALE Project," Proc. EUROSPEECH-95, pp. 1271-1274, 1995

4. L. Lamel, M. Adda-Decher, and J. L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," Proc. EUROSPEECH-96, pp. 185-188, 1996

5. T. Matsuoka "Large-vocabulary continuous-speech recognition using Japanes business newspaper (Nikkei)" DARPA Speech Recognition Workshop, pp. 137 - 142, Feb. 1997

6. Allev a,F., Huang, X., and Hw ang, M. "An Improved Search Algorithm for Continuous Speech Recognition." In IEEE International Conference on Acoustics, Speec h, and Signal Processing, 1993.

7. H. Ney, U.Essen, R.Kneser. "On Structuring Probabilistic Dependences in Stochastic Language Modelling." Computer Speech Language, Vol. 8, pp. 1-38, 1994,

8. Paul, Douglas B. "An Effcient A* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model." In Proceedings of DARPA Speech and Natural Language Workshop, pp 405-409, Feb. 1992,