음절수와 모음 열을 이용한 한국어 연결 숫자 음성인식

윤 재 선[†] · 홍 광 석[††]

요 약

본 논문에서는 음절수와 모음 열 정보를 이용한 한국어 연속 숫자 인식을 제안하였다. 제안한 연속 숫자 인식기는 첫 단계로 발성된 연속 숫자 음성에서 음절수와 구간을 추출하고, 두 번째 단계로 모음 열을 인식한다. 이와 같이 인식된 모음 열 정보를 이용하여 인식 후보를 줄이게 된다. 인식후보 모델은 조음효과에 효과적으로 대처할 수 있는 CV(Consonant Vowel), VCCV, VC단위 HMM(Hidden Markov Model)을 사용하여 연속 숫자 음성인식기를 구성하였다. 실험결과 제안된 방법이 조음효과를 효과적으로 대처하고 연결 숫자 인식에 유효함을 확인하였다.

Connected Korean Digit Speech Recognition Using Vowel String and Number of Syllables

Jeh-Seon Youn[†] · Kwang-Seok Hong[††]

ABSTRACT

In this paper, we present a new Korean connected digit recognition based on vowel string and number of syllables. There are two steps to reduce digit candidates. The first one is to determine the number and interval of digit. Once the number and interval of digit are determined, the second is to recognize the vowel string in the digit string. The digit candidates according to vowel string are recognized based on CV (consonant vowel), VCCV and VC unit HMM. The proposed method can cope effectively with the coarticulation effects and recognize the connected digit speech very well.

키워드 : 연결숫자 음성인식(Connected Digit Speech Recognition), 모음 열(Vowel String), 음절수(Number of Syllables), 은닉 마르코프 모델(HMM)

1. 서 론

컴퓨터의 발전과 통신을 이용한 정보 및 금융 서비스가 확대됨에 따라서 주민등록 번호, 비밀번호, 통장번호, 회원번호 등 많은 분야에서 연속 숫자 음성에 대한 인식을 필요로 하고 있다. 또한 연속 숫자는 키보드 입력뿐만 아니라 음성인식 등의 수단에 의한 입력의 필요성이 증가하고 있다.

연속 숫자 음성은 자음 부분 음성의 경우 경계가 불명확한 곳이 많고, 조음 현상으로 인하여 같은 음소라도 다르게 발음되는 경우가 많다. 따라서 단독 숫자 음성에 비해서 인식률이 훨씬 떨어지게 된다. 이러한 문제에도 불구하고 인간과 기계사이의 통신수단의 자연스러움과 원하는 속도를 얻을 수 있기 때문에 연속 숫자 음성인식의 중요성은 명백하다. 최근의 연속 숫자 음성은 주로 신경회로망이나 HMM을 기반으로 연구되어 왔다[1-8].

화자독립 연속 숫자 음성인식의 접근방법은 일반적으로 크게 두 가지 방식으로 나눌 수 있다[1]. 하나는 분절(segmentation) 없이 매칭(matching)에 기초한 것이고[4-7], 다른 하나는 분절 후에 분류(classification)하는 방식이다[2, 3, 8].

분절을 이용한 기존의 한국어 연속 숫자 음성인식은 주로 음절이나 음소 등의 부 단어(subword) 단위를 이용하여 인식을 하여 왔지만 연속 발성된 경우 음성의 특성상 음절이나 음소 단위로 정확한 분할을 하기가 매우 어려우며 이것이 인식률 저하의 주된 원인이다. 이를 보완하기 위하여 발성된 숫자열의 비교적 안정된 모음구간을 추출하여 분절한 후 앞 음절의 모음구간에서 다음 음절의 모음구간 까지를 인식단위로 하는 반음절(CV, VC)과 반음절 + 반음절(VCCV)을 이용하여 인식성능을 향상하였다[2].

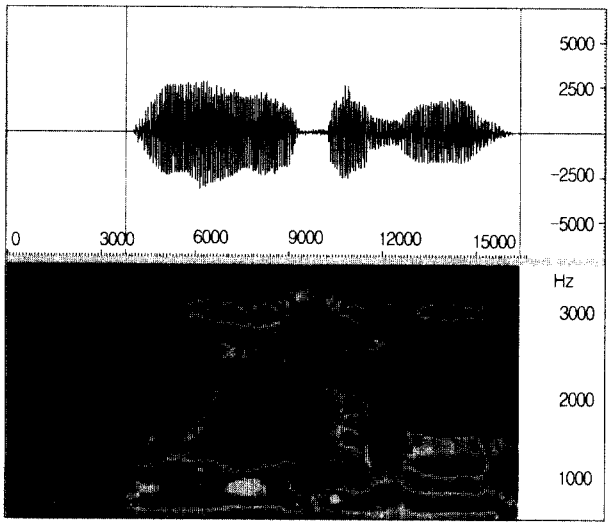
본 논문에서는 연속 발성된 숫자음성에서 비교적 안정된

† 정 회 원 : 보이스텍 음성기술연구소 선임연구원
†† 종신회원 : 성균관대학교 정보통신공학부 교수
논문접수 : 2002년 9월 28일, 심사완료 : 2003년 1월 3일

모음 부의 정보를 이용하여 음절수를 추출하고, 음절의 안정구간에 포함된 모음 열을 추출하여 인식한 후 연결 숫자의 인식 후보를 감축한다. 또한 인식된 모음 열 정보를 이용하여 CV, VCCV, VC 단위의 HMM 모델[2, 3]을 인식 단위로 하는 연속 숫자 음성인식 시스템을 제안하였다. 제안한 방법의 성능 평가는 모음 열 인식 정보를 적용하지 않은 방식[2]과 비교하여 평가하였다.

2. 연속 숫자 음성인식

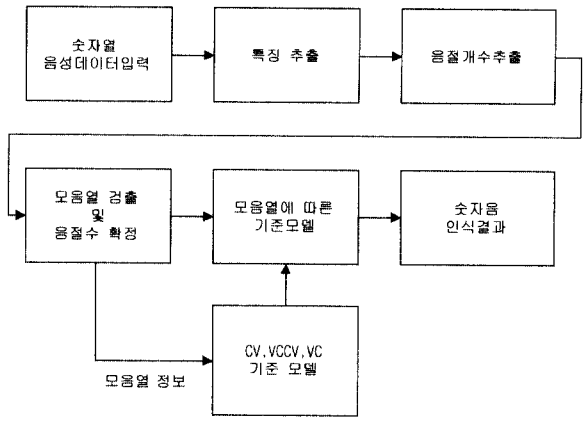
발성된 음성의 경계를 음소나 반 음절 또는 음절 등의 부 단어 단위로 정확하게 구분할 수 있다면 고성능의 연속 숫자 음성인식 시스템의 구성이 가능하다[4, 5]. 한국어 연속 숫자 음성의 경우 0에서 9까지 10개의 단음절의 조합으로 구성되어 있다. 그러나, 연속된 숫자 음성의 발성 시에는 숫자와 숫자사이에 음절의 구분이 쉽지 않게 연결되어 나타나는 경우를 음성 파형 관찰에서 쉽게 확인할 수 있다. 일례로써 (그림 1)에 4연속 숫자음성 5235/오이삼오/라고 발성한 음성 파형과 스펙트로그램을 나타내었다. (그림 1)에서 보는 바와 같이 자연스러운 발성인 경우에는 오와 이 사이, 삼과 오 사이와 같이 연결되어 있어서 정확한 음절 구분이 쉽지 않게 된다[2].



(그림 1) 숫자음성 /5235/의 파형과 스펙트로그램

3. 모음 열 인식을 이용한 연결숫자 음성인식

제안한 연속숫자 음성인식 시스템의 블록 도를 (그림 2)에 나타내었다. 입력된 음성으로부터 먼저 음절 개수를 추출한 후, 분할된 음절 영역으로부터 모음 부를 검출하여 모음 열을 인식한다. 이와 같이 인식된 모음 열로부터 CV, VCCV, VC 기준모델[2, 3]의 결합에 의해 후보 모델을 구성한 후 인식하도록 한다.



(그림 2) 연결 숫자인식 블록도

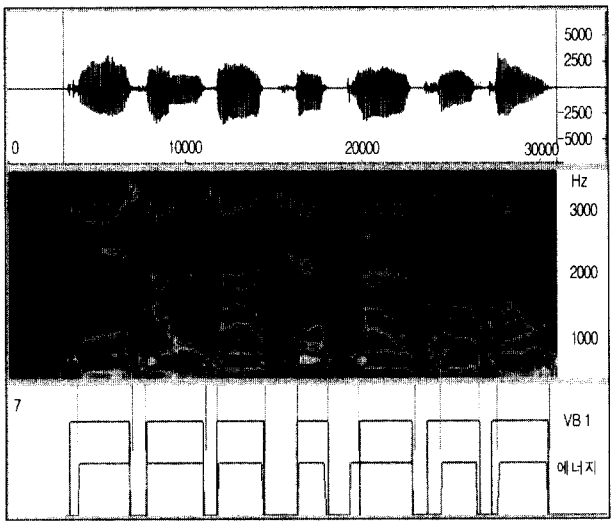
연속 숫자 음성인식 시스템에 사용되는 기준모델은 모음을 기준으로 한 반음절 CV, VC 단위와 VCCV 인식단위를 사용한다. 이때 사용되는 기준모델은 <표 1>과 같다. 숫자음성 /일/과 /이/의 앞쪽 반음절 CV는 /이/이고, /삼/과 /사/의 앞쪽 반 음절 CV는 /사/이기 때문에 중복되는 것을 공유하면 CV 인식단위의 개수는 8개로 되며, /일/과 /칠/의 뒤쪽 반 음절 VC가 /일/이므로 중복되는 것을 공유하면 VC 인식 단위의 개수는 9개로 된다. 마찬가지로 하여 VCCV 인식단위는 72개가 된다.

〈표 1〉 부 단어 단위 모델

인식단위	종 류	개 수
CV	고, 구, 사, 오, 유, 이, 치, 파	8
VC	아, 알, 암, 오, 옹, 우, 욱, 이, 일	9
VCCV	아고, 아구, 아사, 아오, 아유, 아이, 아치, 아파, 알고, 알구, 알사, 알오, 알유, 알이, 알치, 알파, 암고, 암구, 암사, 암오, 암유, 암이, 암치, 암파, 오고, 오구, 오사, 오오, 오유, 오이, 오치, 오파, 옹고, 옹구, 옹사, 옹오, 옹유, 옹이, 옹치, 옹파, 우고, 우구, 우사, 우오, 우유, 우이, 우치, 우파, 욱고, 욱구, 욱사, 욱오, 욱유, 욱이, 욱치, 욱파, 이고, 이구, 이사, 이오, 이유, 이이, 이치, 이파, 일고, 일구, 일사, 일오, 일유, 일이, 일치, 일파	72

3.1 음절 개수 추출

음절 개수 추출은 기준모델 작성시 연속 숫자의 개수에 큰 영향을 미치기 때문에 인식하기 전에 미리 음절 개수를 알고 있다면 인식 오류를 줄일 수 있게 된다. 먼저 입력 데이터로부터 에너지와 영 교차율(zero-crossing rate)을 이용하여 유성음과 무성음 영역을 검출한다. 또한 좀더 정확한 유성음 영역을 검출하기 위해 제 1 포먼트, 제 2 포먼트가 존재하는 215Hz와 2,756Hz사이의 에너지 정보 VB1(vowel band 1) 파라미터를 이용하여 안정된 유성음 영역을 추출하였다. VB1 파라미터의 주파수 대역은 모음영역에서 비교적 큰 에너지를 갖기 때문에 유성음 영역을 검출하는데 유효한 파라미터로 사용될 수 있다.

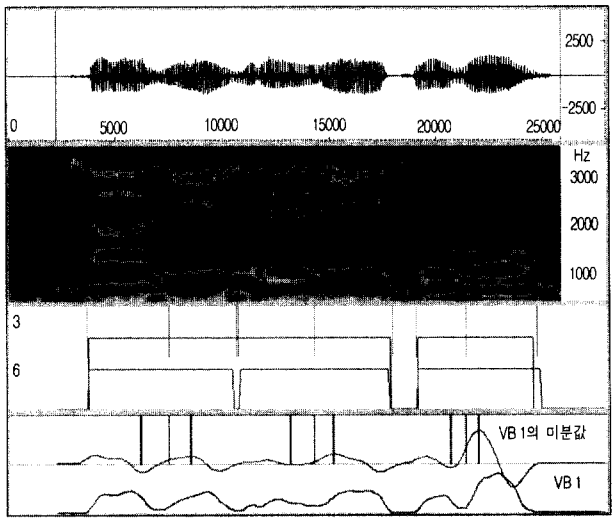


(그림 3) 숫자 /2347890/

(그림 3)은 /2347890/라고 발성을 하였을 경우에 에너지, 영 교차율과 VB1 파라미터의 값을 이용하여 음절 분할의 결과를 나타낸 것이다. 총 7개의 숫자음성이 정확하게 분할됨을 보여주고 있다. 그러나, 연음이 되는 숫자음성인 경우에는 예를 들어 /25/, /42/, /95/사이에서는 음절구별이 정확하게 되지 않는 경우가 있다. 따라서 좀더 정확한 음절의 개수를 추출하기 위해서 본 논문에서는 VB1 파라미터의 미분정보를 이용하여 분할한다. VB1 파라미터의 미분정보를 적용하는 방법은 다음과 같다.

(1) 에너지와 VB1 파라미터를 이용한 첫 번째 후보 분할 영역을 설정한다.
(2) 분할된 영역에서 좌우 8frame의 시간정보를 이용한 VB1 파라미터 미분값을 구한다. FIRST변수와 SECOND 변수를 FALSE로 초기화한다.
(3) 다음의 조건이 만족하면 FIRST변수를 TRUE로 변환하고, i번째 frame값을 M0에 저장한다.
  ① VB1_미분[i] < 0
  ② VB1_미분[i - 2] >= VB1_미분[i - 1]
  ③ VB1_미분[i - 1] >= VB1_미분[i]
  ④ VB1_미분[i] <= VB1_미분[i + 1]
  ⑤ VB1_미분[i + 1] <= VB1_미분[i + 2]
(4) FIRST변수가 TRUE이고, 다음의 조건이 만족하면 SECOND변수를 TRUE로 변환하고, i번째 frame값을 M1에 저장한다.
  ① VB1_미분[i] > 0
  ② VB1_미분[i - 2] <= VB1_미분[i - 1]
  ③ VB1_미분[i - 1] <= VB1_미분[i]
  ④ VB1_미분[i] >= VB1_미분[i + 1]
  ⑤ VB1_미분[i + 1] >= VB1_미분[i + 2]
(5) 다음 조건을 만족하면, 첫 번째 분할 영역에 음절분할을 추가하고 FIRST변수와 SECOND변수를 FALSE로 변환하고 (3)번을 다시 반복 수행한다.
  ① FIRST = TRUE
  ② SECOND = TRUE
  ③ M1 - M0 > 5 frames

(그림 4)는 /421295/라는 연속 숫자음성을 발성한 예를 보여 주고 있다. (그림 3)과는 달리 단어 안에 음절이 연음으로 되어져 있기 때문에 에너지와 VB1 파라미터를 이용하여 분리한 음절의 수가 어절의 개수와 같은 3개만이 검출되었지만 추가로 VB1 파라미터의 미분정보를 이용하여 분할한 결과 총 6개의 음절이 정확하게 분할되었다. 이와 같은 방법을 단어를 인식하는 시스템에 적용하게 되면 음절 개수 정보를 이용하여 단어의 후보를 줄일 수 있는 효과도 있게 된다.

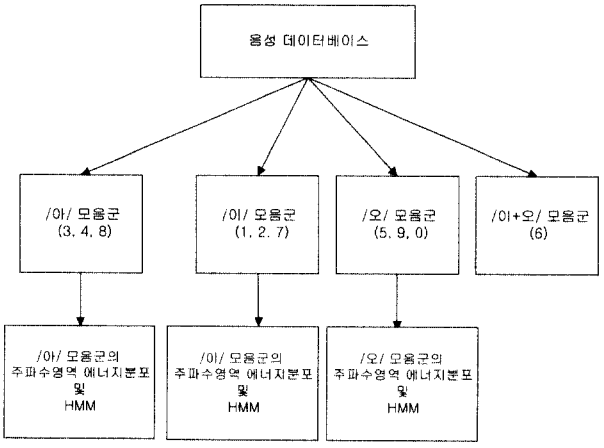


(그림 4) 숫자 /421295/의 음절 분할

3.2 모음 열 인식

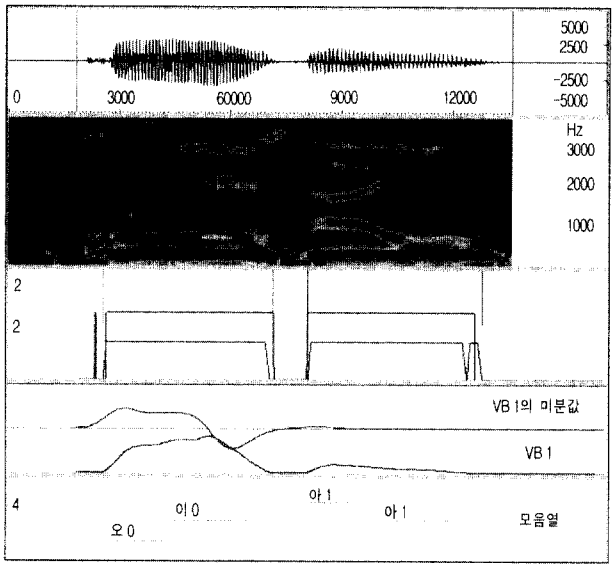
발성된 연속 숫자 음성에서 모음 열을 인식하기 위해서 /삼/, /사/, /팔/은 /아/모음 군에 /공/, /오/, /구/는 /오/모음 군에, /일/, /이/, /칠/은 /이/모음 군으로 구별하고 /육/은 /이+오/군으로 구별하였다. /오/와 /구/를 같은 /오/모음 군으로 분류한 것은 /오/와 /우/의 특성이 유사하여 변별이 쉽지 않았기 때문이다. 또한 /육/은 /이 + 오/나 /이 + 우/의 특성을 보였기 때문에 /이+오/로 구별한다.

모음 열을 구하기 위해서 숫자음성 뿐만 아니라 성명, 단음절, PBW가 포함된 음성 DB로부터 안정된 모음 영역으로부터 각 모음군의 제 1포먼트와 제 2포먼트의 주파수 영역의 에너지 분포와 모음 영역 HMM을 생성하여 기준 모델로 사용하였다[3]. (그림 5)에 모음군의 분류와 그에 따른 기준모델 생성을 나타내었다.



(그림 5) 모음 열을 추출하기 위한 기준모델 생성

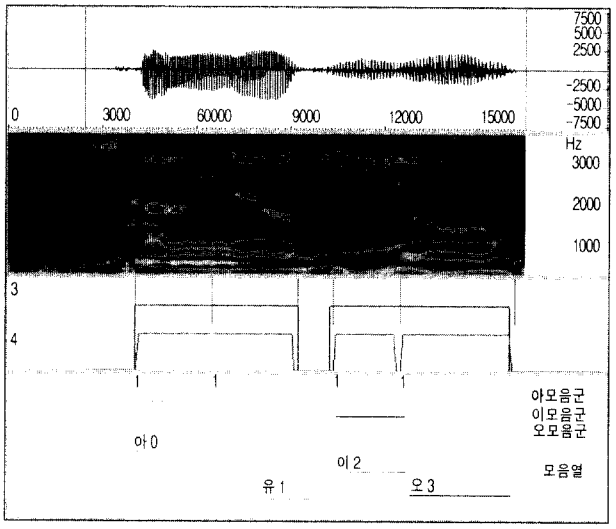
모음 열 추출 방법은 각각의 참조 모델의 모음군의 값이 임의의 기준 값 이상이 되면 모음 후보가 존재한다고 결정하고 인식하도록 하였다. 또한 숫자음성 /오이/, /사이/, /구일/과 같이 종성이 없는 음절과 초성이 /ㅇ/이 오는 음절인 경우에는 음절개수를 정확하게 추출하지 못하는 경우가 있다. 그러나 이 경우에는 앞 음절의 모음 군과 뒤따르는 음절의 모음 군이 각각 다르기 때문에 모음 열 추출 단계에서 음절개수를 추가하여 인식하도록 구성하였다. (그림 6)은 숫자음성 /9142/로써 음절 개수 추출 방법으로는 2개의 음절 개수 후보를 나타내지만, 모음 열 추출단계에서 음절의 개수가 2개에서 4개로 바뀌는 것을 보여주고 있다.



(그림 6) 숫자 /9142/의 음절개수 후보 및 모음 열 추출

음절 개수 추출시 오류가 잦은 연속된 숫자 열인 /이이/, /오오/는 발성 시간 정보를 이용하여 두 개의 영역으로 분할하도록 하였다. 숫자음성 /팔/인 경우에는 /아/모음 다음에 유성종성자음 /ㄹ/이 /이/모음 군으로 맵핑되는 경우가 있기 때문에 규칙을 적용하여 한 음절 안에 끊김 없이 /아/, /이/가 올 경우에는 뒤쪽에 위치한 /이/모음 군을 삭제하도록 구성하였다. 또한 모음 열이 짧은 시간을 갖는 /이/모음 군과 /오/모음 군이 연속적으로 발성될 경우에는 숫자음성 /육/으로 발성된 것으로 하였다.

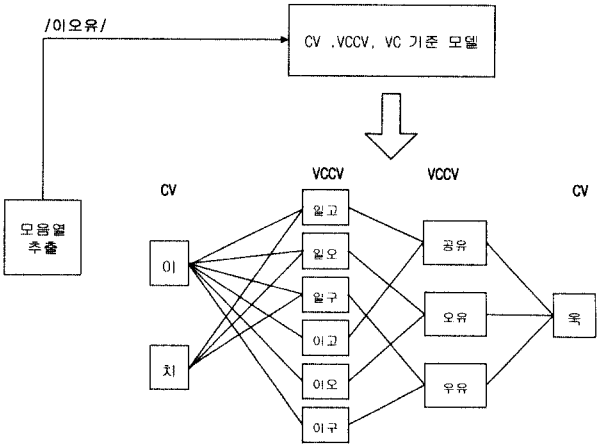
(그림 7)은 숫자음성 /3625/를 음절개수를 4개로 분할한 후, 모음 군을 추출한 결과를 나타낸 것으로써 /6/은 /이/와 /오/모음 군이 연속으로 나타났기 때문에 /유/모음 군으로 결과로 맵핑되는 것을 보여 주고 있다.



(그림 7) 숫자 /3625/의 모음 열 추출

3.3 모음 군에 따른 기준 모델 구성

모음 열 추출 결과에 따라 숫자음성 기준 모델을 결합하도록 함으로써 후보의 수를 줄여 인식 속도를 향상할 수 있도록 구성한다. (그림 8)은 모음 열 /이 오 유/에 따른 기준 모델구성의 결과를 나타내었다. /이/모음 군은 /이/, /일/, /칠/의 음절이, /오/모음 군에는 /공/, /오/, /구/음절이, /유/모음군인 경우에는 /육/음절의 인식 네트워크가 구성된다.



(그림 8) 모음 열 /이오유/의 기준모델 구성

예로써 /이 오 유/의 모음 열을 이용해서 기준모델을 구성할 경우 세 음절의 경우 후보 감축을 위한 별도의 방법을 적용하지 않는다면 구성할 수 있는 모델의 연결은 1,000(=10×10×10×1)개의 연결이 가능하고, 앞에서 순차적으로 제 1후보로 인식하여 가는 경우 30(=10+10+10×1)개의 연결로 인식이 가능하다[2]. 그러나 본 논문에서 제안한 방법은 (그림 8)에서 보는 바와 같이 9(=1×6×1×1+1×3×1×1)개의 연결만으로 가능하기 때문에 인식시간의 단축 효과도 기대된다.

4. 실험 및 결과

모음 열 인식 정보를 이용한 기준모델 구성에 따른 연결 숫자 음성인식 시스템의 유효성을 확인하기 위해서 모음 열 인식에는 연결 숫자음성 뿐만 아니라 성명, 단음절, PBW가 포함된 음성 DB를 이용하여 HMM모델을 구성하였다. 연결 숫자인식을 위한 CV, VCCV, VC 단위모델은 연결 숫자 음성 DB를 이용하여 모델을 생성하였다. 성능 평가를 위해서는 연결 숫자의 모든 조합이 고려된 4 연속숫자 음성 35개를 훈련에 참여하지 않은 화자 5명이 발성한 음성을 이용하여 평가하였다. 4 연속숫자 35개는 <표 2>와 같다. 이를 이용하여 4음절의 음절 개수 추출 정확도, 모음 열 검출 정확도, 인식률을 조사하였으며, 인식률의 비교 대상은 모음 열 정보를 적용하지 않은 기존의 4 연속숫자 음성의 CV, VCCV, VC 단위모델[2]의 경우와 비교 평가하였다.

〈표 2〉 인식실험에 사용한 4연속 숫자

1398, 2409, 6972, 5732, 6843, 5861, 1823, 2934, 5267, 6378, 2538, 1199, 6633, 2244, 5500, 1427, 0287, 3510, 4621, 8194, 7954, 8065, 9176, 9205, 9601, 0712, 3045, 3649, 4156, 7489, 8590, 0316, 8877, 7083, 4750

<표 3>에 음절개수 및 모음 열 추출 정확도를 나타내었다. 정확도는 4자리를 모두 알아맞히는 경우를 백분율로 나타내었다. 모음 열 인식 오류가 발생한 경우는 /칠/, /육/이 포함된 연속 숫자 음성에 나타난 것을 알 수 있다. /칠/인 경우는 /이/와 /일/의 숫자 음성에 비해 주파수 대역이 자음 /ㅊ/ 때문에 많이 달라지는 것을 알 수 있었으며, /육/인 경우에는 /이/와 /오/의 연속된 모음 열이 검출되어야 하지만, 자연스럽게 발성할 경우 /오/의 발음이 약하게 발성되면서 빨리 변화하기 때문에 검출을 하지 못하여 오류가 발생된 것으로 나타났다.

<표 4>에는 모음 열 정보를 이용한 시스템과 기존의 시스템과의 연속숫자 음성인식 결과를 나타내었다. 인식률은 4자리를 모두 알아맞히는 경우를 인식된 것으로 하여 조사하였다. 인식 결과를 살펴보면 모음 열을 이용한 시스템이 25.7%의 인식률의 향상이 있었는데 이는 모음 열을 적용하여 인식후보를 줄임으로써 인식 성능이 향상됨을 알 수 있었다. 특히 기존의 경우는 4자리 발성 음을 3자리 또는 5자리로 인식하는 경우도 다수 있었다. 4자리가 입력된다고 정하여 모델 생성 후 인식하는 경우에는 77.7% 정도의 인식률이 나왔는데 이 경우 제안한 방법의 인식률 향상은 4% 정도로 볼 수 있다.

〈표 3〉 음절개수 및 모음 열 추출 정확도

화 자	음절개수 정확도	모음 열 추출 정확도	모음 열 검출 오류 숫자음
1	34/35	33/35	5732,7954
2	35/35	33/35	5732,5267
3	35/35	31/35	6843,5267, 6378,4750
4	34/35	31/35	6843,5861, 5267,4750
5	35/35	33/35	5732,4156
평 균	98.8%	92%	-

〈표 4〉 4 연속숫자 음성인식 결과

화 자	제안한 모음 열 정보이용	기존의 방법
1	29/35	20/35
2	28/35	19/35
3	28/35	21/35
4	29/35	18/35
5	29/35	20/35
평 균	81.7%	56%

5. 결 론

본 논문에서는 발성된 연속 숫자음성에서 음절수와 모음 열을 추출하고, 추출된 모음 열에 해당하는 후보 모델을 VCCV 단위 HMM을 이용하여 인식하는 한국어 연속숫자 음성인식 시스템을 구현하였다. 음절수와 모음 열 정보를 이용하여 인식후보의 수를 줄였기 때문에 인식 시간의 단축을 가져오게 되고, 후보가 줄어드는 관계로 인식률 면에서도 좋은 성능이 나타남을 알 수 있었다.

인식 성능저하의 주된 원인은 음절수 추출과 모음 열 인식에서 발생하는 오류인데 인식 성능의 개선을 위해서는 모음 열 정보를 추출할 때 좀더 세밀한 규칙을 만들어서 적용하면 더 좋은 성능을 보일 것이라고 생각된다.

참 고 문 헌

[1] T. Ukita, E. Saito, T. Nitta and S. Watanabe, "A Speaker-Independent Connected Digit Recognition System Conca-

tenating Statistically Discriminated Words," IEEE Tran. on Signal Processing, Vol.40, No.10, pp.2414-2424, Oct., 1992.
[2] 윤재선, 홍광석, "반음절 단위 HMM을 이용한 연속 숫자 음성 인식", 한국음향학회지, 제17권 제5호, pp.73-78, 1998.
[3] 윤재선, 홍광석, "VCCV 단위를 이용한 어휘독립 음성인식 시스템의 구현", 한국음향학회지, 제21권 제2호, pp160-166, 2002.
[4] 김순협 외 4인, "음소 단위에 의한 한국어 연속 숫자음 인식에 관한 연구", 한국음향학회지, 제8권 제3호, pp.5-15, 1989.
[5] O. W. Kwon and C. K. Un, "Context-dependent word duration modelling for Korean connected digit recognition," Electron. Lett., Vol.31, No.19, pp.1630-1631, Sep., 1995.
[6] 양진우, 김순협, "HMM과 연결 숫자음의 후처리를 이용한 음성 다이얼링에 관한 연구", 한국음향학회지, 제14권 제5호, pp. 74-82, 1995.
[7] 박현상 외 3인, "Diphone 단위의 Hidden Markov Model을 이용한 한국어 단어인식", 한국음향학회지, 제13권 제1호, pp.14-23, 1994.
[8] 정광우, 홍광석, "MLP-VQ와 가중 DHMM을 이용한 연결 숫자음 인식에 관한 연구", 대한 전자공학회논문지, 제35권 제8호, pp.96-105, 1998.

윤 재 선

e-mail : sunhci@voicetech.co.kr
1996년 성균관대학교 전자공학과 공학사
1998년 성균관대학교 전자공학과 공학석사
2002년 성균관대학교 정보통신공학부 공학박사
2002년~현재 보이스텍 음성기술연구소 선임연구원
관심분야 : 음성 인식 및 화자 인식

홍 광 석

e-mail : kshong@skku.ac.kr
1985년 성균관대학교 전자공학과 학사
1988년 성균관대학교 전자공학과 석사
1992년 성균관대학교 전자공학과 박사
1990년~1993년 서울보건전문대학 전산정보처리과 전임강사
1993년~1995년 제주대학교 정보공학과 전임강사
1995년~현재 성균관대학교 정보통신공학부 부교수
관심분야 : 음성인식 및 합성, HCI