연결 숫자음 인식 시스템의 구현과 성능 변화

윤영선(한남대), 박윤상(보이스피아), 채의근(공주대)

<차 례>

<Abstract>

A Study on the Implementation of Connected-Digit Recognition System and Changes of its Performance

Young-Sun Yun, Yoon-Sang Park, Yi-Geun Chae

In this paper, we consider the implementation of connected digit recognition system and the several approaches to improve its performance. To implement efficiently the fixed or variable length digit recognition system, finite state network (FSN) is required. We merge the word network algorithm that implements the FSN with one pass dynamic programming search algorithm that is used for general speech recognition system for fast search. To find the efficient modeling of digit recognition system, we perform some experiments along the various conditions to affect the performance and summarize the results.

* Keywords: korean digit recognition, speech recognition, gender model, duration model

1. 서 론

일반적인 연속 음성 인식은 음소 모델(phoneme model)이나 단어 모델(word model)과 같은 인식 단위(recognition unit)를 모델링한 후, 탐색 단계에서 언어 모델을 이용하여 인식 단위들의 연결 정보를 파악한 후 최적의 단어 열을 구한다. 그러나 숫자음 인식은 탐색 단계에서 탐색 공간(search space)을 줄일 수 있는 언어 정보를 이용하기 어렵다는 단점이 있다. 외국어의 경우 숫자음이라 할지라도 각 숫자음 간의 변이는 일반적인 연속 음성 인식에서의 단어 변이 수준으로 파악되기 때문에, 언어 정보를 사용할 수 없다는 단점을 제외하고는 연속 음성 인식의 음향학적 모델링을 그대로 적용할 수 있다. 한국어의 경우 숫자음은 일(一), 이(二), 삼(三) 등과 같은 한자(漢字) 숫자음과 하나, 둘, 셋 등과 같은 한글 숫자음으로 분류할 수 있다. 두 종류 모두 각 숫자음마다 유사성이 연속 음성 인식에서의 단어 유사성보다 높다. 특히 한자 숫자음은 숫자음이 단일 음절로 이루어졌기 때문에 그 인식 방법은 연속 음성 인식보다 더 어렵게 생각되고 있다. 따라서 일반적인 숫자음 인식은 자릿수를 고정하거나 주민등록번호나 신용카드 번호와 같이 적용되는 환경에 맞는 위치 정보를 이용하여 숫자음 인식의 오류를 줄이고자 하는 연구가 진행되고 있다.

일반적으로 숫자음 발성은 3연 또는 4연으로 발성되고 있으며, 거의 대부분 6연 또는 8연 발성을 넘지 않는다. 사람들이 숫자음을 발성할 때 6연 또는 8연 이상의 발성음은 발성과정에서 자연스럽게 3연 또는 4연 발성으로 구분되어 발성되기 때문에, 6연이나 8연 발성은 연결 숫자음에서 발성 가능한 최장 연속 발성음으로 간주할 수 있다. 따라서 일반 음성 DB를 수집하는 과정에서는 10연이나 16연 연속 발성 숫자음의 경우, 일부 간투사를 허용하고 있다[1]. 따라서 본 논문은 8연 숫자음을 대상으로 연결 숫자음 인식 시스템을 구축하며, 다양한 조건 하에서 인식 성능에 영향을 미칠 수 있는 요인과 그에 따른 성능 변화를 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 연결 숫자음 인식 시스템에서 숫자음을 탐색하는 과정에 대해 살펴보고, 3장에서는 숫자음 데이터 베이스의 구성과 다양한 조건에서의 숫자음 인식 실험 및 그 성능을 살펴본다. 마지막으로 4장에서 본 연구의 요약 및 결론을 맺도록 하겠다.

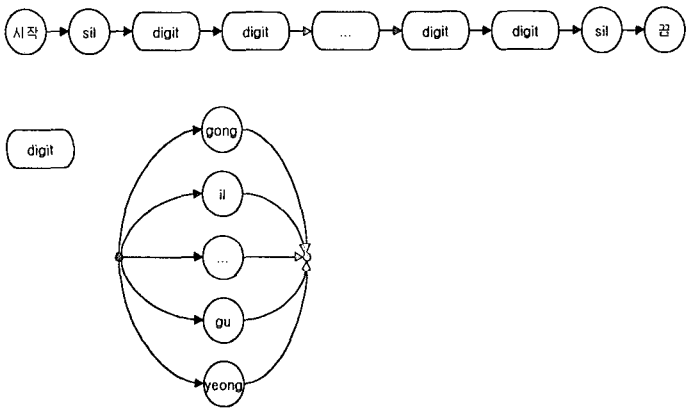
2. 연결 숫자음 인식 시스템의 구성

본 장에서는 연결 숫자음 인식 시스템에서 자릿수를 고정하기 위하여 적용되는 단어 그래프(word graph)를 이용한 유한 상태 네트워크(FSN; Finite State Network)의 구현 방식을 설명하고, 연속 음성 인식에서 널리 사용되는 1단계 동적

프로그램(OPDP; One-Pass Dynamic Program) 방식과 유한 상태 네트워크의 결합 알고리즘을 설명한다.

2.1. 유한 상태 네트워크

고정 길이의 연결 숫자음 인식을 위해서 연결 단어를 유한 상태 네트워크로 표현하여야 한다. FSN 표현하는 방법에는 여러 가지가 있을 수 있으나, 본 연구에서는 HTK (Hidden Markov Toolkit)[2]에서 사용하고 있는 단어 네트워크(Word Network)의 형식을 따라 유한 상태 네트워크를 구현하였다.



<그림 1> 8연결 숫자음의 유한 상태 네트워크

단어 네트워크는 노드(node) 리스트와 아크(arc, 호) 리스트로 구성되어 있다. 노드는 단어를 표현하고 아크는 단어간의 전이(transition)을 의미한다(<그림 1> 참조). HTK의 표준 격자 형식(SLF; Standard Lattice Format)에서 노드와 아크의 정의는 줄 단위로 표현되며, 여러 필드로 구성된다. 단어 단위의 SLF 형식을 기술하는 것이 어렵지 않더라도 시간이 많이 걸리는 작업이기 때문에, HTK에서 제공하는 HParse를 이용하여 확장 BNF (extended Backus-Naur Form)에서 간단하게 SLF 형식을 구할 수 있다. 확장 BNF를 이용하여 8연속 연결 숫자음 인식을 정의하기 위해서는 다음과 같이 표현할 수 있다.

<표 1> HParse의 확장 BNF를 이용한 8연속 연결 숫자음

```
digit = gong | il | i | sam | sa | o | yug |
        chil | pal | gu | yeong;
( sil $digit $digit $digit $digit $digit $digit $digit $digit sil )
```

일반적으로 동일한 단어의 반복이 많은 경우에는 단어 네트워크를 공유할 수 있기 때문에 부분 네트워크(sub network) 개념을 도입하여 SLF 형식을 정의할 수 있다. 이 경우에 부분 네트워크를 먼저 기술하고 다음 단계에서 부분 네트워크를 이용하여 다시 노드를 구성하여야 한다.

인식 단계에서는 이 SLF 파일을 읽어서 단어 네트워크를 구성하고, 단어간의 전이가 발생하는 경우에 현재 단어 노드와 연결된 단어 노드로 전이한다. 본 연구에서는 빠른 검색 방법으로 널리 사용되는 OPDP 기법으로 유한 상태 네트워크를 구현하였다.

2.2. OPDP와 FSN의 결합

OPDP는 각 프레임에서 최적의 조건을 검사하는 알고리즘으로써 Vintsyuk과 Bridle, Brown, Nakagawa에 의하여 독립적으로 연구되었다. 이들 알고리즘들은 서로 독립적으로 연구되었지만 거의 동일한 방식의 알고리즘이며, 재귀적인 DP 계산을 포함하지 않기 때문에 계산량이나 메모리 면에서 효율적인 알고리즘으로 널리 사용되고 있다.

특히 입력 음성과 참조 패턴간의 거리와 누적 거리를 계산하는 계산량은 입력 음성의 길이에 독립적이다. 따라서, 입력 음성의 길이와 평균 참조 패턴의 길이가 일정하다면 계산량은 어휘의 크기 n에 비례하게 되어 $O(n)$ DP로도 불린다. 그러나 여러 개의 경로가 한 상태에서 병합되는 경우, 최대 확률 값을 가지는 경로만 보존하고 나머지 경로는 보존하지 않아 여러 후보 경로를 얻지 못하는 단점이 있다[3].

OPDP는 각 프레임에서의 최적 확률값을 이용하는 방법이며, FSN은 특정 단어에서 다른 단어로 전이하는 경로를 제한한다. 따라서, 이 두 방법을 결합하기 위해서는 FSN을 구성하는 단어 네트워크에서 현재 탐색 중인 노드를 저장하는 지시자가 필요하며, 이 지시자는 새로운 경로가 생성되면 새로운 지시자로 복사된다. 이 알고리즘은 기존의 Token Passing Algorithm[2]과 거의 동일한 방식으로 동작하나, OPDP 알고리즘을 병행하여 적용하였기 때문에 특정 프레임에서 동일한 단어로 시작하는 경로가 두 개 이상인 경우, 최적의 경로를 포함하는 지시자만 유지되고 나머지는 제거된다. 이 과정을 반복하여 입력 음성이 종료되었을 경우에 FSN의 마지막 종료 노드에 도달한 지시자가 가지는 경로가 최적 경로로 채택된다. 이 알고리즘은 <표 2>와 같이 간단하게 기술될 수 있다.

<표 2> OPDP와 FSN의 결합 알고리즘

단계 1 : 초기화 (initialization) (t=0)
○ FSN의 시작 가능한 단어의 초기 상태 우도 값(likelihood)을 초기화

단계 2 : 동적 정합 탐색 (dynamic programming search)
◎ 모든 프레임에 대해 탐색 시작 (t=1, ..., T-1)
● 현재 지시자가 가리키는 FSN의 단어 노드에 대해서
○ 단어간 전이 단계
- 현재 지시자가 가리키는 FSN의 단어 노드 다음에 관측 가능한 단어들을 검색하여 현 지시자 이동 (단어 지속시간 모델링 가능)
- 다음 경로에 이미 단어가 추가된 경우, 기존의 저장된 확률값과 비교하여 높은 확률값을 갖는 경로 선택
○ 단어 내부의 전이 단계
- 상태간 전이 실행
○ 경로 저장

단계 3 : 경로 역추적 (path back tracking)
○ FSN의 최종 노드에 경로가 형성되었으면, 대응되는 단어들 중, 가장 높은 확률값을 가지는 단어에서 경로 역추적 시작
○ FSN의 최종 노드에 도달된 경로가 없으면, 현재 단어 중 높은 확률값을 가지는 단어에서 경로 역추적 (탐색 오류)

제안된 알고리즘의 파악된 문제점은 입력 음성이 종결되더라도 FSN의 최종 노드에 지시자가 도착하지 못하는 경우이다(탐색 오류). 이것은 인식 단위가 음향학적으로 제대로 모델링되지 못한 경우 발생하는 현상으로 특정 단어가 입력 음성에 지배적으로 반응하는 경우이다. 이 현상은 OPDP 알고리즘의 단점을 그대로 반영하고 있으며 일반적으로 단어 지속 모델이나 상태 지속 모델을 이용하여 해결한다.

3. 실험 및 결과

본 장에서는 연결 숫자음 인식 시스템의 성능 향상을 가져오기 위하여, 여러 조건에서 인식 실험을 하고 그 결과를 정리한다. 실험에 사용된 숫자음 데이터는 (주)보이스피아에서 수집한 8연 숫자음 1500문장을 남·여 각각 15명씩 100문장을 3세트씩 발성한 9,000문장을 학습에 사용하였으며, 평가용으로는 8연 숫자음 500문장을 남·여 각각 5명씩 100문장을 3세트씩 발성한 3,000문장을 사용하였다. 평가용 문장에서 청취 후 잘못 녹음된 20문장을 제외하여 총 2,980문장을 평가에 사용하였다. 각 음성은 조용한 사무실 환경에서 16kHz샘플링으로 녹취되었으며, 20ms의 분석구간과 10ms의 구간이동, 24차의 필터뱅크 에너지에 의한 12차 MFCC로 변환되었다. 추출된 MFCC는 다시 22차 Liftering과 CMN (Cepstral Mean Normalization)을 통과하였으며, 정규화된 log energy와 1차 미분 계수 13차를 더하여 26차가 최종 음성 특징으로 채택되었다. 그리고 이 26차 음성 특징은 분절 특징의 기본 특징으로 사용되었으며, $N=1, R=1$인 경우에는 일반 HMM과 동일한 구성을 가지게 된다[4].

3.1 인식 단위 결정

숫자음 인식에서 사용될 수 있는 인식 단위는 음소 모델과 단어 모델이 있다. 인식 단위를 결정하기 위하여 단음소와 삼음소 모델을 3개의 상태로 모델링하여 실험하였다. 휴지(sil) 모델을 제외한 각 음소 모델별 발생 모델 개수 및 훈련 데이터 수는 <표 3>과 같다.

<표 3> 음소 모델 수와 평균 발생 빈도

음소 구분	모델 수	평균 빈도
단음소	13	12,040
삼음소	309	506

단음소 모델과 삼음소 모델을 이용하여 일반 HMM을 이용하여 연결 숫자음 인식을 한 결과는 <표 4>에 보인다. 표에서 보인 것처럼, 단음소 모델을 이용하는 경우 인접한 음소 변화에 따른 음운 현상을 적절히 표현하지 못하기 때문에, 단어 정확률(word accuracy)은 비슷한 성능을 보이더라도, 문장 인식률(sentence correct)에서 많은 차이가 남을 알 수 있다. 표에서 N과 R은 분절의 길이와 회귀 차수를 의미하고, M은 혼합 밀도(mixture)의 수를 의미한다. 따라서, N=1, R=1, M=4는 일반 HMM의 혼합 밀도의 수가 4임을 표시한다.

<표 4> 음소 모델별 인식 결과 (N=1, R=1, M=4)

모델	단어 정확도 (word accuracy %)	문장 인식률 (sent. correct %)
단음소	83.5	31.1
삼음소	84.8	38.4

다음으로 단어 모델을 이용하여 실험하였다. 이 경우, 단어 당 상태 수를 결정하기 위하여 음소 당 세 개의 상태로 모델링하였다. 따라서 11개의 숫자음 단어는 구성 음소 수에 따라 상태가 3개, 6개, 9개의 구조를 가지게 된다. 단어 모델은 앞뒤 단어를 고려하는 것에 따라 음소 모델과 마찬가지로 한 단어 모델(mono-word), 두 단어 모델(bi-word), 세 단어 모델(tri-word)로 구분하게 된다. 두 단어 모델과 세 단어 모델은 모델 수를 줄이기 위하여 앞 단어의 종성음과 뒷 단어의 초성음을 고려하여 동일 음소는 동일군으로 묶어 실험을 하였다. 단어 모델별 빈도 수는 <표 5>에 정리되어 있다.

<표 5> 단어 모델 수와 평균 발생 빈도

단어 모델	모델 수	평균 빈도
한 단어	11	6,521
두 단어	99	725
세 단어	885	80

세 단어 모델의 경우, 빈도 수가 10이하인 경우도 발생하여, 빈도수가 10이하인 경우에는 모델에서 제외하였다. 일반적으로 고단위 단어 모델의 문장 허용 범위가 전체 평가 문장의 발생 범위를 충분히 포함하지 못하는 경우가 발생할 수 있기 때문에 저단위 단어 모델을 이용하여 연결 숫자음의 허용 범위를 넓히고 있으나(back-off 모델 이용), 본 실험에서는 back-off 모델 방식을 사용하지 않고 순수한 문장 허용 범위를 조사하였다. 일반적으로는 한 단어 모델보다 두 단어, 세 단어 모델 순으로 인식 성능이 좋은 것으로 발표되었으나, 본 실험에 사용된 데이터베이스가 작아 각 인식 단위별로 발생 빈도수가 많은 한 단어 모델의 성능이 가장 좋았다.

<표 6> 각 단어 모델별 인식 결과 (N=5, R=3, M=4)

인식 모델	단어 정확도 (%)	문장 인식률 (%)
한 단어	91.7	56.0
두 단어	88.6	50.8
세 단어	85.3	39.8

위의 결과로부터 한 단어 모델을 이용하는 것이 현재 데이터베이스의 크기에서는 최적의 인식 성능을 보이는 것으로 판단하여 실험을 진행하였다. 일반적으로 숫자음의 구성을 살펴보면, 단음소 단어(이, 오)와 이음소 단어(일, 사, 육, 구, 영), 삼음소 단어 (삼, 칠, 팔, 공)로 분류할 수 있기 때문에, 단어 모델을 구성하는 상태 수에 따른 인식률의 변화를 생각할 수 있다. 또한 혼합 밀도 수가 증가할수록 인식 성능도 따라 향상되기 때문에 혼합 밀도의 수를 10으로 고정시키고, 각 구성 상태 수의 변화를 실험하였더니, <표 7>과 같은 결과를 얻었다. 표의 시스템 구성은 앞에서 설명한 바와 같이 각 단어를 구성 음소 수에 따라 구분하여 "단음소-이음소-삼음소 단어모델의 상태 수"를 6-6-9로 표시하고 있다.

<표 7> 한 단어 모델의 상태 수 변화에 따른 성능 변화

시스템 구성	단어 정확도 (%)	문장 인식률 (%)
6-6-9	93.9	66.4
6-8-10	94.4	68.1
8-10-12	94.8	70.5
10-10-10	94.8	70.3
10-12-14	94.8	70.0
12-14-16	94.3	69.0

위 실험 결과에서 단어 별 상태 수의 결정에 따른 성능 변화를 알 수 있으며, 단음소 단어, 이음소 단어, 삼음소 단어의 상태 수가 각각 8, 10, 12 일 때 가장 성능이 우수함을 알 수 있었다.

3.2 성별 인식 단위

음소 모델이나 단어 모델을 이용하여 인식 단위를 결정하였을 경우에, 남성이나 여성의 변이가 큰 경우, 혼합 밀도를 증가시키는 경우보다 성별을 구별하여 모델링하는 방법이 사용되기도 한다. 성별을 구별하는 모델을 적용하게 되면, 동일

한 성별에 의한 연속적인 상태 변이를 모델링을 할 수 있기 때문에, 혼합 밀도를 증가시키는 것보다 선호되고 있다. 즉, 혼합 밀도를 증가시키게 되면 상태의 전이에 따라 성별에 따른 음향학적 분포가 일관성이 있게 연결되지 않으나 성별 분포에 따라 모델을 별도로 구성하게 되면, 탐색 공간에서 일관성을 유지할 수 있기 때문이다.

성별을 구별한 음소 모델과 단어 모델의 성능 변화는 <표 8, 9>에 정리되어 있다. <표 8>은 성별을 구별하지 않은 경우의 전체 인식률과 남성과 여성을 분리하여 실험한 인식률을, <표 9>는 성별을 적용한 경우의 전체 인식률, 남성, 여성 인식률을 보이고 있다.

<표 8> 성별을 적용하지 않은 경우의 성능 변화

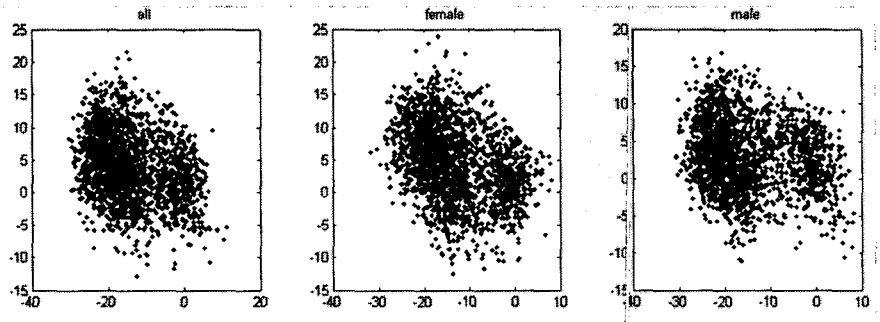
시스템		전체		남성		여성	
		단어	문장	단어	문장	단어	문장
삼 음소 모델		84.8	38.4	83.4	34.5	86.3	42.2
단어 모델	6-6-9	93.9	66.4	94.2	67.3	93.7	65.5
	6-8-10	94.4	68.1	94.7	69.9	94.0	66.3
	8-10-12	94.8	70.5	95.0	71.9	94.6	69.1
	10-10-10	94.8	70.3	95.0	71.9	94.6	68.8
	10-12-14	94.8	70.0	95.0	71.2	94.6	68.8
	12-14-16	94.3	69.0	94.8	72.2	93.8	65.8

<표 9> 성별을 적용한 경우의 성능 변화

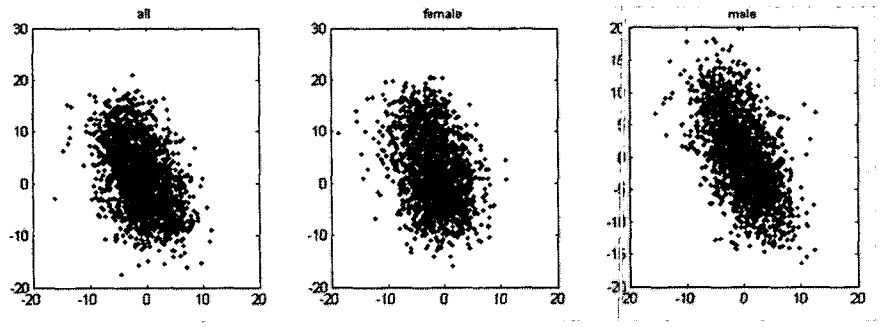
시스템		전체		남성		여성	
		단어	문장	단어	문장	단어	문장
삼 음소 모델		86.2	42.1	83.8	35.2	88.6	49.1
단어 모델	6-6-9	94.3	68.4	93.9	66.9	94.6	69.9
	6-8-10	94.9	71.2	94.8	71.3	95.0	71.0
	8-10-12	95.1	72.2	95.1	72.6	95.1	71.9
	10-10-10	95.4	73.5	95.6	74.8	95.2	72.1
	10-12-14	95.6	74.1	95.6	75.2	95.4	73.0
	12-14-16	94.5	68.8	94.2	67.0	94.8	70.6

위 표에서 삼 음소 모델은 $N=1, R=1, M=4$ 인 경우의 음소 단위 모델이며, 단어 모델은 $N=5, R=3, M=10$인 경우의 한 단어 모델을 의미한다. 위 실

험 결과에서 음소 모델인 경우나 단어 모델인 경우 성별을 구별한 경우의 인식 성능이 더 향상됨을 알 수 있다. 그러나 음소 모델과 단어 모델에서 남·여 성별에 따른 성능 변화가 다름을 알 수 있다. 음소 모델을 이용한 경우나, 단어 모델에서 상태 수가 작은 경우, 또는 상태 수가 아주 많은 경우에서는 남성에 비하여 여성의 인식 성능이 더 높음을 알 수 있다. 이것은 남성과 여성의 음향학적 분포가 서로 다르다는 것을 의미하며, 적절하지 않은 상태 수는 인식 성능의 저하를 발생한다는 것을 알 수 있다(단어 모델 12-14-16).



<그림 2> 'ㅣ' 음소의 성별 분포도 (전체, 여성, 남성)

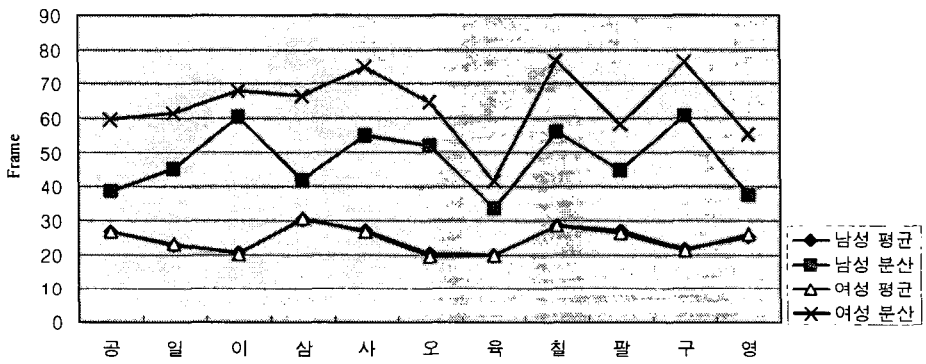


<그림 3> 'ㅊ'음소의 성별 분포도 (전체, 여성, 남성)

위 <그림 2, 3>은 MFCC 1차와 2차 특징의 분포를 보이고 있다. 여기에서 여성 음성 특징은 남성 음성 특징에 비해서 좀 더 밀집한 형태를 보이고 있다. 물론 시간적인 분포를 배제하였기 때문에 정확한 판단은 힘들지만 위 그림으로부터 남성의 변이가 더 크다는 것을 알 수 있다.

3.3. 단어 지속 모델

지속 모델링은 탐색 과정에서 특정 상태 또는 단어의 왜곡된 감쇠 현상이나 이들 요소가 지배적으로 음향학적 분포를 보이는 현상을 보정하기 위해서 널리 사용되는 방법이다[8]. 본 연구에서는 2장에서 설명한 연결 숫자음 인식 시스템의 탐색 알고리즘에서 탐색 오류를 보정하고, 인식 성능을 향상시키기 위한 방법으로 단어 지속 모델을 사용한다. 단어 지속 모델을 사용하기 위해서는 학습 데이터에 대한 단어의 평균 길이 통계가 필요하며 <그림 4>에 정리되어 있다.



<그림 4> 성별 지속시간 분포도

위 그림에서 살펴보면, 각 숫자음은 구성 음소에 따라 그 길이도 달라짐을 알 수 있다. 따라서 기존 인식 시스템에서 단어 지속 모델을 적용할 경우 인식 성능의 변화를 예측할 수 있다. 지속시간 분포도를 살펴보면 남성이나 여성의 경우 평균 지속시간은 비슷한 시간을 보이나, 여성의 지속시간의 변이가 더 큼을 알 수 있다.

단어 지속 모델을 적용하기 위하여 OPDP와 FSN의 결합된 탐색 과정에서 단어 전이에 따른 penalty 값으로 단어의 지속시간에 따른 정규화 분포를 적용하였다. 실험 결과는 <표 10>에 나열된다.

<표 10> 지속 모델 적용에 따른 성능 변화

시스템		성별 구분 없음				성별 구분			
지속 모델		미적용		적용		미적용		적용	
		단어	문장	단어	문장	단어	문장	단어	문장
삼 음소 모델		84.8	38.4	-	-	86.2	42.1	89	50.4
단어 모델	6-6-9	93.9	66.4	94.2	67.6	94.3	68.4	94.6	70.1
	6-8-10	94.4	68.1	94.5	68.8	94.9	71.2	95.0	71.9
	8-10-12	94.8	70.5	94.6	69.6	95.1	72.2	95.1	72.3
	10-10-10	94.8	70.3	94.7	69.9	95.4	73.5	95.3	73.1
	10-12-14	94.8	70.0	94.6	69.6	95.6	74.1	95.3	73.6
	12-14-16	94.3	59.0	94.1	68.2	94.5	68.8	94.3	68.2

위 실험 결과에서 알 수 있듯이 인식 단어의 상태 길이가 작은 경우에는 지속 모델을 적용한 경우 성능 향상이 있음을 알 수 있다. 그러나 인식 단어의 상태 길이가 긴 경우에는 성능 향상이 없음을 알 수 있다. 이 결과로부터 단어의 상태 길이가 작은 경우, 평균 단어 길이보다 작은 숫자음의 삽입 가능성을 줄여 성능이 향상되었다고 판단되며, 상태 길이가 긴 경우에는 전체 연결 숫자음에서 삽입 오류가 감소되어 지속 모델링의 효과가 적은 것으로 판단된다. 특히 성능 향상이 발생한 삼음소 모델의 경우, 성능의 변화가 남·여 다른 현상을 보였다(<표 11>).

<표 13> 남·여 별 성능 변화 및 상대적 증가율

	남성		여성	
지속 모델 적용 여부	단어	문장	단어	문장
미 적 용	83.8	35.2	88.6	49.1
적 용	87.1	44.3	91.0	56.5
성능 향상	3.9	25.9	2.7	15.1

<그림 4>와 <표 11>에서 알 수 있듯이 지속시간의 평균이 남성과 여성 비슷하다고 하더라도 여성 화자의 지속시간 변이가 크기 때문에, 인식 성능 향상이 남성이 더 크게 나옴을 알 수 있다.

3.4. 검토

본 장에서는 연결 숫자음 인식 성능에 따른 다양한 조건을 실험하고 그에 따른 영향을 살펴보았다. 먼저 학습 데이터가 작은 경우에는 음소 모델 기반의 인식 시스템에 비하여 단어 기반의 인식 시스템의 성능이 우수하였으며, 단어 기반이라 할지라도 한 단어 모델의 인식 단위가 더 효과적이었다. 또한, 성별에 따라 모델을 다르게 적용한 경우, 남·여 성별에 따라 그 효과가 다름을 알 수 있었다. 여성 숫자음의 경우 남성이 발성한 숫자음에 비하여 유사성이 강하기 때문에 음소 기반의 인식 시스템인 경우 성능 차이가 컸으며(남성 0.7%, 여성 6.9%), 단어 기반의 인식 시스템인 경우 전체적으로 남성이나 여성 모델의 성능이 향상됨을 알 수 있다. 또한, 지속 모델을 적용한 경우 단어의 상태 수에 따라 그 성능의 변화가 다름을 알 수 있었으며, 남성에 비해 여성의 단어별 지속시간의 변이가 커 남성 음소 모델에 지속시간을 적용하였을 경우 성능 향상의 폭이 큼을 알 수 있었다.

음소 모델이나 단어 모델인 경우 단어 별 상태 수에 따라 성능의 변화가 커 본 연구에서는 평균 단어 길이에 기초해 음소 모델의 상태 수를 다르게 적용한 결과 성능의 변화가 급격하다는 것을 알 수 있었다. 삼음소 모델 기반의 $N=5, R=2, M=4$인 환경에서 실험한 결과는 <표 12>에 정리되었다. 이 실험에서는 삼음소 모델을 적용하여 '이'와 '오'에 해당하는 음소는 상태 수를 6개로 결정하고 나머지 음소에 대해서는 4개의 상태를 할당하였다. 그 결과 모든 음소에 대해 동일하게 3개의 상태를 할당한 시스템의 성능보다 우수함을 알 수 있어, 각 음소별 또는 단어별 상태 수를 적절하게 배분하는 것이 성별 모델이나 지속 모델 못지 않게 중요함을 알 수 있다.

<표 12> 삼음소 기반의 상태 수 변화에 따른 성능 변화
(성별 모델, 지속시간 모델 적용, N=5, R=2, M=4)

시스템	전체		남성		여성	
	단어	문장	단어	문장	단어	문장
음소별 동일 상태 수(3)	91	57.7	89.4	52.5	92.5	62.8
음소별 다른 상태 수 (4,6)	94.9	73.5	93.8	69.2	96.0	77.8

4. 요약 및 결론

본 연구에서는 연결 숫자음 인식 시스템을 개발하기 위하여 간단하게 1단계 동적 정합 방식과 유한 상태 네트워크를 결합한 탐색 알고리즘을 사용하였으며, 인식 시스템의 성능에 영향을 줄 수 있는 다양한 조건에 대해 실험을 하였다. 실험 결과 소규모 데이터베이스인 경우 음소 모델에 비해 단어 모델의 성능이 더 우수함을 알 수 있었다. 또한 성별 모델을 적용한 결과 음소 모델을 이용한 경우에는 남성보다 여성의 성능 변화가 컸으나, 단어 모델을 적용한 경우 남·여 모두 비슷한 성능 향상을 가져옴을 알 수 있다. 지속 모델을 적용한 경우, 단어의 상태 수가 작은 환경에서 더욱 효과적임을 알 수 있었다. 성별이나 지속 모델을 적용한 경우도 인식 성능에 영향을 주나, 각 숫자음의 평균 길이에 따른 상태 수의 변화가 인식 성능에 많은 영향을 주는 것을 알 수 있다. 이런 실험 결과는 앞으로 더 낳은 인식 시스템의 개발에 많은 도움을 줄 수 있을 것으로 여겨진다. 추후 연구로는 한국어 숫자음 인식에 적합한 상태 수 모델링과 음소별 가중치에 따른 성능 변화에 대한 연구가 필요할 것으로 보인다.

참 고 문 헌

[1] ETRI, "음성 DB 수집 방법", http://voice.etri.re.kr/db/d-collectionMethod.asp, 2002.
[2] S. J. Young et al., *The HTK Book*. 2001.
[3] 오영환, *음성언어정보처리*, 홍릉과학출판사, 1998.
[4] S. J. Young, N. H. Russell, J. H. S. Thornton, "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems", *Technical Report TR*, Vol. 38, Cambridge University Engineering Department, 1989.
[5] 윤영선, 오영환, "모수적 궤적 기반의 분절 HMM을 이용한 연속 음성 인식", *한국음향학회지*, 19권, 3호, pp.35-44, 2000.
[6] 윤영선, 오영환, "분절 특징 HMM의 매개 변수 수의 감소에 관한 연구", *한국음향학회지*, 19권, 7호, pp.48-52, 2000.
[7] 윤영선, "분절 특징 HMM을 이용한 영어 음소 인식", *한국정보과학회지*, 29권, 3호, pp.167-179, 2002.
[8] L. Rabiner, B-H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

접수일자: 2003년 2월 7일
수정일자: 2003년 3월 7일
게재결정: 2003년 3월 8일

▶ 윤영선(Young-Sun Yun)
주소: 대전시 대덕구 오정동 133번지
소속: 한남대학교 정보통신공학과
전화: 042) 629-7569
FAX: 042) 629-7843
E-mail: ysyun@mail.hannam.ac.kr

▶ 박윤상(Yoon-Sang Park)
주소: 대전시 유성구 구성동 400번지 동문창업관 3108호
소속: (주)보이스피아
전화: 042) 863-2171
E-mail: yspark@voicepia.co.kr

▶ 채의근(Yi-Geun Chae)
주소: 충남 공주시 옥룡동 326
소속: 공주대학교 멀티미디어정보 · 영상공학부
전화: 041) 850-6210
FAX: 041) 858-0580
E-mail: ygchae@kongju.ac.kr