반음절쌍과 변형된 연쇄 상태 분할을 이용한 연속 숫자음 인식의 성능 향상

김동옥* · 박노진**

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

Dong-Ok Kim* · No-Jin Park**

이 논문은 2005년도 정보통신기능대학 연구비를 지원받았음

요 약

본 논문에서는 언어모델과 음향모델을 개선함으로써 단위 숫자음의 인식성능 최적화에 대해 설명한다. 언어모델은 한국어 단위 숫자음 문장의 문법적 특징을 분석하고, FSN 노드를 두음절로 구성하여 오 인식률을 감소시켰다. 음향모델은 단음절로 구성되어 발성기간이 짧고 조음이 많이 생기는 불명확한 음소, 음절의 분할로 인한 오 인식을 줄이기 위해 인식단위를 반음절쌍으로 하였다. 인식단위의 특징을 효과적으로 모델링하기 위해 특징레벨에서 K-means 알고리즘[4]으로 클러스터링 하여 상태를 분할하는 변형된 연쇄 상태 분할방법을 이용하였다. 실험 결과 제안된 언어모델의 적용 후 동일 문맥종속 음소모델에서 10.5%, 음향모델에서 인식단위를 반음절쌍으로 하였을 경우 문맥종속 음소모델에 비해 12.5%, 변형된 연쇄 상태분할을 하였을 경우 1.5%의 인식률을 향상시킬 수 있었다.

ABSTRACT

This paper describes an optimization of a language model and an acoustic model that improve the ability of speech recognition with Korean unit digit. Recognition errors of the language model are decreasing by analysis of the grammatical feature of korean unit digits, and then is made up of fsn-node with a disyllable. Acoustic model make use of demi-syllable pair to decrease recognition errors by inaccuracy division of a phone, a syllable because of a monosyllable, a short pronunciation and an articulation. we have used the k-means clustering algorithm with the transformed successive state splitting in feature level for the efficient modelling of the feature of recognition unit . As a result of experimentations, 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.

키워드

Demi-syllable pair, Transformed successive state splitting

* 한국정보통신기능대학

** 서정대학

접수일자 : 2005. 6. 20

Ⅰ. 서 론

본 논문에서 이용한 방법은 언어모델과 음향모델의 개선이다. 먼저 언어모델을 개선하기 위하여 한국어 단위 숫자음 문장의 구조를 파악하고, 문맥적 규칙을 발견하여 적용하였다. 단위 숫자음 문장이 숫자음 한 음절과 단위음 한음절로 구성된 단어의 반복적인 발성으로 나타남을 인식하고, FSN 노드를 두음절로 구성하여 단음절로 이루어진 한국어 단위 숫자음 인식에 있어서의 오 인식 요인을 줄이고자 하였다.

음향모델의 개선을 위해서는 한국어 숫자음과 같이 단음으로 구성되어 발성구간이 짧고, 연속음성인식에서 부정확한 음절, 음소의 구분으로 인한 오 인식을 줄이기 위하여 반음절쌍(demi-syllable pair) 인식단위로 모델링하고, 반음절쌍 모델같이 하나의 모델에 변이음적 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.

Ⅱ. 한국어 연속 숫자음 모델

2.1 단음절 FSN

FSN은 인식하고자 하는 단어의 연결 관계를 네트웍으로 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 비행기표 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 1 은 FSN의 예를 보였다.[2]

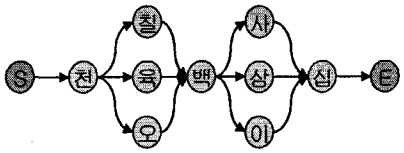


그림 1. FSN의 예
Fig. 1 Example of FSN

2.2 반음절쌍 음향모델 제안

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

음절의 구분이 명확하지 않은 제한된 도메인에서의 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 모델링 하여 반음절쌍(demi-syllable pair) 모델을 구성한다. 그림 2. 에서 반음절쌍 모델의 모습을 보여준다.

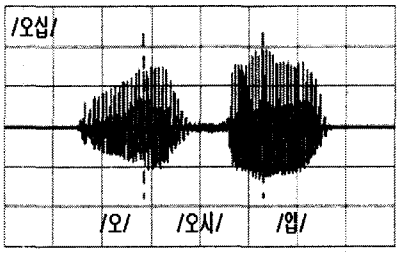


그림 2. 반음절쌍 모델의 예 /오십/
Fig 2. Example of Demi-syllable Pair Model /오십/

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

표 1. 반음절쌍 모델의 구성
Table. 1 Composition of Demi-syllable Pair Model

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

2.2.1 인식과정

연속 단위 숫자음 인식에서 인식 대상이 되는 가변 숫자음 문장의 인식은 반음절쌍 모델에 의하여 그림 3와 같은 구조로 나타난다. 문장의 처음과 끝은 반음절로 나타나며 중간은 반음절쌍이 반복적으로 나타난다. 그림 3 (a)는 반음절쌍 모델기반 FSN을 보여주고 있으며,(b)는 단위 FSN 이 인식하는 두음절 숫자음을 보여주고 있다.

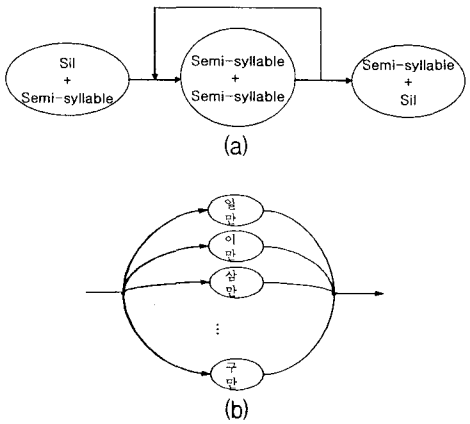


그림 3. FSN 구조
Fig. 3 Structure of FSN

표 2. 반음절쌍 인식단위
Table. 2 Definition of Demi-syllable Pair Recognition Unit

구분 / 단어	인식단위				
	반음절	반음절쌍	반음절쌍	반음절쌍	반음절
일	S1D	-	-	-	E1D
이	S2D	-	-	-	E2D
십	STD	-	-	-	ETD
백	SBD	-	-	-	EBD
삼십	S3D	E3DSTD	-	-	ETD
사백	S4D	E4DSBD	-	-	EBD
육만	S6D	E6DSMD	-	-	EMD
칠백사	S7D	E7DSBD	EBDS4D	-	E4D
오십육	S5D	E5DSTD	ETDS6D	-	E6D
팔만삼천	S8D	E8DSMD	EMDS3D	E3DSCD	ECD
구십주	S9D	E9DSTD	ETDSJD	-	EJD
백원	SBD	EBDSWD	-	-	EWD

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

정의된 모델과 인식과정을 연속 단위 숫자음 문장 "구만 사천 원"에 적용하여 그림 4에 나타내었다.

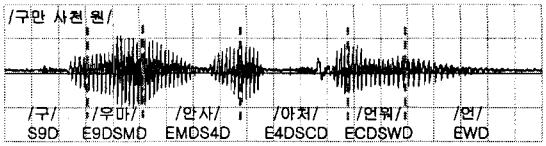


그림 4. 연속 단위 숫자음 /구만사천원/에서 반음절쌍의 예
Fig. 4 Example of Continuous Unit Digits /구만사천원/

Ⅲ. 연쇄상태분할

3.1 변형된 연쇄상태 분할

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

3.1.1 초기 모델 학습

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

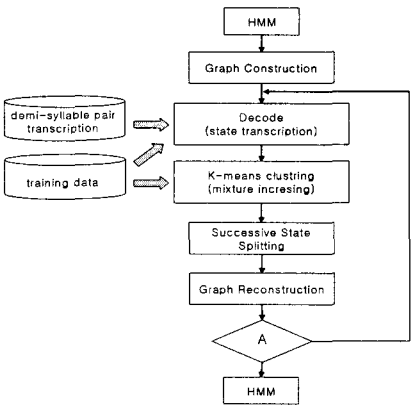


그림 5. 변형된 연쇄 상태 분할 과정
Fig. 5 Transformed Successive State Splitting

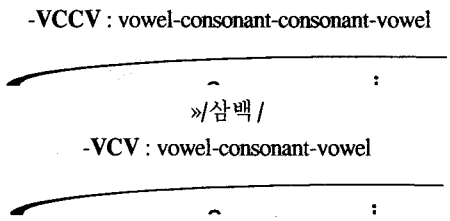


그림 6. 반음절쌍모델
Fig. 6 Demi-syllable model

다음으로 아래와 같은 단계를 거쳐 그룹별로 초기 모델을 학습한다.

단계1.

6가지로 분류한 모델에서 각각의 음소(phone)의 개수를 클러스터 개수 K로 정하고 K-means 알고리즘[4]으로 클러스터링 한다. 표3에서 구분한 모델의 음소의 개수로 K개의 임의의 클러스터 센터 $Z_1(1), Z_2(1), \ldots, Z_k(1)$을 설정하고 Euclidean 거리 측정법에 의해 ①클러스터 중심값과 벡터의 거리값 $Distance = \sqrt{(X_2 - X1)^2 + (Y_2 - Y_1)^2}$ 을 구하여 각각의 벡터를 클러스터에 분배한다. ②클러스터링 된 벡터들과 중심점과의 거리의 평균값 $Z_j(t+1) = \frac{1}{N_j}\sum x$을 기준으로 새로운 클러스터 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($Z_j(t+1) = Z_j(t)$) 클러스터링 과정을 끝낸다.

3.1.2 K-means에 의한 variation and temporal modeling

초기 모델을 연쇄 상태 분할하여 최적의 상태 네트워크를 가지는 모델을 구성하고자 한다. SSS와의 차이점은 각각의 반음절쌍 모델이 포함하는 음소의 개수로 K-means 클러스터링 하여 구한 가우시안 각각에 상태를 할당하고, 시간방향으로 정렬하여 모델별로 초기 모델을 정의하고, 2혼합수 가우시안을 구할 때 K-means 알고리즘으로 클러스터링 하였다는 점이다.

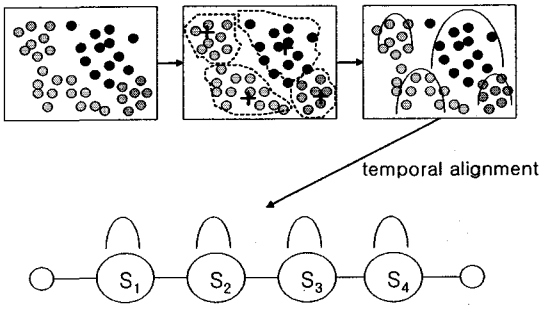


그림 6. VCCV 그룹의 초기 모델
Fig. 6 Training of an initial model

1혼합수 가우시안을 가지는 상태를 유사도에 따라 발음특성이 반영된 변이방향 혹은 시간방향으로 배열함으로써, 변이음과 이중모음 등의 형태가 나타나는 반음절쌍 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 7은 K-means 클러스터링에 의한 가우시안 상태 분할 과정을 보여주고 있다.

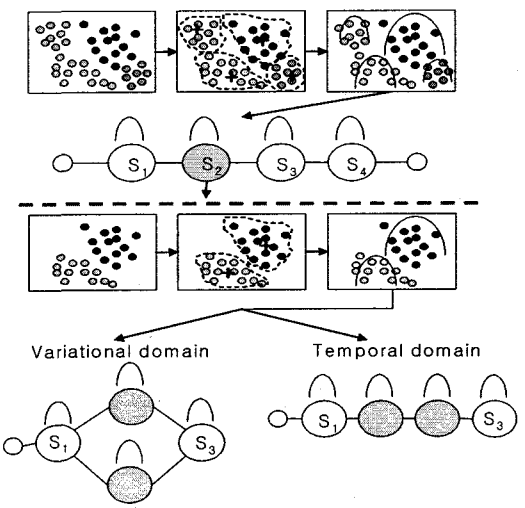


그림 7. K-means 클러스터링에 의한 변형된 연쇄 상태분할
Fig. 7 Transformed State Splitting by K-means Clustering

Ⅳ. 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 3, 표 5와 같다.

표 3. 증권거래용 숫자음 음성 데이터베이스(DB1)의 구성
Table.3 Digits Speech Database for Stock Trading(DB1)

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 4. 단위 숫자음 인식용 음성데이터베이스(DB2)의 구성
Table. 4 Speech Database for Unit Digits(DB2)

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	-

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

표 5. 음성 데이터베이스의 구성
Table. 5 Composition of Speech Database

DB / 내용	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 6. 음성 데이터베이스의 음절별 등장 횟수
Table.6 Number of Speech Database by Syllable

	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

4.2 실험 결과

4.2.1 두음절 FSN과 반음절쌍 음향모델에 의한 인식성능의 변화

본 논문의 실험을 위한 인식기로는 HTK(Hidden Markov Toolkit)[3]기반의 인식기를 사용 하였으며, 두음절 FSN의 타당성을 검증하기 위하여 기존의 FSN[1][2]과 비교실험 하였다. 기존의 FSN과 제안된 두음절 기반 FSN의 인식성능을 비교하기 위하여 모노폰, 트라이폰, 반음절쌍 음향모델에서 각각 비교 실험한 결과는 표7

과 같다. 반음절쌍 음향모델의 경우 기존의 FSN에서는 음향모델 단위가 언어모델보다 크므로 실험대상에서 제외시켰다.

본 실험에서 단어 인식률은 문장내 포함된 단어의 개수 274개에 대한 인식률이며, 문장 인식률은 단위 숫자음 60문장에 대한 인식률이다. 실험결과, 두음절 기반 FSN은 사전수가 15개에서 120개로 증가하였지만, 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 오인식 되었을 경우 단어의 오 인식으로 나타나므로 문장 인식률의 변화와 상관관계가 없기 때문이다.

표 7. 제안된 두음절 FSN과 기존 FSN의 인식률
Table. 7 Recognition rate of the proposed Disyllable and existing FSN

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

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, 상태수에 따른 인식률의 변화를 알아보았다. 본 실험에서는 앞서 실험결과 인식성능이 우수했던 반음절쌍 음향모델을 이용하여 상태 분할 후, 두음절 FSN 언어모델을 적용하여 인식실험을 하였다.

표 8. 변형된 연쇄 상태 분할에 의한 인식률(%)
Table. 8 Recognition Rate by Transformed Successive State Splitting

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	-

실험결과 두음절 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 반음절쌍으로 모델링함으로서 문맥종속적인 음소모델보다 최고 12.7%의 인식성능을 향상시킬 수 있었다. 반음절쌍 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.

향후 반음절쌍 모델 간 특징레벨에서의 공유성을 이용하여 유사도의 평가가 이루어진다면, 인식 성능의 향상에 기여할 것으로 보인다.

V. 결 론

본 논문에서는 단위 숫자음의 인식을 위해 독립적인 방법으로 모델을 개선하였다. 한국어 숫자음의 특성상 모든 숫자음이 단음절로 구성되어 있고, 음절과 음절 사이에 조음이 많아 음소, 음절의 경계 분할이 부정확하여 오인식률이 높은 특성을 보완하기 위하여 비교적 음향적 특성이 강한 모음구간에서 경계를 분할하는 반음절쌍 인식모델을 이용하였고, 변이음적 특성을 효과적으로 모델링하기 위하여 K-means 알고리즘을 이용한 연쇄 상태 분할을 하여 음향모델을 개선하였다.

제안된 두음절 기반 FSN 언어모델을 다양한 음향모델에서 실험한 결과 같은 음향모델일 경우 사전수의 증가에도 불구하고 인식성능이 향상됨을 확인할 수 있었다. 또한 반음절쌍 음향모델을 그룹으로 나누어 초기모델을 작성한 후 분산값을 기준으로 분할 대상을 결정하고 전체 상태수를 일정 간격으로 두고 상태 분

할과 인식률 측정을 반복한 결과 고정된 상태 모델에 비하여 인식 성능이 향상되었음을 알 수 있었다.

본 연구 결과를 증권거래, 홈쇼핑 등의 음성기반 상거래에 적용할 경우 대어휘, 대화형 음성인식에 효과적인 기존의 구성과 단위 숫자음 인식에 있어서 독립적으로 최적화된 구성을 상호혼용 및 결합하면, 전체 시스템의 성능향상에 기여할 것으로 기대된다.

참고문헌

[1] X. Huang, A. Acero, H.W. Hon, "Spoken language processing", Prentice Hall PTR, New Jersey, pp.1-5,558-560,655 2001.

[2] Daniel jurafsky & James h. Martin, "SPEECH and LANGUAGE PROCESSING", Prentice Hall, New Jersey, p.33-53, 2002.

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

[4] L.R. Rabiner, B.H. Juang, "Fundamentals of speech recognition", Prentice Hall, New Jersey, chap. 6,pp.15-23,125-128,321-324 1993.

[5] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE , Volume: 77 Issue: 2 , pp. 257 -286, Feb. 1989.

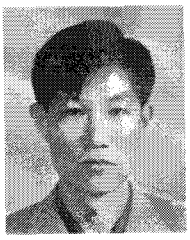
[6] J. Takami, S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling", ICASSP-92., p. 573 -576, Mar., 1992.

[7] A. Kannan, M. Ostendorf, J.R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition", Speech and Audio Processing, IEEE Transactions on , Volume: 2 Issue: 3 pp.453 -455, Jul. 1994.

저자소개

김 동 옥(Dong-Ok Kim)

서울산업대학교 전자공학과 공학사
광운대학교 전자통신공학과 공학석사
한국항공대학교 통신정보공학과 공학박사
현재 : 한국정보통신기능대학 교수
※관심분야 : 이동통신시스템, 디지털통신시스템, W-CDMA, 디지털 통신시호처리



박 노 진(No-jin Park)

서울산업대학교 매체공학과 공학사
광운대학교 전자통신공학과 공학석사
광운대학교 전자통신공학과 공학박사
현재 : 서정대학 정보통신과 전임강사
※관심분야 : 통신신호처리, 음성인식, 채널 코딩, DMB, WCDMA