Journal of Korea Multimedia Society Vol. 9, No. 1, January 2006(pp. 23-32)

반음절쌍과 변형된 연쇄 상태 분할을 이용한 연속 숫자 음 인식의 성능 향상

이수정[†], 서은경[††], 최갑근[†††], 김순협[††††]

요 약

본 논문에서는 언어모델과 음향모델을 개선함으로써 단위 숫자음의 인식성능 최적화에 대해 설명한다. 언어모델은 한국어 단위 숫자음 문장의 문법적 특징을 분석하고, Finite State Network(FSN) 노드를 두 음절로 구성하여 오 인식률을 감소시켰다. 음향모델은 단 음절로 구성되어 발성기간이 짧고 조음이 많이 생기는 불명확한 음소, 음절의 분할로 인한 오 인식을 줄이기 위해 인식단위를 반음절 쌍으로 하였다. 인식단위의 특징을 효과적으로 모델링하기 위해 특징부분에서 K-means 알고리즘으로 군집화 하여, 상태를 분할하는 변형된 연쇄 상태 분할방법을 이용하였다. 실험 결과 제안된 언어모델의 적용 후 동일 문맥종속 음소모델에서 10.5%, 음향모델에서 인식단위를 반음절 쌍으로 하였을 경우 문맥종속 음소모델에 비해 12.5%, 변형된 연쇄 상태분할을 하였을 경우 1.5%의 인식률을 향상시킬 수 있었다.

Performance Improvement of Continuous Digits Speech Recognition Using the Transformed Successive State Splitting and Demi-syllable Pair

Lee, Soo Jeong[†], Seo, Eun Kyoung[††], Choi, Gab-Keun[†††], Kim, Soon Hyob[††††]

ABSTRACT

This paper describes the optimization of a language model and an acoustic model to improve speech recognition using Korean unit digits. Since the model is composed of a finite state network (FSN) with a disyllable, recognition errors of the language model were reduced by analyzing the grammatical features of Korean unit digits. Acoustic models utilize a demisyllable pair to decrease recognition errors caused by inaccurate division of a phone or monosyllable due to short pronunciation time and articulation. We have used the K-means clustering algorithm with the transformed successive state splitting in the feature level for the efficient modelling of feature of the recognition unit. As a result of experiments, 10.5% recognition rate is raised in the case of the proposed language model. The demi-syllable pair with an acoustic model increased 12.5% recognition rate and 1.5% recognition rate is improved in transformed successive state splitting.

Key words: Language model(언어모델), Demi-syllable pair(반음절쌍), Transformed successive state splitting(변형된 연쇄 상태분할)

※ 교신저자(Corresponding Author) : 이수정, 주소 : 서울시 노원구 월계1동(139-701), 전화 : 031)717-9979
FAX : 031)716-9930, E-mail : leesoo86@apet.co.kr
접수일 : 2005년 10월 18일, 완료일 : 2006년 1월 13일
[†]준회원, 광운대학교 음성신호처리 박사과정
[††]준회원, 광운대학교 대학원 컴퓨터공학과
(E-mail : ekseo@acanettv.com)
[†††]광운대학교 대학원 컴퓨터공학과 박사과정
(E-mail : cocomm@kw.ac.kr)
[††††]광운대학교 컴퓨터공학과 교수
(E-mail : kimsh@daisy.kw.ac.kr)
※ 이 논문은 2005년도 광운대학교 연구년에 의하여 연구되었음

1. 서 론

연속음성인식은 자연어 형태의 문장을 인식하는 것으로 대용량의 어휘사전, 음향모델, 문장 내 단어 간의 전후관계인 언어모델을 구성하여 인식과정이 진행된다.

음향모델은 음성 신호를 나타내는 표현 방법으로 가장 많이 사용하는 모델의 형태는 은닉 마코프 모델(Hidden Markov Models)이다[1-3]. 이 모델은 샘플을 기반으로, 필요한 확률 변수를 추정하여 인식 대상 어휘를 모델링 하는 확률적 처리 기법이다. 음성인식에서의 샘플은 음성의 특징 매개변수로써 신호처리 기법을 이용하여 추출하며, 이 특징 변수들로만 추정한 모델을 음향 모델이라 한다[4]. 언어모델은 음성인식기의 문법이라 할 수 있는데, 그 종류는 인식단어간의 연결 관계를 망구조로 표현하는 구조적 모델링 방법인 FSN(Finite State Network)[1-2]과 단어 간의 연결 관계를 확률로서 표현하는 n-gram 방법이 있다[2]. 연속숫자음의 인식에서는 숫자음의 특성상 단어 간 전후 관계가 무의미하며, 임의의 숫자음 문장 표현이 가능해야 하므로 구조적 문법인 FSN을 사용한다.

본 논문에서 이용한 방법은 언어모델과 음향모델의 개선이다. 먼저 언어모델을 개선하기 위해 한국어 단위 숫자음 문장구조를 파악하고, 문맥적 규칙을 발견하여 적용하였다. 단위 숫자음 문장은 숫자음 한 음절과 단위음 한 음절로 구성되었다. 단 음절로 이루어진 한국어 단위 숫자음 인식에서 오 인식 요인을 줄이고자 하였다. 음향모델의 개선을 위해 한국어 숫자음은 단음으로 구성되어 발성구간이 짧아 음절, 음소 구분이 어렵다. 그러므로, 연속음성인식에서 오인식을 줄이기 위해 반음절 쌍(demi-syllable pair)을 인식단위로 모델링 하였다. 반음절 쌍 모델같이 하나의 모델에 변이음 특성이 많이 들어갈 경우 고정된 개수의 상태로 각 음성의 특성을 잘 반영할 수 없으므로, 상태를 분할함으로써 각 발성의 시간적 특성이 인식 모델에 반영되도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 연속 숫자음 인식 모델을 설명하고, 3장에서는 상태 분할에 대해 설명한다. 4장에서는 제안된 내용의 검증을 위한 실험과정을 설명하고 결과에 대해 고찰한다. 마지막으로 5장에서 본 논문의 결론을 내리고자 한다.

2. 한국어 연속 숫자음 모델

2.1 단음절 FSN

FSN은 인식하고자 하는 단어의 연결 관계를 망구조로 표현하는 것으로서 자유도가 낮고, 사람의 언어는 다양한 변이가 있어서 FSN을 벗어나는 경우가 많다. 특히 대화체 언어로 가는 경우에는 그 정도가 특히 심하다. FSN을 음성인식에 사용하는 경우는 비행기 표 예약, 기차표 예매 등과 같은 작은 태스크에서 사용자의 자유를 제한하는 경우에 사용된다. 받아쓰기 또는 자연 발화에 의한 대화 음성을 인식하고자 하는 경우에는 형식 문법으로는 언어현상을 모두 고려할 수 없다. 그림 1은 FSN의 예를 보였다.[2]

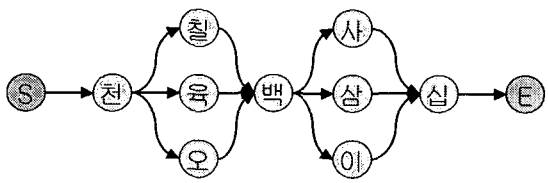


그림 1. 유한상태의 예

2.2 두 음절 기반 FSN

한국어 단위 숫자음은 특정한 규칙을 가지고 발음되어 이것을 이용하여 인식률을 향상시킬 수 있다. 그림 2와 같은 숫자음 문장의 문법적 특징을 구조적으로 모델링 하는 방법으로 FSN을 이용한다. FSN에서 나타나는 각 노드는 인식단어이다. 한국어 단위 숫자음 문장을 구성하는 단어가 "숫자음+단위음"으로 구성된 점을 착안하고, 기본 인식단어를 두 음절로 구성하여 문법적 규칙을 추가 하였다. 문법적 제약을 크게 할수록 인식기의 탐색 범위를 줄일 수 있으므로 인식률이 높아진다.

"*" : (sil), 일, 이, 삼, 사, 오, 육, 칠, 팔, 구

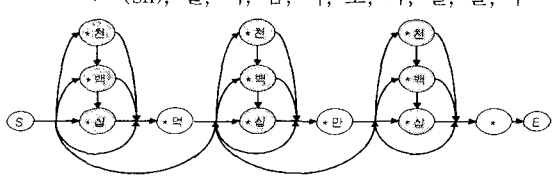


그림 2. 한국어 단위 숫자음 문장의 구조

그림 2는 한국어 연속단위 숫자음의 구조를 보여준다. 숫자음의 발성은 left-to-right 방식으로 다양한 경로를 통해 각각의 노드를 지나며, 가변적인 숫자문장을 표현할 수 있다.

2.3 반 음절 쌍 음향모델 제안

기존HMM[3,5] 음향모델(mono, triphone)을 사용하는 한국어 연속 숫자음의 인식은 숫자 음이 단 음절로 구성되어 있고, 연속 음성인식의 특성상 발성시에 숫자와 숫자사이에 음절의 구분이 명확하지 않고, 연결되어 나타나는 경우가 있어 인식에 어려움이 많다.

음절의 구분이 명확하지 않은 제한된 영역에서의 인식에 음절의 안정구간인 모음을 기준으로 분할하여 모델링을 하였다. 모음을 기준으로 하나의 음절을 양분하여 반음절 단위[8]로 나누고, 앞 음절의 뒷 반음절과 뒤 음절의 앞 반음절을 하나의 쌍으로 모델링하여 반음절 쌍(demi-syllable pair) 모델을 구성한다. 그림 3에서 반음절 쌍 모델의 모습을 보여준다.

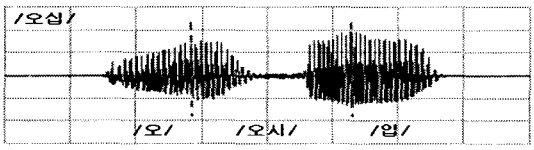


그림 3. 반음절 쌍 모델의 예 /오십/

반음절 쌍 인식단위로 모델링 할 경우 모델은 표 1과 같이 6가지 형태로 나타난다.

표 1. 반 음절 쌍 모델의 구성

종류	구성
V	vowel
VV	vowel-vowel
VC	vowel-consonant
VCCV	vowel-consonant-consonant-vowel
CV	consonant-vowel
VCV	vowel-consonant-vowel

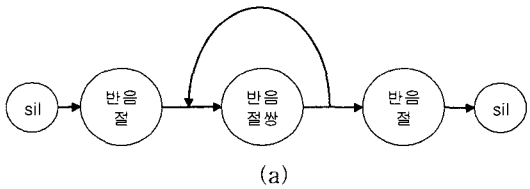
표 2. 반 음절 쌍 인식단위

S :음절의 시작부분,E:음절의 끝부분, D 숫자
T:십, B:백, C:천, M:만, A:억, J:주, W:원

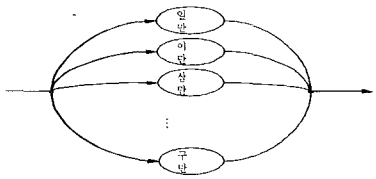
2.3.1 인식과정

연속 단위 숫자음 인식에서 인식 대상이 되는 가변 숫자음 문장의 인식은 반음절 쌍 모델에 의하여 그림 4와 같은 구조로 나타난다. 문장의 처음과 끝은 반음절로 나타나며 중간은 반음절 쌍이 반복적으로 나타난다. 그림 4 (a)는 반음절 쌍 모델기반 FSN을 보여주고 있으며, (b)는 단위 FSN 이 인식하는 두 음절 숫자음을 보여주고 있다.

정의된 모델과 인식과정을 연속 단위 숫자음 문장 "구만 사천 원"에 적용하여 그림 5에 나타내었다.



(a)



(b)

그림 4. 유한상태 구조

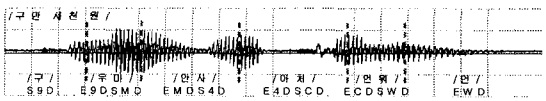


그림 5. 연속 단위 숫자음 /구만사천원/ 반 음절 쌍의 예

3. 연쇄 상태 분할

3.1 Successive State Splitting

연쇄 상태 분할(SSS:Successive State Splitting) [6,7]은 처음부터 정밀한 음향모델을 작성하지 않고 반복학습을 통해 상태의 분할과 파라미터의 추정을 점진적으로 정확하게 학습하는 방법이다. 먼저 유사 음소단위를 기본으로 하여 전체모델을 연결한 구조의 초기모델로서 각각의 모델은 하나의 상태와 그 상태를 처음부터 끝까지 연결하여, 전체 학습 음성 데이터로부터 생성한다. 상태분할은 경로분할을 동반하는 문맥방향과 경로분할을 동반하지 않는 시간

방향이 있는데, 출력확률의 우도에 따라 한 방향으로만 수행된다. 문맥방향으로 분할 할 때는 경로분할에 동반된 각각의 경로에 할당된 문맥 클래스도 같이 분할된다. 따라서, 문맥 클래스의 분할에 포함된 모든 상태 중에서 학습 데이터에 대한 누적우도 확률이 높은 쪽 상태를 분할하도록 선택한다. 이러한 상태분할을 반복하며 Hidden Markov Network[6] 구조를 결정한다. 그러나, 학습음성 데이터를 이용하여 상태분할과 파라미터 추정을 반복하기 때문에 최종모델을 학습하는데 필요한 계산량이 상대적으로 증가하는 문제점이 있다. 따라서, 불특정화자의 대규모 학습 음성 데이터에 의해 재추정하는 방법을 이용하고 있고, 기본모델의 구조결정에 이용한 학습 음성 데이터가 특정화자의 한정된 음성 데이터이므로 대 어휘 연속음성인식을 위한 음향모델을 작성하기 위해 많은 어려움이 있다[9].

3.2 변형된 연쇄상태 분할

본 논문에서는 유사도가 최대에 이르는 시점까지 모델별 상태분할을 진행하되 K-means 군집화[4] 방법을 이용하여 가우시안을 구하고, 각각을 하나의 상태에 할당한 뒤 시간방향으로 정렬하여 상태 분할을 반복하는 방법으로 모델링 하여 인식 성능을 향상시키고자 하였다.

3.1 초기 모델 학습

K-means 군집화에 의해 다중 혼합 가우시안을 구하고, 각각의 가우시안을 하나의 상태로 할당하여 초기 모델을 학습한다. 먼저 다중 혼합 가우시안을 구하기 위해 각각의 모델을 분석하여 K개의 임의의 군집 센터를 결정한다. 표 3과 같이 가변 숫자음 문장에 나타날 수 있는 묵음(silence)을 제외한 반 음절 쌍 모델 168개를 각각의 모델이 나타날 수 있는 형태인 V, CV, VCV, VCCV, VV, VC 6가지로 분류한다. 한국어에서는 VCCV, VCV형태로 나타난다.

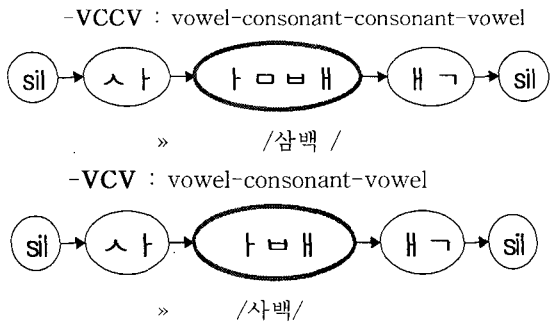


그림 7. 반음절 쌍 모델

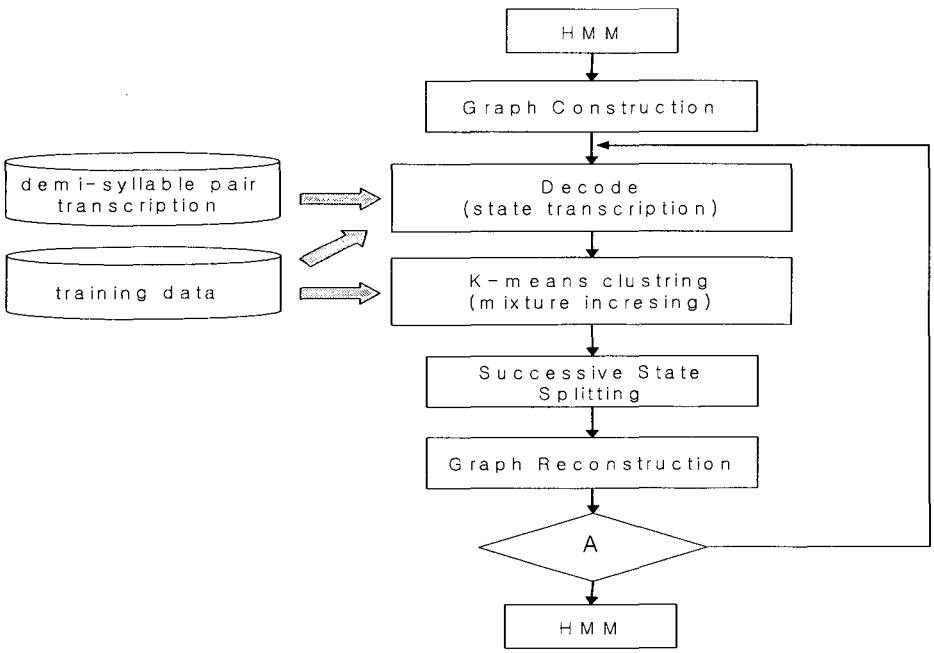


그림 6. 변형된 연쇄 상태 분할 과정

표 3. 반 음절 쌍 모델의 구분

요소 / 그룹	phone 개수	인식모델
V	1	SAD, S6D, S5D, S2D, S1D, EJD, E9D, E5D, E4D, E2D
CV	2	STD, SMD, SCD, SBD, S9D, S8D, S7D, S4D, S3D
VCV	3	ETDSWD, ETDSAD, ETDS6D, ETDS5D, ETDS2D, ETDS1D, EMDSWD, EMDSAD, EMDS6D, EMDS5D, EMDS2D, EMDS1D, ECDSWD, ECDSAD, ECDS6D, ECDS5D, ECDS2D, ECDS1D, EBDSWD, EBDSAD, EBDS6D, EBDS5D, EBDS2D, EBDS1D, ...
VCCV	4	ETDSMD, ETDSJD, ETDSCD, ETDSBD, ETDS9D, ETDS8D, ETDS7D, ETDS4D, ETDS3D, EMDSTD, EMDSJD, EMDSCD, EMDSBD, EMDS9D, EMDS8D, EMDS7D, EMDS4D, EMDS3D, ECDSTD, ECDSMD, ECDSJD, ECDSBD, ECDS9D, ECDS8D, ECDS7D, ...
VC	2	EWD, ETD, EMD, ECD, EBD, EAD, E8D, E7D, ...
VV	2	E9DSWD, E9DSAD, E5DSWD, E5DSAD, E4DSWD, ...

다음으로 아래와 같은 단계를 거쳐 그룹별로 초기 모델을 학습한다.

단계1.

6가지로 분류한 모델에서 각각의 음소(phone)의 개수를 군집 개수 K로 정하고 K-means 알고리즘[4]으로 군집화 한다. 표3에서 구분한 모델의 음소의 개수로 K개의 임의의 군집 센터 $Z_1(1), Z_2(1), \ldots, Z_k(1)$ 을 설정하고 Euclidean 거리 측정법에 의해 ①군집화 중심 값과 벡터의 거리 값 $Distance = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$ 을 구하여 각각의 벡터를 군집에 분배한다. ②군집 된 벡터들과 중심점과의 거리의 평균값 $Z_j(t+1) = \frac{1}{N_j}\sum x$ 을 기준으로 새로운 군집 중심점을 계산하고 중심점을 새로 결정한다. 이러한 과정을 반복하다가 중심점의 이동이 있으면($Z_j(t+1) \neq Z_j(t)$) ①, ②의 과정을 반복하고 이동이 없으면($Z_j(t+1) = Z_j(t)$) 군집화 과정을 끝낸다.

단계2.

각각 군집화 된 특징에 대하여 다중 혼합 가우시안을 구한다.

단계3.

하나의 가우시안을 하나의 상태에 할당하여 단일 가우시안 상태로 만든다.

단계4.

각각의 상태를 유사도에 따라 시간방향으로 정렬하여 초기모델을 완성한다.

그림 8은 음소의 개수 4인 VCCV그룹의 초기 모델이다.

3.2 K-means에 의한 variation and temporal modeling

초기 모델을 연쇄 상태 분할하여 최적의 상태 네트워크를 가지는 모델을 구성하고자 한다. SSS와의 차이점은 각각의 반음절 쌍 모델이 포함하는 음소의 개수로 K-means 군집화 하여 구한 가우시안 각각에 상태를 할당하고, 시간방향으로 정렬하여 모델별로 초기 모델을 정의하고, 2혼합 수 가우시안을 구할 때 K-means 알고리즘으로 군집화 하였다는 점이다. 1혼합 수 가우시안을 가지는 상태를 유사도에 따라 발음특성이 반영된 변이방향 혹은 시간방향으로 배열함으로써, 변이음과 이중모음 등의 형태가 나타나

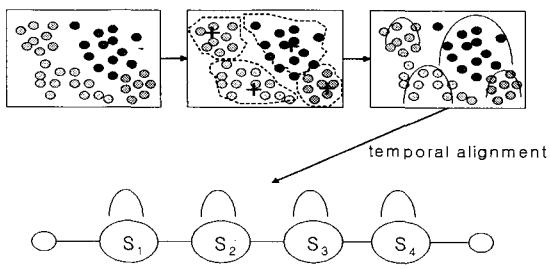


그림 8. VCCV 그룹의 초기 모델

는 반음절 쌍 모델의 최적화된 상태 네트워크를 구하고자 하였다[7,8]. 분산이 제일 큰 S2를 분할할 상태로 결정하고, 선택된 상태를 2개로 분할한다. 그 후 분할된 상태의 배치를 변이방향과(병렬)과 시간방향(직렬)으로 나눈 후 둘 중에 유사도가 큰 것이 최종 인식된다. 그림 9는 K-means 군집화에 의한 가우시안 상태분할 과정을 보여주고 있다.

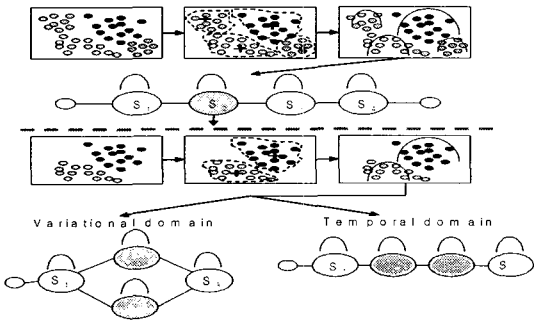


그림 9. K-means 군집화에 의한 변형된 연쇄 상태분할

4. 실험 및 분석

4.1 훈련 데이터베이스와 테스트 데이터베이스

4.1.1 훈련 데이터베이스 분석

본 논문에서 연속 숫자음의 인식을 위해 두 가지의 음성 데이터베이스를 사용하였다. 첫 번째는 조음현상을 고려하여 숫자음과 단위음을 1음절에서 4음절까지의 길이를 갖도록 다양하게 구성한 음성 데이터베이스이며, 두 번째는 숫자음 1음절과 단위음 1음절로 구성한 두음절 기반 음성 데이터베이스이다. 음성 데이터베이스 첫 번째 것을 가리켜 DB1, 두 번째 것을 가리켜 DB2라 한다. DB1과 DB2의 내용을 정리해 보면 표 4, 표 5와 같다.

DB1과 DB2는 증권거래용으로 숫자음 데이터베이스로 증권거래량을 나타내는 단위음 '주'와 주식가격을 나타내는 단위음 '원'을 포함한다. DB1의 인식 가능한 숫자음은 10부터 수천억 단위까지 10배수의 숫자이다. DB2는 1부터 수천억 원 단위까지 인식범위 내의 자연수 모두를 인식할 수 있도록 구성되어 있다.

4.1.2 테스트 데이터베이스 분석

테스트 데이터베이스는 표 8에서 나타낸 것과 같이 274단어로 구성된 연결단위 숫자음 60문장을 훈련 데이터베이스에 참여하지 않은 화자 20명의 발성분을 사용하였다.

훈련 데이터베이스와 테스트 음성 데이터베이스의 신호 처리는 표 9와 같다.

표 4. 증권거래용 숫자 음 음성 데이터베이스(DB1)의 구성

단어형태(음절수)	개수	예
숫자음(1)	9	일, 이, 삼, …
숫자음(1) + '원'	5	십원, 백원, 천원, …
숫자음(1) + '주'	5	십주, 백주, 천주, …
숫자음(1) + 단위음(1) + '원'	41	이십원, 삼백원, …
숫자음(1) + 단위음(1) + '주'	41	사십원, 사백주, …
단위음(1)+숫자음(1)+ 단위음(1)+'원'	59	십오만원, 백육십원, …
단위음(1)+숫자음(1)+ 단위음(1)+'주'	59	백칠십원, 만사천주, …
합계	219	-

표 5. 단위 숫자 음 인식용 음성데이터베이스(DB2)의 구성

단어형태(음절수)	개수	예
단위음(1)	7	십, 백, …
숫자음(1)	9	일, 이, 삼, …
숫자음(1)+단위음(1)	45	이십, 삼십, …
단위음(1)+숫자음(1)	45	십일, 십이, …
단위음(1)+단위음(1)	20	백십, 만천, …
숫자음(1)+'원'	9	일원, 이원, …
단위음(1)+'원'	5	십원, 백원, …
숫자음(1)+'주'	9	일주, 이주, …
단위음(1)+'주'	5	십주, 백주, …
합계	154	-

4.2 실험 결과

4.2.1 두 음절 FSN과 반음절 쌍 음향모델에 의한 인식성능의 변화

본 논문의 실험을 위한 인식기로는 HTK(Hidden Markov Toolkit)[3]기반의 인식기를 사용 하였으며, 두 음절 FSN의 타당성을 검증하기 위하여 기존의

표 6. 음성 데이터베이스의 구성

내용 \ DB	DB1	DB2	DB1 + DB2
화자 수	200	150	350
단어 수	**219**	**154**	354
인식범위	10 ~ 9천9백9십9억 9천9백9십9만 9천9백9십	1 ~ 9천9백9십9억 9천9백9십9만 9천9백9십9	1 ~ 9천9백9십9억 9천9백9십9만 9천9백9십9
인식간격	10배수	1배수	1배수
인식대상 수	인식범위/10	인식범위	인식범위
DB size	12K*219*화자수	12K*154*화자수	12K*219*화자수 + 12K*154*화자수
sampling rate	8k	8k	8k
resolution	16bit	16bit	16bit
mono-phone 개수	27	27	27
tri-phone 개수	**190**	**243**	281

표 7. 음성 데이터베이스의 음절별 등장 횟수

	일	이	삼	사	오	육	칠	팔	구	십	백	천	만	억	주	원	합계
DB1	5	23	23	23	23	23	23	23	23	95	78	75	98	2	105	105	747
DB2	13	13	13	13	13	13	13	13	13	29	29	29	29	29	15	15	292

표 8. 테스트 음성 데이터베이스

	금액 단위 숫자음	거래량 단위 숫자음
1	구만사천원	구만사천주
2	구만사천삼백오십사원	구만사천삼백오십사주
3	칠만오천원	칠만오천주
4	칠만오천오백육십이원	칠만오천오백육십이주
5	오십육만원	오십육만주
6	오십육만구천이백삼십이원	오십육만구천이백삼십이주
⋮	⋮	⋮
27	십육만삼천원	십육만삼천주
28	십육만삼천삼백팔십오원	십육만삼천삼백팔십오주
29	삼천사백이십만원	삼천사백이십만주
30	삼천사백이십만오천삼백이십원	삼천사백이십만오천삼백이십주

FSN[1,2]과 비교실험 하였다. 기존의 FSN과 제안된 두 음절 기반 FSN의 인식성능을 비교하기 위하여 모노폰, 트라이폰, 반음절 쌍 음향모델에서 각각 비교 실험한 결과는 표 10과 같다. 반음절 쌍 음향모델의 경우 기존의 FSN에서는 음향모델 단위가 언어모델보다 크므로 실험대상에서 제외시켰다.

본 실험에서 단어 인식률은 문장 내 포함된 단어의 개수 274개에 대한 인식률이며, 문장 인식률은 단위 숫자음 60문장에 대한 인식률이다. 실험결과, 두 음절 기반 FSN은 사전수가 15개에서 120개로 증가

표 9. 데이터베이스의 신호 처리

설정		설정값	
Environment		조용한 사무실	
Sampling rate		8khz	
Quantization		16bit	
Preemphasis coefficient		0.97	
Window	Shape	Hamming	
	Size	25 ms	
	Shift step	10 ms	
Feature	MFCC	26	12
	Normalized energy		1
	1st order derivatives		13

표 10. 제안된 두 음절 FSN과 기존 FSN의 인식률

음향모델 / FSN	사전수	mono-phone		tri-phone		demi-syllable pair	
		단어	문장	단어	문장	단어	문장
기존의 FSN	15	88.0	55.2	88.4	66.0	-	-
제안된 FSN	120	86.5	69.8	89.1	76.5	97.6	89.2

표 11. 변형된 연쇄 상태 분할에 의한 인식률(%)

문맥 / 상태수		11000	14000	17000	20000	23000
SSS	단어	95.4	98.0	98.2	95.2	94.5
	문장	87.8	89.6	91.0	80.7	77.8
Transformed SSS	단어	97.2	97.4	98.5	94.8	-
	문장	88.7	90.3	92.5	82.3	-

하였지만, 음향모델이 동일할 경우 문장인식률에서 모노폰은 14.6%, 트라이폰은 10.5% 인식 성능이 향상됨을 알 수 있었다. 단어인식률의 변화는 문장인식률의 변화와 상이한 결과를 보였는데, 이는 단어의 수가 기존의 FSN은 274×2이고, 제안된 FSN은 274이며, 제안된 FSN에서는 기존 FSN의 한 단어에 해당하는 한 음절만 오 인식 되었을 경우 단어의 오 인식으로 나타나므로 문장 인식률의 변화와 상관관계가 없기 때문이다.

4.2.2 변형된 연쇄 상태 분할에 의한 모델의 인식 성능 변화

일반적인 연쇄 상태 분할에 의한 모델과 변형된 연쇄 상태 분할에 의한 모델 개선 후, 상태 수에 따른 인식률의 변화를 알아보았다. 본 실험에서는 앞서 실험 결과 인식성능이 우수했던 반음절 쌍 음향모델을 이용하여 상태 분할 후, 두 음절 FSN 언어모델을 적용하여 인식실험을 하였다.

실험결과, 일반적인 연쇄 상태 분할 방법으로 상태를 분할하여 모델링 하였을 경우 상태수가 17000개일 때 최대 문장 인식률 91.0%를 보였으며, 그 이상 분할하였을 경우 인식률이 떨어졌다. 변형된 연쇄 상태 분할에 의해 모델링 하였을 경우 최대 문장 인식률 92.5%로 상태수가 같을 때 1.5%의 인식 성능이 향상됨을 확인했다. 실험결과 두 음절 기반 FSN으로 언어모델을 개선하였을 때, 사전수가 8배로 늘어났음에도 불구하고, 문맥종속 음소모델에서 문장 인식률이 10.5% 증가함을 확인할 수 있었다. 조음 및 변이음 특성이 강하고, 분절이 어려워 오 인식률이 높은 한국어 숫자음의 경우 인식 단위를 음소보다 긴 구간인 반 음절 쌍으로 모델링 함으로서 문맥종속적인 음소모델보다 최고 12.7%의 인식성능을 향상시

킬 수 있었다. 반음절 쌍 모델링에 있어서 변형된 상태분할에 의하여 문장 인식률을 3.3% 높일 수 있었다. 상태 분할은 모델별 초기 모델의 상태수와 유사도의 변화를 고려하여 3000회씩 간격을 두고 진행한 후 인식 실험을 하였다. 유사도의 가장 큰 기준은 인식률이므로 인식률이 가장 좋은 곳을 최적의 지점이라고 판단하여 이를 이용하였다. 상태수를 증가시킬수록 원래의 고정된 상태 모델보다 향상된 인식률을 갖는 모델을 생성할 수 있었다. 그러나 상태수를 17000개 이상 증가시킬 경우 데이터의 부족으로 인해 인식률이 큰 폭으로 떨어졌다.

반음절 쌍 모델은 앞부분과 끝부분이 다른 모델과 특징을 공유하는 형태이므로, 인식 단 에서 이러한 특성을 이용하여 인식후보 모델의 수를 제한할 수 있다. 전후 인식모델의 관계가 한정됨으로써 인식 단에서 마치 벽돌을 쌓는 방법처럼 인식된 결과 모델에 대한 상관관계를 확인하여 검증작업이 가능하게 되는 것이다. 향후 반음절 쌍 모델 간 특징레벨에서의 공유성을 이용하여 유사도의 평가가 이루어진다면, 인식 성능의 향상에 기여할 것으로 보인다.

5. 결 론

본 논문에서는 단위 숫자음의 인식을 위해 독립적인 방법으로 모델을 개선하였다. 한국어 연속 단위 숫자음의 인식 성능을 향상시키기 위해 언어모델과 음향모델을 개선하기 위한 방법을 제안하였다. 먼저 언어모델을 개선하기 위해 한국어 단위 숫자음을 분석하여 단위 숫자 음은 한 음절의 숫자음과 한 음절의 단위음으로 구성된 단어로 이루어진 문장이라는 규칙성을 발견하여 FSN 언어모델에 적용하였다. 다음으로 한국어 숫자음의 특성상 모든 숫자음이 단음절로 구성되어 있고, 음절과 음절 사이에 조음이 많아 음소, 음절의 경계 분할이 부정확하여 오 인식률이 높은 특성을 보완하기 위하여 비교적 음향적 특성이 강한 모음구간에서 경계를 분할하는 반음절 쌍 인식모델을 이용하였고, 변이음적 특성을 효과적으로 모델링하기 위하여 K-means 알고리즘을 이용한 연쇄 상태 분할을 하여 음향모델을 개선하였다.

훈련 데이터베이스와 테스트 데이터베이스를 각각 구성하여 실험하였으며 제안된 두 음절 기반 FSN 언어모델을 다양한 음향모델에서 실험한 결과 같은 음향모델일 경우 사전수의 증가에도 불구하고 인식 성능이 향상됨을 확인할 수 있었다. 또한 반음절 쌍 음향모델을 그룹으로 나누어 초기모델을 작성한 후 분산 값을 기준으로 분할 대상을 결정하고 전체 상태수를 일정 간격으로 두고 상태 분할과 인식률 측정을 반복한 결과 고정된 상태 모델에 비하여 인식 성능이 향상되었음을 알 수 있었다. 상태수를 증가시킬수록 인식률이 향상되었으나 임계값 이상 분할하였을 경우 데이터의 부족으로 인해 인식률이 큰 폭으로 떨어졌다.

본 연구 결과를 증권거래, 홈쇼핑 등의 음성기반 상거래에 적용할 경우 대 어휘, 대화형 음성인식에 효과적인 기존의 구성과 단위 숫자음 인식에 있어서 독립적으로 최적화된 구성을 상호혼용 및 결합하면, 전체 시스템의 성능향상에 기여할 것으로 기대된다. 본연구의 실험을 웹 기반 클라이언트/서버 환경에서 이루어져 앞으로 전화기 및 휴대전화기를 이용한 이동환경에서의 다양한 모델의 혼용을 위한 연구가 필요하다.

참 고 문 헌

[1] X. Huang, A. Acero, and H.W. Hon, *Spoken language processing*, Prentice Hall PTR, New Jersey, 2001.

[2] Daniel Jurafsky & James H. Martin, *SPEECH and LANGUAGE PROCESSING*, Prentice Hall, New Jersey, 2002.

[3] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Ver.3.2)*, Cambridge University Engineering Department, 2002.

[4] L.R. Rabiner and B.H. Juang, *Fundamentals of speech recognition*, Prentice Hall, New Jersey, 1993.

[5] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257~286, Feb. 1989.

[6] 윤재선, 홍광석, "반음절 단위 HMM을 이용한 연속 숫자 음성인식," 한국음향학회지 제17권 제5호, pp. 73~78, 1998.

[7] J. Takami, S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *ICASSP-92*, pp. 573~576, Mar. 1992.

[8] A. Kannan, M. Ostendorf, and J.R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 3, pp. 453~455, Jul. 1994.

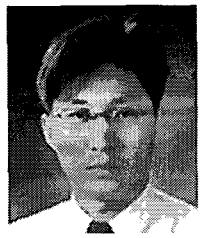
[9] 오세진, 황철준, 김범국, 정호열, 정현열, "결정 트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘을 이용한 음성인식에 관한연구," 한국음향학회 제21권 제2호, pp. 200~202, 2002.

이 수 정

1992년 3월~1997년 2월 한국방송통신대학교 전자계산학과 졸업 (이학사)
1997년 3월~2000년 2월 광운대학교 공학석사
2004년 3월~현재 광운대학교 박사과정(음성신호처리)전공
1993년 3월~2000년 4월 한국방송통신대학교 교육매체 개발연구소
2001년 9월~2004년 8월 대한상공회의소 경기인력개발원 전임강사
2004년 3월~2005년 2월 서정대학 인터넷정보과 겸임교수
2005년 7월~현재 (주)APET 부설 기술연구소 기술개발팀장/책임연구원

서 은 경

2002년 광운대학교 전자물리학과 졸업
2004년 광운대학교 대학원 컴퓨터공학과 졸업



최 갑 근

2001년 2월 광운대학교 정보통신대학원 졸업(공학석사)
2004년 3월 광운대학교 대학원 컴퓨터공학과(박사과정)
1999년 3월 (주) 이오리스 부설 기술연구소
2003년 10월 (주)삼성블루텍 S/W 개발그룹
2004년 1월 GM Corporation 개발팀

김 순 협

1974년 2월 울산대학교 전자공학과 (공학사)
1976년 2월 연세대학교 대학원 전자공학과(공학석사)
1983년 2월 연세대학교 대학원 전자공학과(공학박사)
1979년 3월~현재 광운대학교 컴퓨터공학과 교수
2000년 12월 한국음향학회 회장역임
2001년 1월~현재 한국음향학회 명예회장