통계적 모델에 의한 연속 숫자음의 인식 기술개발

이 강 성* 안 태 옥* 김 순 협**
*광운대학교 전자계산기 공학과 박사과정 **광운대학교 전자계산기 공학과 교수

Development of Continuous Spoken Digit Recognition System using Statistical Model

G. S. LEE, T. O. ANN S. H. KIM

요 약

본 연구는 통계적 모델에 의한 연속 숫자음의 인식에 관한 것으로, 4연속 숫자음을 인식 대상으로하여 실험한다. 시스템은 크게 음향 음성 처리부 및 어휘 해석부 두 부분으로 나뉜다. 음향 음성 처리부에서는 입력 음성으로부터 특징 벡터인 12차의 LPC cepstrum계수를 구하여, 프레임 레이블링과 소음소 레이블링(phone labelling)을 한다. 프레임 레이블링인 베이스 분류법을 이용하였으며, 소음소 레이블링은 프레임 레이블과 사후확률(posteriori probability)로 부터 이루어 졌다. 어휘 해석부분에서는 소음소 단위를 입력으로 받아 음운규칙을 통해 작성된 소음소 망을 거쳐 연속 숫자음 출력을 얻도록 했다.

본실험은 화자 3명이 발음한 35개의 4연속 숫자음을 인식 대상으로 하였으며, 4연속 숫자음을 평가단위로 80%의 인식률을 얻었고, 각 숫자음의 음절을 단위로 95%의 인식율을 얻어 제시한 알고리즘의 유효성을 입증하였다.

I 서론

연속음 인식의 적용분야중에서 숫자음은 가장 보편적인 대상어휘 이므로 반드시 해결해야 할 인식 대상이다. 따라서 본 실험은 연속 숫자음을 인식 대상어휘로 선정하였다. 우선 연속음 인식 대상이 결정 되었으면 인식의 기본단위(recognition unit)를 설정해야 하는데, 단어를 인식의 기본단위로 하는 것은 바람직하지 않다. [2] 숫자음은 음소의 수가 적고 어휘수가 적어 문맥에 따른 음질의 변화가 비교적 적다는 점에서 음소를 표준음성 추출의 기본단위로 설정했다. 하지만 음소나 변이음은 음의 변별적 특징 (distinctive feature)을 구별하기 위한 단위이므로 실제적으로 변화하는 음성현상과 일대일로 대응하지는 못한다. 따라서 음소사이의 과도음이나 각 음소의 변화음등의 음향현상과 대응되는 최소단위를 소음소(Phone)라 하고 본 논문에서는 이를 인식의 최소단위로 한다[1][3]. 즉, 다시말하면 음소단위로 추출한 표준음성을 이용하여 입력음성을 소음소(phone) 단위로 분리하고, 이를 연속 숫자음 인식의 기본 단위로 삼는 것이다.

본 논문에서의 전체적인 인식절차를 그림 1 에 보인다.

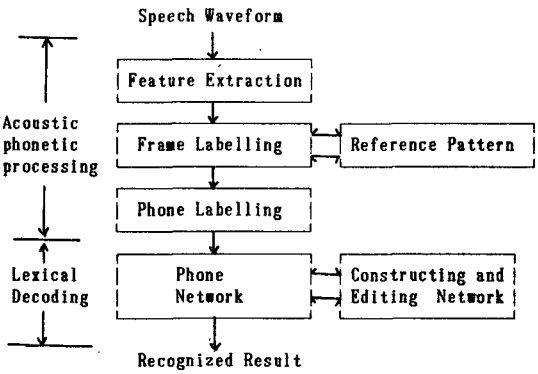

그림 1 인식 절차의 구성도.

본 논문에서는 통계적으로 구성한 음소 표준모델을 사용하는 베이스 분류법을 이용하여 각 프레임에 레이블링 한것을 소음소 단위로 레이블링하는 방법을 제시한다. 또한 소음소 레이블링에는 오류정보가 포함되기 마련인데 예기치 못한 음성의 오류는 치명적인 어휘해석의 오류를 낳게 하므로 이들 정보를 고려하는 것은 매우 중요하다. 이 문제의 해결을 위하여 소음소를 노드로 하는 소음소망(phone network)을 이용하여 어휘해석을 한다.

II 표준음소의 작성

1. 인식의 기본단위

숫자음인 경우에 고려해야할 대상음소 수가 데어휘인 경우에 비해 대폭 줄어든다. 본 실험에서 설정한 음소의 종류를 표 1에서 보인다.

설정된 표준 음소			
번호	분 류	음소	예
1	모 음	[i]	일, 이, 칠
2	(vowels)	[a]	삼, 사, 팔
3		[o]	공, 오
4		[u]	구
5		[yu]	육
6	자 음	[m]	삼
7	(consonants)	[ŋ]	공, 육육
8		[l]	일, 칠, 팔
9		[s]	삼, 사
10		[c^h]	칠
11		[p^h]	팔
12		[k]	구, 공
13	묵 음 (silence)	[q]	

표 1 표준음소의 종류.

3. 표준 패턴

각 음소 표준 패턴은 12차의 LPC cepstrum 계수의 평균 벡터, 공분산 행렬의 역행렬 과 행렬식 값으로 하였다. 이들은 Gaussian 확률 근사를 위해 필요한 정보로서 이후 각 음소에 대한 표준 파라미터로 사용된다.

4. 특징 추출

LPC cepstrum 계수를 특징 벡터로 한다. 다음에 특징 벡터를 구하는 절차 를 보인다.

(1) Filter (4.5 LPF)
(2) AD conversion (12bit, 10 KHz)
(3) pre-emphasis ($1-Z^{-1}$)
(4) Hamming windowing (12.8 ms window for every 6.4 msec step)
(5) LPC *cepstrum analysis (order 12)*

III 레이블링

1. 프레임 레이블링

각 프레임의 레이블링은 베이스 식별법 (Bayesian calssification method) [4] [5] 에 근거한다.

$$P(i|x) = \frac{P(x|i)\,P(i)}{\sum_{k=1}^{N} P(x|k)\,P(k)} \qquad \text{--------- (3.1)}$$

단, $P(x|i)$ 는 확률함수, $P(i)$ 는 사전(priori) 발생확률, N 는 표준 음소의 갯수 x 는 입력 벡터이다. 만약 확률 분포가 Gaussian 분포를 따른다면 위식의 $P(x|i)$를 다음식의 multivariate Gaussian 분포로서 근사할 수 있다.

$$P(x|i) = \frac{1}{(2\pi)^{n/2}\,|Vi|^{1/2}} \exp\{-\frac{1}{2}(x-\mu i)^t\, Vi^{-1}\,(x-\mu i)\} \qquad \text{--------- (3.2)}$$

이때 n은 차원 수, Vi는 공분산 행렬이며 μi는 평균 벡터이다. 또한 식 (3.1) 에서 P(i)는 음성 특징의 사전 확률 (priori probability) 을 나타내는데 모든 음성 특징에 대하여 같다고 가정하면 분모항은 모두 같은 값을 갖는다. 따라서 $P(i|x)$는 단지 $P(x|i)$에만 의존하게 된다.

그림 2에 베이스 분류법 으로 분류한 예를 보인다. 그림의 한 개의 점은 확률 1 을 17개의 gray level로 나타낸 것이며 진할수록 높은 확률을 나타낸다. 또 스펙트로그래프 와도 함께 비교한다.

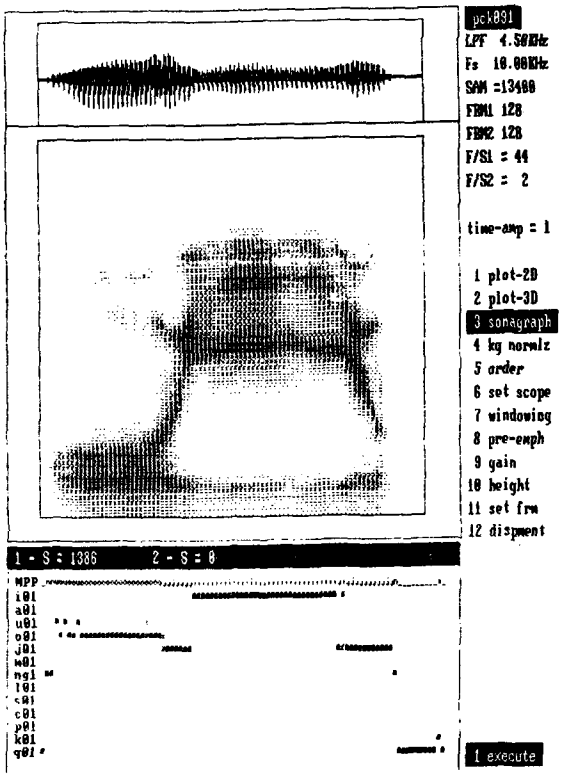

그림 2 스펙트로 그래프 와 프레임 레이블링. pck091.dat (5267) 의 a) 파형 b) 스펙트로 그래프 c) 베이스 분류법에 의한 프레임 레이블링

2. 소음소 레이블링 (phone labelling)

추출된 각 프레임마다에 대한 확률 분포는 종종 오류 정보를 포함하거나 불안정한 형태를 취하게 되는데 이를 일차적으로 구분하여 같은 음성적 성질을 갖는 영역으로 나눌 필요가 있다.

2.1 대략적인 구분화 작업 (Coarse Segmentation)

이 단계에서는 프레임 레이블을 이용하여 대략적인 변이음단위의 구분화 작업을 한다. 어느 점에서 시작하여 임계갯수(THn) 이상의 연속 프레임이 하나의 레이블을 갖는 구간으로 자르는 것이다. 설정된 구간내에 두개 이상의 레이블이 존재 할 때 이 구간을 안정구간이라하고 한개만의 레이블이 존재 할 때 안정 구간이라 한다.
흐름도를 그림 3에 보인다.

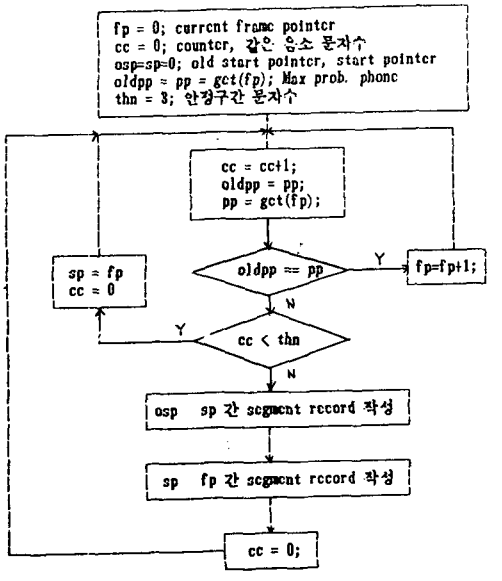

그림 3 대략적인 구분화 알고리즘 흐름도.

2.2 세분 절차 (Refine procedure)

앞단에서 얻어진 세그먼트 중에서 불안정한 것 만을 취하여 그것이 몇개의 소음소 구간으로 다시 구분될 수 있는지의 여부를 타진한다. 불안정 구간에 존재 할 수 있는 소음소 구간의 수는 대체로 파라미터 THn에 의존하는데 THn이 클수록 다수개의 구간이 존재 할 수 있다. 본 실험에서 설정한 것과같이 THn이 3 일경우 최대 2개의 구간이 존재한다고 보면 충분하다. 이 구간 분할의 문제는 Brandt's GLR method[6]의 기본 개념을 적용함으로서 해결하였다.

2.3 정정 (Correction)

이 단계에서는 설정된 소음소가 타당한지의 여부를 결정한다. 이 단계에서는 종종 불안정한 세그먼트(segment, phone)를 입력으로 받게되는데 이를 보정할 필요가 있다.

먼저 [p], [s], [c], [k] 등의 자음은 음부터 끝까지 같은 음성적 성질을 갖지 않으므로 하나의 자음이 여러개의 자음 세그먼트로 분류되는데 그 패턴은 대략 일정하다. 예를들어 /s/는 [c]+[s], [s]+[c]+[s], [c]+[s]+[p] 등의 패턴을 갖으며, /p/ 는 [s]+[p], /c/ 는 [p]+[c], [s]+[c], /k/ 는 [p], [p]의 패턴을 갖는다. 무성 자음구간의 세그먼트들을 위의 패턴과 시간적인 정보를 보조적으로 이용하면서 비교, 대치 시킨다.

두 번째로 [yu]의 세그먼트에 대해서 살펴보면, II 에서 살핀바와 같이 [yu]는 [i] 에서 [u] 로의 과도상태를 추출하여 만든 표준 음소이지만 시간적인 변화 파라미터는 포함 하고 있지 않고있다. 따라서 다음의 경우에 [yu] 가 나타날 수 있다.

1. 음소 [yu]의 경우
2. 한음소에서 다른 음소으로가는 과도상태
3. 불명료한 발음으로 인해 안정 상태인 구간이 잘못 분류되는 경우

위의 세가지 경우를 분리하는 방법으로는 분석적인 방법과 구문적인 방법을 생각할 수 있다. 분석적인 방법은 구간에 스펙트럼 변화와 [yu]의 시간변화를 고려한 표준패 턴과의 매칭거리를 파라메터로 이용한다. 구문적인 방법으로 는 [yu] 세그먼트의 주변환경을 고찰함으로서 [yu] 세그먼트의 상태를 결정하는 방법이다. 전자의 방법을 [yu] 세그먼트 성격규정을 위해 사용해 보았는데 몇가지 부 적당한 이유로 단지 구문적인 방법만을 적용하였다. 일차로 임계 프레임 갯수이하가 되는 세그먼트는 앞 세그먼트와 같은 세그먼트로 치환하고 [yu] 세그먼트 다음에 소음소 [q], [ ], [i], [o] 가 나오지 않는 세그먼트[yu]는 구문에 맞게 [yu]를 대치한다. 그리고나서 소음소망을 통과하는 어휘분석을 통한 인식 단계에서 [yu] 세그먼트가 '육'으로 치환될때 최소음절 길이와 비교해서 만일 '육' 음절 길이가 이것보다 작으면 실험적으로 표를 만들어 세그먼트의 성격을 판별한다. 이것이 과도 세그먼트로 판별될 경우 앞 세그먼트 에 포함시켰다.

세번째로 모음의 연접 (juncture) 에 관해 고찰해보면 숫자음에서 문제가 되는 연접은 [i],[o] 의 동일 모음 폐쇄 연접 (close juncture)이 일어날 경우이다. 같은 [i] 이면서도 '1'의 [i]와 '2'의 [i]는 조금 다르다. '2'의 [i]는 이완모음(lax vowel), '1'의 [i]는 긴장모음 (tense vowel)에 속한다고 할 수 있다. 그래서 뒤 세그먼트가 [i]인가를 검사해서 [i]이라면 [i]의 긴장모음에 해당하는 부분을 파워의 dip 정보를 이용하여 추출한다.

자연스럽게 발음 했을 경우 같은 이완모음 [i], [o] 끼리의 연접은 에너지 상태 변화로는 구분 할 수가 없으므로 여기서는 평균 길이를 이용하여 구분한다. 이완모음 [i], [o]의 평균길이를 lmi, lmo 라 하고 세그먼트 [i], [o]의 길이를 li, lo라하면, 모음 [i], [o]의 수 ni, no 는 다음과 같이 계산하여 가까운 정수를 취한다.

$$ni = li / lmi$$
$$no = lo / lmo$$

IV 어휘 해석

1. 소음소망 (phone network) 의 작성

음성신호를 적당한 단어 혹은 음절의 열로 변환하기 위하여 앞단에서 얻어진 단위의 이동 경로를 지정해주는 소음소망을 작성한다. 소음소망의 작성은 기본적으로 세단계를 거쳐 작성된다. 첫 번째로는 열개 숫자음의 변이음을 노드로 개별적인 브랜치를 만든다. 두번째로는 각 단어의 같은 기호로 시작하는 노드를 묶어 트리를 형성한다. 마지막으로는 각 단어의 제일 마지막 노드로부터 시작해서 10개의 각 단어 시작 노드로 연결하는데 음운학적 규칙을 적용한다 (그림 4).

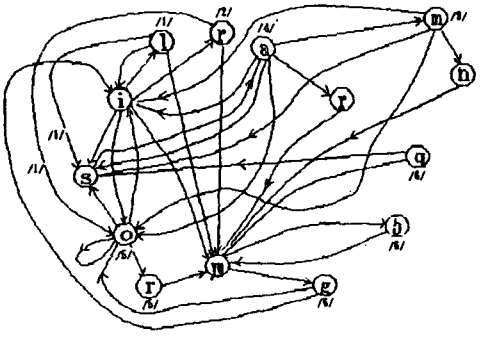

그림 4 소음소망 : 음운규칙을 적용하여 단어의 끝 노드에서 모든 시작 노드로 연결한 예.

오류정보가 포함된 세그먼트 계열을 인식하기 위해서는 모든 정보를 고려한 적당한 천이상태를 정의하지 않으면 안된다. 본실험에서 수정하여 얻은 소음소망을 그림5에 보인다.

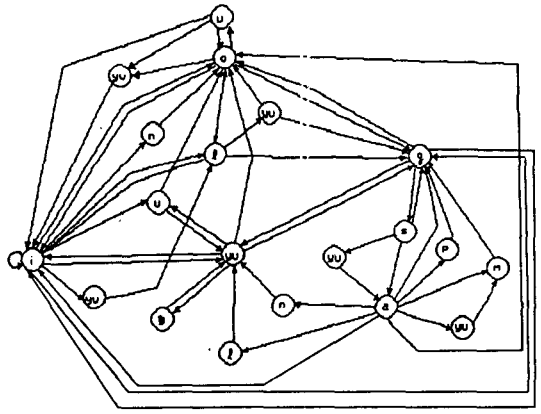

그림 5 수정된 소음소망.

V 실험결과 및 고찰

본 실험에서는 한국 전자 통신 연구소의 신호처리 연구실에서 88년 1월에 작성한 4연속 숫자음을 음성 대상자료로 선정하였다. 이 자료는 가능한 모든 조합을 고려해 만든 것이며 표 2 에 그 내용을 보인다. 녹음은 방음실에서 됐으며 발성 속도는 약 4.26 음절/초 이다. 필터는 4.5 LPF 이며 12 Bit 10KHz 샘플링된 데이타이다.

본 실험에서는 화자 3명이 두번 발성한 자료를 이용하였다. 첫번째 발성은 음소 표준패턴 추출용으로 사용하였으며 나머지 한개 발성을 인식 실험용으로 이용하였다.

표준 패턴을 위한 음소의 레이블링은 확대된 시간축상의 파형과 3차원 스펙트럼, 스펙트로그래프, 파워를 이용하였으며, D/A를 통하여 확인하여 Hand segmentation을 하였다.

실험 결과는 단음절을 기준으로 했을 경우 95%였고, 4연속 숫자음을 기준으로 했을 경우는 80%가 나왔다. 표 3 에 오인식 실험 결과를 보인다.

특징벡터로는 12차 LPC cepstrum 계수를 사용하였으며 프레임 레이블링과 소음소 레이블링은 결과를 볼 때 우수함을 알 수 있었다. 어려웠던 점은 소음소 레이블링을 거쳐 나온 [yu] 세그먼트의 정확한 성격규정을 위해서는 분석적인 방법도 함께 도입되야 함을 알 수 있었다. 또한 동일모음의 연접이 일어날 경우 이의 분리에 대한 알고리즘이 미흡하였다.

no.	contents	no.	contents	no.	contents
1	0287	13	1823	25	5500
2	5732	14	6378	26	6972
3	9601	15	8877	27	9861
4	4156	16	3510	28	3649
5	1199	17	8065	29	0316
6	1398	18	2934	30	7083
7	6843	19	7489	31	8194
8	0712	20	2244	32	9205
9	5267	21	4621	33	1427
10	6633	22	9176	34	2538
11	2409	23	3045	35	4750
12	7954	24	8590		

표 2 4연속 숫자 발성표.

오인식 결과	수정사항
2400 (2409)	0 --> 9
8064 (8065)	0 --> 5
x489 (7489)	x --> 7
4x41 (4621)	x --> 6, 4 --> 2
3040 (3045)	0 --> 5
9861 (5861)	9 --> 5
4799 (4750)	99 --> 50
x799 (1199)	x --> 1
7489 (1389)	1 --> 7
5297 (5267)	6 --> 9
1409 (2409)	2 --> 1
2405 (2409)	5 --> 9
7823 (1823)	1 --> 7
4177 (4621)	177--> 621
9990 (8590)	99 --> 85
79083 (7908)	9 --> x
8994 (8194)	9 --> 1
9105 (9205)	1 --> 2
1378 (6378)	1 --> 6
3945 (3045)	9 --> 0
4연속음별 인식률 : 80% 음절별 인식률 : 95%	

x 는 인식 안된 것임.

표 3 인식률

VI 결론

본 연구은 3인이 두번 발성한 모든 조합을 고려한 35종의 4연속 숫자음을 대상으로 인식 실험을 하였다. 음향-음성처리로써 연속음에서 추출하여 만든 음소 표준 패턴을 이용, LPC cepstrum 계수를 특징 벡타로 하는 프레임에 베이스 분류법을 이용하여 사후 확률을 구하고 프레임 레이블링을 하고, 그 결과를 이용하여 소음소 레이블링 하였으며, 어휘 해석으로는 소음소 망을 이용해 연속음 인식을 하였다.

음소 표준패턴은 세명이 처음 발음한 연속음에서 추출하였으며, 나머지 1 회 의 자료는 실험 평가용으로 사용하였다. 실험결과로는 4연속숫자음 단위로 80%의 인식율을 얻었으며, 음절 단위로는 95%의 인식율을 얻어 본 알고리즘의 유효성을 입증하였다.

결과를 볼 때 베이스 분류법을 이용한 프레임 레이블링과 소음소 레이블링 알고리즘이 우수함을 알 수 있었다. 또 세그먼트의 오분류에 의한 에러 정보를 소음소망에서 흡수함으로 우수한 결과를 내었다. 앞으로 더 진행해야 할 연구에서는 과도음 [yu]의 성격 규정을 위해 분석적 방법을 도입 하고, 동일 모음 연접 문제를 해결해야 할 것이다.

본 실험은 음성 다이얼링 시스템에의 응용을 목적으로 한것이나 Home banking 시스템등 많은 숫자음 인식을 이용하는 시스템에 적용될 수 있으며 대어휘 연속음 인식 연구에도 유용하게 이용될 수 있을 것이다.

참고문헌

1. Jean-Paul Halton, Automatic Speech Analysis and Recognition, Reidel Publishing Company, 1982.

2. 은 종관, "음성인식 기술현황," 한국 음향학회지 Vol. 7 No. 1, 1988.

3. Ronald A. Cole, Perception and Production of Fluent Speech, Lawrence Erlbaum Associates, Inc., Publishers, 1980.

4.

5. Sei-ichi Nakagawa, "Speaker - Independent Phoneme Recognition in Continuous Speech by a Statistical Method and a Stochastic Dynamic Time Warping Method," Technical Report of Carnegie - Mellon University, CMU-CS-86-102, 1988.

6. Regine Andre-Obrecht, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals," IEEE, Trans. on Acoustics, Speech, Signal Processing, Vol. 36, No. 1, Jan. 1988.

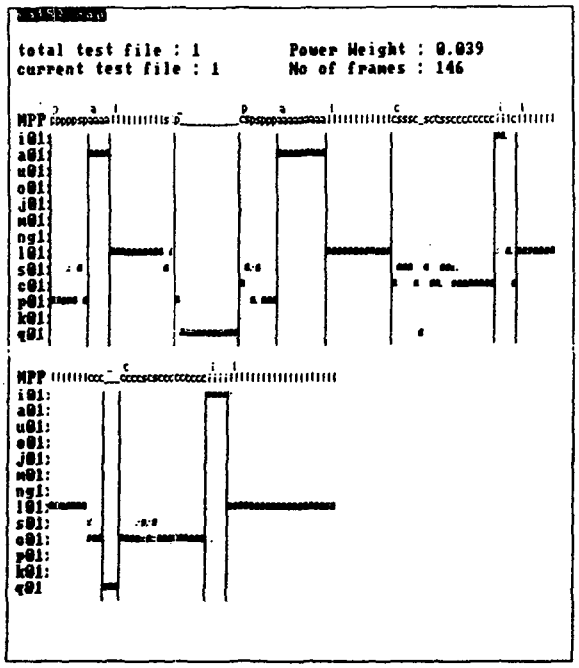

그림 6 소음소 테이블링 결과의 예. 발음내용 : 8877