1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

확률적 모델을 이용한 연속 숫자음 인식에 관한 연구

이주승*, 이성권*, 김순협*

* 광운대학교 컴퓨터공학과

A Study on Continuous Digits Speech Recognition using Probabilistic Models

Ju-Sung Lee*, Seong-Kwon Lee*, Soon-Hyob Kim*

* Kwangwoon University
jsleesr@explore.kwangwoon.ac.kr

요 약

본 연구는 음소 단위의 CHMM(Continuous Hidden Markov Model)을 이용한 한국어 연속 음성인식에 관한 내용이다. 연구실 환경에서 음성으로 전화를 걸기 위하여 연속 숫자음 인식을 수행하였다. ETRI 445 데이터를 사용하여 초기의 모델은 ML(Maximum Likelihood) 추정법을 이용하여 작성하였고 적응화를 위해 최대 사후 확률 추정법을 사용하였다. 연속 숫자음의 인식을 위하여 한국어 숫자음 음성의 음향학적 특성을 고려하여 발성 사전을 작성하였고, 음절 단위로 되어있는 한국어 숫자음의 모든 경우를 고려하여 복수개의 단어를 사전에 등록하였다. 또한 숫자음의 앞 뒤 연음현상을 고려하여 작성한 21 종류의 7자리 숫자음과 이를 음절 단위로 세그먼트한 숫자음을 DB로 사용하여 적응화를 수행하였다. 이의 효율성을 입증하기 위하여 ETRI에서 작성한 35종류의 4연속 숫자음 목록을 대상으로 인식실험을 수행하였다.

I. 서론

음성은 인간이 가지고 있는 기본적인 능력 중에서 가장 중요한 것 중 하나로서 우리가 속박감을 거의 느끼지 않고 자유롭게 구사할 수 잇는 가장 자연스럽고 효과적인 정보교류의 수단이라 말할 수 있다. 또 음성에 의해 표현되는 말은 인간과 인간사이의 의사소통의 수단으로서 뿐만 아니라 논리적으로 사물을 생각하는 경우에 있어서도 중요한 역할을 한다. 이 음성이 인간과 기계와의 통신, 즉, 정보의 교환수단으로도 사용되어 오고 있다. 이러한 점에서 음성을 자동으로 인식할 수 있다면 우리의 일상생활에 큰 변화를 가져다 줄 것이며 우리의 삶을 더욱 윤택하게 하는 요소가 될 수 있다. 이와 더불어 개인용 컴퓨터의 보급의 가속화, 컴퓨터에 의한 신호처리 기술과 정보처리 기술의 급속한 발전과 더불어 음성을 통한 인간과 기계와의 직접적인 커뮤니케이션을 위한

Man-Machine Interface의 중요성도 강조되고 있다. 또 음성을 통한 기계와의 보다 편한 정보 전달에 대한 욕구는 계속되어 오늘날 가시화 되어가고 있으며 연구도 활발히 진행되어 가고 있다. 컴퓨터의 발전과 통신을 이용한 정보 및 금융 서비스가 확대됨에 따라 전화번호, 주민등록번호, 각종 신용카드의 비밀번호, 통장번호 등 많은 분야에서 무제한의 연속 숫자열에 대한 인식을 필요로 하고 있으며, 우리의 일상 생활에 밀접하게 사용될 때 많은 장점을 얻을 수 있다.

Ⅱ. 음성 데이터베이스

국외의 경우 음성 데이터베이스가 이미 공통으로 이용 가능하도록 구축되어 있으나, 국내의 경우에는 지금까지 각자가 부분적으로 제작하여 사용하였다. 따라서 데이터량 및 이용형태가 제한되고, 시스템의 성능 및 분석 방식의 평가도 객관적인 평가가 이루어지고 있지 않은 실정이다. 그러므로 본 연구에서는 초기 HMM 작성을 위한 음성 데이터를 한국 전자 통신 연구소에서 작성한 PBW(Phoneme Balanced Word) 445 단어 음성 데이터베이스 중 14인의 2회 발성 중에서 1회분 총 6,230 단어를 수작업에 의해 이루어진 유사 음소단위 레이블링 정보를 이용하여 구성하였고 적응화와 인식에 있어서는 무반향 방음부스에서 작성된 ETRI 445 DB 와는 달리 일반적인 연구실 환경에서 남성화자 5인에 의해 4회씩 발성된 데이터와 적응화를 위한 남성 화자 8명의 각 2회씩 발성된 데이터와 5인의 화자가 2번씩 발성한 단독 숫자음을 포함한 3180 단어의 세그먼트한 숫자음 데이터를 사용하여 실험하였다. 마이크는 SONY사의 콘덴서 마이크를 사용하였다. 인식 대상 단위는 본 연구실에서 앞 뒤 숫자음 음절을 고려한 21종류의 연속 숫자음을 화자 적응용으로 실험하였고 이의 적합성을 입증하기 위해 ETRI에서 만든 4연속 숫자음에 인식 실험하였다.

Ⅲ. 음성의 특징 파라미터 추출

실험에 사용된 데이터는 A/D 변환된 후 Pre-emphasis 필터를 통과한 후 16ms 길이의 해밍윈도우를 거쳐 구간 분석된다. 이때 각 구간은 5ms마다 중첩된다. 여기서 자기 상관 계수를 구하고 20차 LPC 계수를 구한 후 14차 LPC 캡스트럼 계수를 구하고 나서 10차의 회귀계수를 구하여 특징 파라미터로 한다. 여기에 음소 지속 시간정보를 추가로 이용하였다.

표 1. 35 종류의 연속 숫자음 표

ETRI 4 연속 숫자음(35단어)									
1	0287	8	5732	15	9601	22	4156	29	1199
2	1398	9	6843	16	0712	23	5267	30	6633
3	2409	10	7954	17	1823	24	6378	31	8877
4	3510	11	8065	18	2934	25	7489	32	2244
5	4621	12	9176	19	3045	26	8590	33	5500
6	6972	13	5861	20	3649	27	0316	34	7083
7	8194	14	9205	21	1427	28	2538	35	4750

Ⅳ. 숫자음 발음 사전 구성

한자리 숫자 단위로 인식할 경우에는 음소간의 조음 현상이 숫자를 구별하는데 큰 어려움을 주지 않지만, 연속어로 인식 범위가 확장될 경우에는 음소간, 단어간 조음 현상 때문에 단위숫자의 경계가 상당히 모호해진다. 또 우리말 숫자음은 숫자음간 혼동이 많은 단음절로 구성되어 있다. 다섯 개의 모음과 다섯 종류의 자음군으로 구성되어 있으며 연음이나 동시 조음 효과에 의하여 '일'과 '이' , '일'과 '칠', '삼'과 '사', '오'와 '구' 그리고 '구'와 '공'이 빈번한 오인식을 일으킨다.[표 2]

일반적으로 조음 단위 사이의 천이는 보통 묵음과 모음, 모음과 묵음, 자음과 모음, 모음과 자음, 모음과 모음 사이에서 발생하는데 자음과 자음 사이에서는 거의 음향학적으로 중요한 변이는 발생하지 않는 것으로 알려져 있다.

단음절로 구성되어 있는 숫자음은 초성 자음, 중성 모음, 종성 자음의 결합으로 이루어지고 종성 자음이 나타난 음절 뒤의 초성 자음이 존재하지 않고 바로 모음이 오는 경우 반드시 연음 또는 동시 조음 효과가 발생하게 된다. 이러한 경우의 모든 예를 가정하여 정상적인 발성 이외의 결과를 정리하면 [표 3]과 같이 나타낼 수 있다.

이러한 점을 고려하여 숫자음 발성 사전에 이러한 음절을 포함하여 구성하였다.

표 2. 우리말 숫자음에서의 상호 혼동음 및 그룹

일	↔	이	이 영	↔	영
일	↔	칠	이 이	↔	이
삼	↔	사	구 오	↔	구
오	↔	구	이 일	↔	일
구	↔	공	오 오	↔	오
			칠 일	↔	칠

표 3. 우리말 숫자음에서의 동시 조음 효과

Table 3. Syllables generation by co-articulation result

대상 숫자	1	2	3	4	5
발음 기호	/ih l/	/ih/	/s aa m/	/s aa/	/ao/
조음 결과	/m ih l/ /l ih l/ /g ih l/	/m ih/ /l ih/ /g ih/	/ss aa m/	/ss aa/	/g ao/ /l ao/
대상 숫자	6	7	8	9	0
발음 기호	/ju gp/	/ch ih l/	/p aa l/	/g~ up/	/g~ ao ng/
조음 결과	/n ju gp/ /l ju gp/ /ju ng/	/ch ih/	/p aa/	/gg up/ /k up/	/gg ao ng/

V. System 구성

[그림 1]은 이용한 HMM 모델은 4상태 3출력 1 혼합분포의 연속출력확률 이산 시간제어 HMM이다. [그림 2]는 전체 시스템의 구성도를 나타낸다. 초기의 HMM의 모델 작성은 ML 추정법을 사용하였으며, 적응화에는 MAP 추정법을 사용한다. 인식의 기본 단위로는 묵음을 포함한 48개의 유사 음소단위를 이용한다.

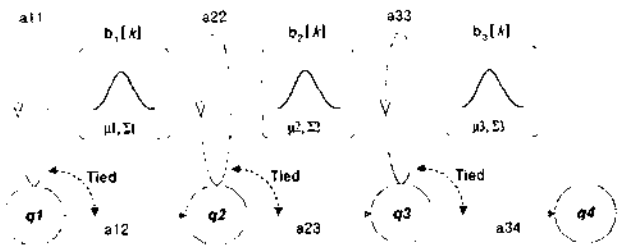


그림 1. 4상태 3출력 1 혼합분포 CHMM

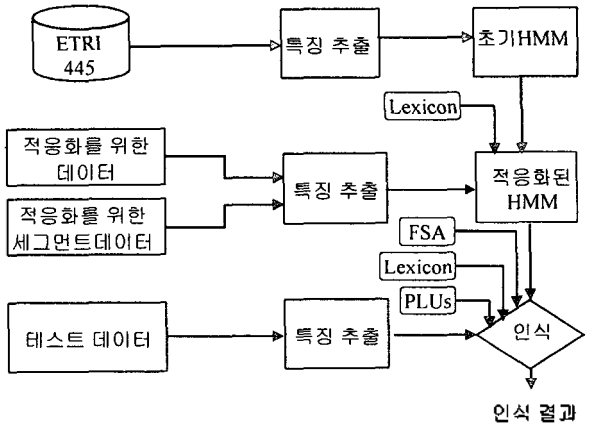


그림 2. 전체 인식 시스템

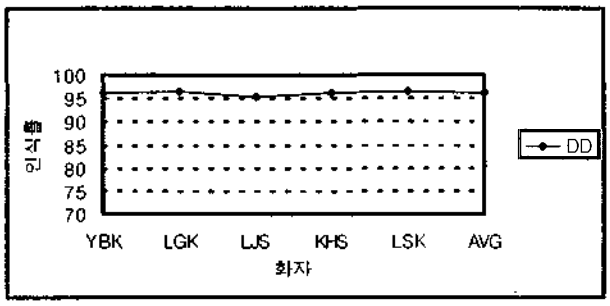
VI. 실험 및 검토

본 연구에서는 앞에서 언급했던 우리말의 숫자음 발성에 따른 음향학적 고찰을 통하여 숫자음 조합에서 발생될 수 있는 음운 현상을 고려하여 발음 사전에 등록하였다. 숫자음의 연음 및 조음 현상을 고려하여 작성한 21종류의 7자리 연속 숫자음 DB로 적응화를 하였고, 연속으로 발성되는 숫자음에 대해 처음에 위치한 숫자음과 우리말의 '에' 다음에 오는 숫자음은 선행하는 숫자음에 대해 영향을 받지 않는다는 가정을 하여 문법적으로 제한을 주었다. 이의 효율성을 입증하기 위하여 ETRI에서 작성한 35종류의 4연속 숫자음 목록을 대상으로 인식 실험을 하였다. 그리하여 연구실 환경에서 실시간으로 ETRI 4연속 숫자음에 대하여 약 96%의 인식률을 얻음으로 발성 사전의 구축에 대한 효율성을 증명하였다.

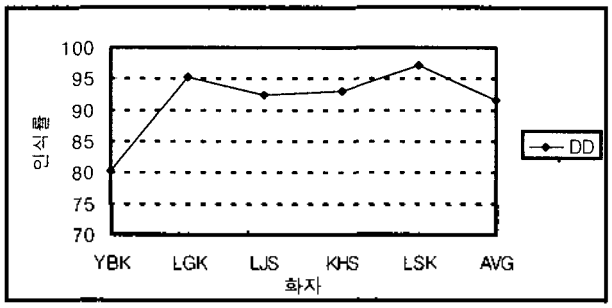
VII. 결론

7 연속 숫자음의 인식률이 특정화자에 대해서만 오차가 크게 나타나고, 4 연속 숫자음과 비교하여 높은 인식률을 보이고 있다. 그리고 숫자음에 대해서는 몇몇 특정 음소에 대해서만 학습이 이뤄지므로 다른 특정 응용분야에 비해 학습이 잘 되는 것을 볼 수 있다. 앞으로 더욱 많은

화자에 의한 학습이 이루어지고 전화 번호 인식이나 기타 여러 분야로의 확장, 우리말 숫자음에 대한 연구와 후처리 부분을 고려한다면 응용 분야에 다양하게 적용 가능할 것으로 보인다.



(a) 4 연속 숫자음 인식



(b) 7 연속 숫자음 인식

그림 3. 연속 숫자음 인식

[참고문헌]

[1] John R. Deller, Jr. John G. Proakis and John H. L. Hansen, "Discrete-Time Processing of Speech Signals", Macmillan Publishing Company, 1993

[2] SAEED V.VASEGHI "Advanced Signal Processing and Digital Noise Reduction",pp.111-139 Wiley Teubner

[3] J.K. Baker, "The DRAGON System - An Overview", IEEE Trans. Acoust. Speech, Signal Processing, ASSP-23(1), pp. 24-29, February 1975

[4] Kai-Fu Lee, Raj Reddy, "Automatic Speech Recognition",KLUWER ACADEMIC PUBLISHERS

[5] MIN ZHOU "A Study on Stochastic Models for Spoken Language"Ph.D Thesis Toyohashi Univ. 1996

[6] M. K. Ravishankar, "Efficient algorithms for speech recognition", Ph.D. thesis, Computer Science Department, Carnegie Mellon University, May 1996

[7] 오 세진, "문맥자유문법을 이용한 한국어 연속음성인식에 관한 연구", 영남대학교 대학원 석사학위 논문, 1997, 12

[8] Sadaoki Furui "Digital Speech Processing, Synthesis, and Recognition" , MARCEL DEKKER, INC.1991

[9] Douglas O'sShaughnessy "Speech Communication Human and Machine"Addison Wesley Publishing Company. 1987