전화망에서의 한국어 연속숫자음 인식 실험

강점자, 김갑기
한국전자통신연구원, 네트워크 연구소, 음성정보연구센타

The Recognition Experiment of Korean Connected Digit in the Telephone Network

Jeom-Ja Kang, Kap-kee Kim

Speech Technology Research Center, Network Laboratory, ETRI
E-mail : jjkang @etri.re.kr

요 약

본 논문에서는 전화망 환경에서의 한국어 숫자음 인식을 위한 특징 파라미터 추출, 음향 모델링 방식을 결정하기 위하여 HTK 툴을 사용한 4 연숫자음 인식 실험 결과를 기술한다. 또한, 실험 결과를 토대로 빈번하게 발생하는 숫자음에 대해서 오류율을 분석하였다. 숫자 모델로는 left context biword 모델과 triword 모델을 사용하였으며, 상태수와 mixture 수를 바꾸어 인식 실험을 수행한 결과, triword 모델이 biword 모델보다 인식율이 높은 것으로 나타났으며, substitution 에러율은 " 이<->일" 에서 가장 높은 에러가 발생하는 결과를 얻을 수 있다.

1. 서론

음성정보처리산업은 인간과 기계와의 인터페이스 수단으로 음성을 매개로 하는 정보처리 기술을 이용하여 다양한 정보전달 서비스를 개발하고, 제공함으로써 기업의 부가가치를 창출하고 인간 생활의 질을 향상시키는 미래산업[1]으로 다보스 포럼 및 MIT 의 미래예측에서 21 세기 정보화사회를 선도하는 10 대 유망 기술로 선정된 바 있다. 또한 음성을 이용하여 다양한 응용분야(전화망, 인터넷,단말기, PDA, 홈오토, 가전등)에 적용되고 있다.

그러나 다양한 분야에서 음성을 응용하여 적용하려는 시도가 활발히 이루어지고 있으나, 인식엔진의 성능저하로 인하여 사용자의 요구사항에 밀도는 수준이다. 특히 한국어 연속숫자음을 이용한 음성응용서비스(예, 음성다이얼링, 음성인식텔레뱅킹 등)에서 주민등록번호, 카드번호, 회원 고유번호, 전화번호, 계좌번호, 금액 등을 많이 이용하게 되는데, 한국어 숫자음이 단음절로 되어 있어 상호간 유사한 발음이 많은 특성으로 인하여 인식에 어려움이 많다. 현재 연속 숫자음 인식을 수행하고 있는 일부 국내 업체는 외국 기술을 도입하여 연속 숫자음 인식을 수행하고 있는 실정이다.

본 논문에서는 전화망 환경에서의 한국어 연속숫자음 인식에 적합한 특징 파라미터 추출 및 인식단위 선정을 위하여 HTK 툴을 사용한 실험결과와 실험 결과를 바탕으로 빈번하게 발생하는 숫자음에 대해서 오류율을 분석하였다. 실험에 사용한 숫자음 모델로는 left context biword 모델과 triword 모델을 사용하였다. 그리고, 각각의 모델별로 상태수와 mixture 수를 다르게 구성하여 인식 실험을 실시하고, 인식 성능을 비교 검토하였다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 숫자음 인식 실험 과정에 대해서 기술하고, 제 3 장에서는 biword 모델과 triword 모델 별로 인식 실험 결과와 결과 분석을 기술한다. 제 4 장에서는 결론을 맺는다.

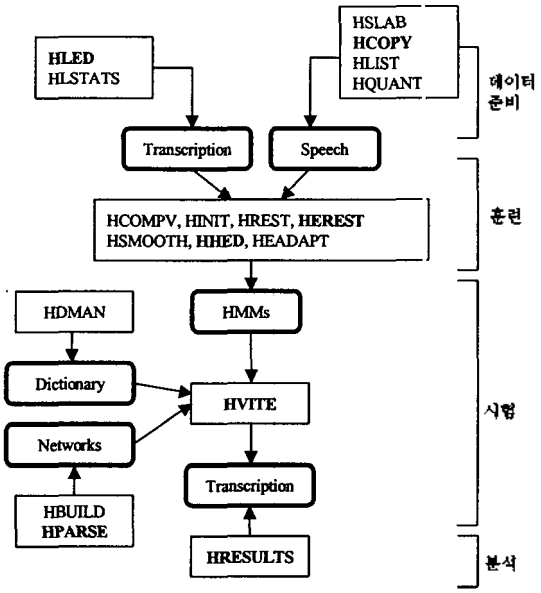
2. 숫자음 인식 실험 과정

본 논문에서 숫자음 인식실험을 위해 사용한 인식기는 연속확률분포 HMM(Hidden Markov Model)을 이용한 HTK 툴을 사용하였다. 따라서 본 장에서는 HTK 툴의 개요와 인식 실험 과정을 기술한다.

2.1 HTK 툴 개요

HTK 는 HMM 을 만들기 위한 툴로, 음성 데이터를 사용하여 인식 결과를 시뮬레이션 하기 위해 일반적으로 사용되어지는 툴이다. 즉 이것은 동일한 음성 데이터와 전사(transcription)파일을 사용하여 자기가 만든 음성인식기의 결과와 HTK 툴 사용한 음성 인식 결과를 비교해 볼 수 있는 툴로 사용될 뿐만 아니라, 본 논문에서와 같이 특징 추출 파라미터를 결정하거나,

음향모델링 결정을 위해 요긴하게 사용되어지는 툴이다. HTK 는 현재까지 3.0 버전이 나와 있으나, 본 논문에서는 triword 모델 인식 실험을 위해서는 windows 2.2 를 사용하고, biword 인식 실험을 위해서는 Linux 2.2 를 사용하였다. 버전 2.2 와 3.0 간에는 적응(adaptation) 기능을 제외하고는 동일하다. 다음의 (그림 1)은 HTK 툴에서 인식 실험을 하기 위한 처리 단계를 나타낸 것이다. 처리 단계는 크게 데이터 준비, 훈련, 시험, 분석 단계로 나누어지고, 각 단계별 실험 과정은 2.2 절에서 기술한다. 각 단계별로 요구되는 툴의 기능은 HTK 매뉴얼[2]을 참조하면 어려움없이 수행가능하다. 굵게 표시한 명령어는 본 실험을 위해 사용한 명령어를 나타낸 것이다.



(그림 1) HTK 도구의 처리 단계

2.2 인식 실험 과정

HTK 를 사용하여 biword, triword 모델 각각에 대해 인식 실험을 하기 위해서는 먼저 실험에 사용할 데이터를 준비하고, 준비된 데이터에 대해서 특징 파라미터를 추출하고, 전사파일을 준비한다. 특징 파라미터 추출과 전사파일 생성이 완료되면 mono-word 모델 훈련, biword 또는 triword 모델의 훈련, tired-state biword 또는 tired-state triword 모델의 훈련/인식/결과, biword 또는 triword 모델 각각에 대해 mixture 수를 다르게 구성하여 훈련/인식/결과를 출력하면 인식 실험 과정이 완료된다. 상태수를 다르게 하여 실험하고자 할 때에는, 상태 수를 변경하여 위의 과정을 반복하여 실험한다. 본 실험에서의 상태 수는 3 개, 5 개, 7 개, 9 개, 11 개를 사용하였다.

1) 인식 실험 데이터 준비

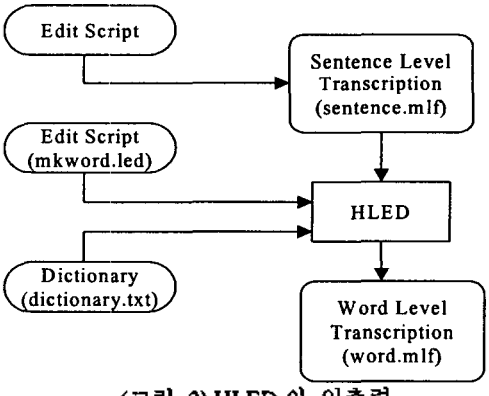
인식 실험에 사용한 숫자음 데이터베이스는 보이스웨어 데이터베이스로 전화망 환경에서 연속되는 4 개의 숫자음을 녹음한 파일로 나이는 20 대, 30 대, 40 대 남녀로 구분된다. 숫자 음성 DB 로는 영, 공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구의 11 개의 숫자 음성이다. 데이는 8KHz u-law 데이터를 8KHz 16bit PCM 데이터로 변환한 것이다. 실험에 사용한 훈련 데이터수는 49,799 개, 시험(test) 데이터 수 10,068 개이다.

2) 음성 특징 파라미터 추출

8KHz 16bit PCM 데이터로부터 음성 특징 파라미터를 추출하기 위해서는 configuration 파일의 각 아이템에 적당한 값을 설정하고, HTK 의 Hcopy 명령어를 이용한다. Configuration 파일에서 사용한 음성 특징 파라미터는 TARGETKIND=MFCC_0_Z_D_A 로 구성하였다. 이것은 Static MFCC(Mel Frequency Cepstral Coefficients) 13 차(MFCC_0)와 Delta 13 차(_D), Acceleration 13 차(_A)를 합한 총 39 차를 사용하고, 채널왜곡 보상을 위해 CMN(Cepstral Mean Normalisation, _Z)을 적용하였다. Pre-emphasis 는 0.97 로 지정하고, 윈도우 사이즈는 20msec, 윈도우 오버랩은 10msec 씩 이동하며 추출하였다.

3) 전사(transcription)파일 생성

word 모델에서 가장 기본이 되는 mono-word 모델에 대해서 훈련을 하기 전에 (그림 1)에서와 같이 word 단위의 전사파일이 필요하다. 전사파일의 생성은 HLED 명령어를 사용하면 되는데, 먼저 4 연 숫자음이 발성된 문장(sentence) 단위의 전사파일과 발음사전을 준비해야 한다. 다음의 (그림 2)는 전사파일을 생성하기 위해 필요한 입력파일과 전사파일 생성이 끝난 후의 출력파일을 나타낸 것이다.

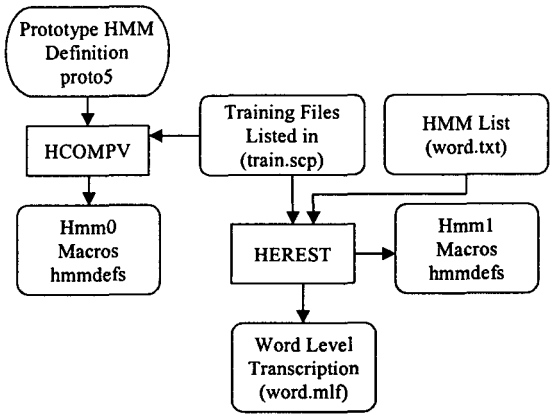


(그림 2) HLED 의 입출력

4) mono-word 모델 훈련

mono-word 모델의 훈련 이전에 HMM 모델의 프로토타입을 정의해야 한다. 프로토타입을 정의할 때에 상태수 3,5,7,9,11 에 대해서 실험하고자 한다면, 각각의 상태 수에 적합한 프로토타입을 정의한다. 프로토타입이 정의되면 hmm 모델(hmmdefs)과 macro 를 정의하여 훈련한다. Hmm 모델에는 숫자음 11 개와 silence 1 개를 포함하여 12 개를 사용한다. 다음의 (그림 3)은

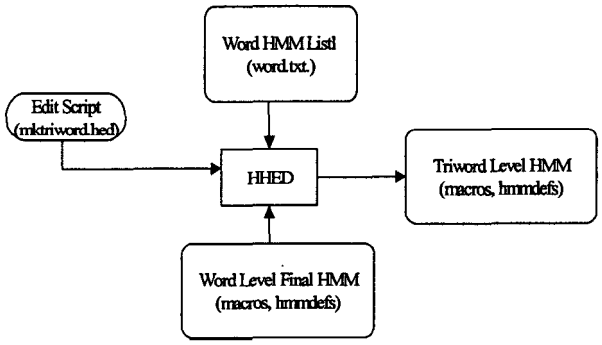
mono-word 모델을 훈련하기 위한 입출력과 사용하는 명령어를 나타낸 것이다.



(그림 3) mono-word 모델 훈련을 위한 입출력

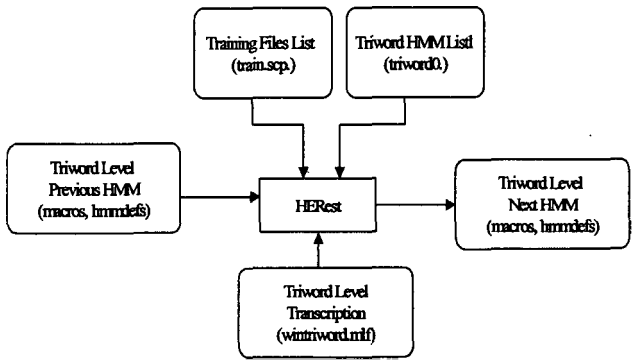
5) biword/triword 모델의 훈련

biword 또는 triword 모델의 훈련 이전에 mono-word 모델의 훈련 단계에서와 같이 biword 또는 triword 단위의 전사파일이 필요하다. 이때 필요한 전사파일은 mono-word 단위의 전사파일을 기본으로 사용하고, HLED 명령어를 사용하여 HMM 리스트 파일과 전사파일을 생성한다. 다음의 (그림 4)는 mono-word 단위의 HMM 모델을 triword 단위의 HMM 모델로 변경하기 위한 입출력을 나타낸 것이다.



(그림 4) triword 단위 HMM 모델 변경을 위한 입출력

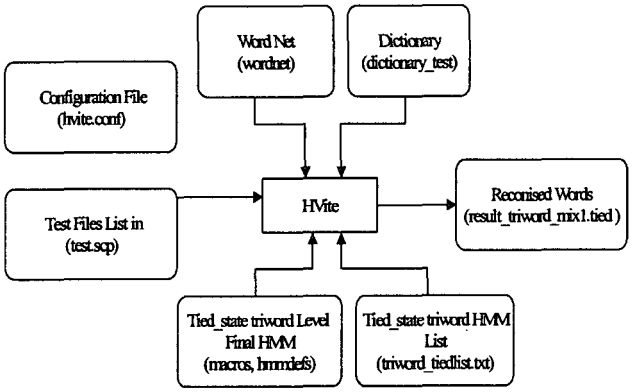
다음의 (그림 5)는 triword 모델 훈련을 위한 입출력을 나타낸 것이다.



(그림 5) triword 모델 훈련을 위한 입출력

6) tired-state biword/triword 모델의 훈련 및 인식

여기에서는 인식기의 성능을 향상시키기 위하여 biword 모델 또는 triword 모델의 훈련 결과를 사용하여 상태를 타잉하고, 훈련을 한 후, 인식 결과를 출력한다. 다음의 (그림 6)은 triword 단위의 인식을 위한 입출력을 나타낸 것이다.



(그림 6) triword 단위 인식을 위한 입출력

7) mixture 수 변경에 따른 biword/triword 모델의 훈련 및 인식

이전까지는 훈련시에 동일한 상태에서 하나의 Gaussian 을 적용한 경우이고, 이제부터는 인식율을 높이기 위하여 Gaussian 을 3 개, 5 개, 7 개를 적용하여 훈련하고 인식된 결과를 출력한다.

3. 숫자음 인식 실험 결과 및 분석

3.1 실험결과

left context biword 모델, triword 모델 각각에 대해 각 모델의 상태수를 3,5,7,9,11 개, 각 상태당 mixture 수를 1,3,5,7 로 변화시켜 인식실험을 실시한 결과, 다음의 <표 1>에서와 같은 인식결과를 나타내었다.

<표 1> biword, triword 모델 인식 결과

모델 / 상태수		Biword 모델		Triword 모델	
		Mixture 5	Mixture 7	Mixture 5	Mixture 7
상태 3	SENT	85.05	86.85	90.74	91.02
	WORD	95.85	96.39	97.40	97.51
상태 5	SENT	86.71	88.33	93.60	93.95
	WORD	96.31	96.81	98.24	98.33
상태 7	SENT	86.78	88.48	94.64	94.81
	WORD	96.36	96.86	98.54	98.57
상태 9	SENT	87.48	88.21	95.08	94.94
	WORD	96.56	96.78	98.68	98.65
상태 11	SENT	88.75	89.21	95.62	95.56
	WORD	96.90	97.05	98.80	98.79

<표 1>에서와 같이 biword 모델에서는 상태 7, 상태 11 의 mixture 수가 7 개 일 때 높은 인식율을 보이고 있고, triword 모델에서는 상태 9, 상태 11 의 mixture 수 5 에서 높은 인식율을 보이고 있다. triword 모델의 인식 결과와 biword 모델의 인식 결과를 비교해 볼 때, triword 모델이 문장은 6.41%, 단어는 1.75%이상 높은 인식율을 보이고 있다. 또한, triphone 의 실험결과[3]와 비교할 때, 문장은 6.73%, 단어는 1.84% 정도 높은 인식율을 보이고 있다.

3.2. 실험 결과 분석

3.1 절의 실험결과를 토대로 HMM 의 상태수는 9 개, 11 개, mixture 수는 5 개, 7 개로 한정하여 인식한 결과를 분석한 것이다. 분석은 각 결과에 대해 Confusion Matrix 를 작성하여 오인식이 빈번히 일어나는 숫자음을 구분하였다. 다음의 <표 2>는 biword 모델의 상태 수 11, mixture 수 5 개에 대한 Confusion Matrix 이다.

<표 2> biword 모델의 Confusion Matrix

Label / Recog.	영	일	이	삼	사	오	육	칠	팔	구	공
영	0	0	0	2	0	1	128	0	0	0	0
일	0	0	263	1	0	8	24	5	0	0	2
이	0	156	0	0	0	1	23	0	0	0	1
삼	2	0	0	0	5	1	2	0	2	0	2
사	4	1	0	71	0	10	0	0	13	6	1
오	2	0	4	1	0	0	28	0	0	124	26
육	27	11	8	0	1	14	0	0	1	6	0
칠	0	26	2	1	1	1	6	0	16	3	0
팔	0	2	0	10	41	1	0	2	0	1	3
구	0	0	0	1	0	50	4	0	0	0	45
공	0	2	0	0	0	9	6	0	0	6	0

<표 3>은 biword 모델에서 에러의 발생 빈도수가 5% 이상인 숫자음에 대한 것이다. 에러의 유형으로는 삽입, 삭제, 대체 에러로 나타나는데, 이들 중 삽입과 삭제 에러는 전체의 6-7% 정도 발생한다. 이에 비해 대체 에러는 93-94% 정도 발생한다. <표 3>에서 보는 바와 같이 60%이상이 몇 개의 숫자음에 집중적으로 발생함을 알 수 있다. Biword 모델에서 "이<->일"이 에러율이 가장 높은 것은 것으로 나타났다.

<표 3> biword 모델의 state 9, 11 에서의 에러율

Label/ mis-recognition	(occurance - percentage)				total percentage
	9state-5mixture	9state-7mixture	11state-5mixture	11state-7mixture	
삼/이	166 - 12.8%	154 - 12.8%	156 - 13.36%	151 - 13.53%	13.13%
이/일	215 - 16.6%	231 - 19.2%	203 - 17.39%	218 - 19.53%	18.13%
삼/사	76 - 5.86%	73 - 6.06%	71 - 6.08%	75 - 6.72%	6.17%
오/구	68 - 5.25%	60 - 4.98%	50 - 4.28%	47 - 4.21%	4.7%
육/영	168 - 12.97%	138 - 11.47%	128 - 10.96%	94 - 8.42%	11.04%
구/오	137 - 10.57%	134 - 11.13%	124 - 10.62%	112 - 10.03%	10.6%
total percentage	64.09%	65.66%	62.72%	62.45%	

다음의 <표 4>는 triword 모델의 상태 수 11, mixture 수 5 개에 대한 Confusion Matrix 이다.

<표 4> triword 모델의 Confusion Matrix

Label / Recog.	영	일	이	삼	사	오	육	칠	팔	구	공
영	0	0	1	0	0	0	42	0	0	0	0
일	0	0	110	0	0	0	8	4	0	1	1
이	0	56	0	0	0	0	7	0	0	0	0
삼	6	0	0	0	2	0	1	1	2	0	1
사	2	0	0	17	0	4	1	0	7	3	0
오	2	1	1	1	1	0	1	0	0	39	8
육	14	5	3	0	0	2	0	0	0	0	0
칠	0	19	0	0	0	0	1	0	5	0	1
팔	1	1	0	4	8	0	0	1	0	0	5
구	0	0	0	0	0	22	1	0	0	0	12
공	2	0	0	0	0	3	1	1	0	9	0

<표 5>는 triword 모델에서 에러의 발생 빈도수가 5% 이상인 숫자음에 대한 것이다. 에러의 유형도 biword 와 같고, 에러율도 biword 와 마찬가지로 60% 이상이 몇 개의 숫자음에 집중적으로 발생한다. triword 모델에서도 "이<->일"이 에러율이 가장 높고, biword 모델과 비교해서도 이 숫자음에 대해서는 에러율이 더 높은 것으로 나타났다.

<표 5> triword 모델의 state 9, 11 에서의 에러율

Label/ mis-recognition	(occurance - percentage)				total percentage
	9state-5mixture	9state-7mixture	11state-5mixture	11state-7mixture	
일/이	60 - 11.9%	61 - 11.8%	56 - 12.3%	55 - 12.11%	12.02%
이/일	131 - 25.9%	140 - 27.1%	110 - 24.3%	120 - 26.4%	25.92%
오/구	25 - 4.9%	26 - 5.0%	22 - 4.8%	28 - 6.1%	5.3%
육/영	52 - 10.3%	46 - 8.9%	42 - 9.2%	43 - 9.4%	9.45%
구/오	39 - 7.7%	43 - 8.3%	39 - 8.6%	35 - 7.7%	8.07%
total percentage	60.7%	61.1%	59.2%	61.71%	

4. 결 론

본 논문에서는 한국어 4 연숫자음에 대하여 biword, triword 모델로 구분하여 인식 실험을 하는 과정과 인식 결과 및 결과를 분석하였다. 인식결과는 triword 모델의 성능이 좋았으며, 오류의 유형에서는 "이<->일"이 가장 많은 오류율을 가짐을 알 수 있었다. 추후 이러한 실험 결과를 바탕으로 인식 단위 결정이나 숫자음 인식의 문제점을 파악하고, 성능을 개선하는데 사용될 수 있을 것으로 기대한다.

참고문헌

[1] 김춘석, "음성정보처리산업 현황", 음성인터넷 서비스기술 워크샵, pp.15-33, 2002.6.11.

[2] Steve Young 외, "The HTK Book", 2000. 7.

[3] 송화전, 김형순, "한국어 연결숫자인식을 위한 음향 모델링 방식 비교", 2001 년도 한국음향학회 학술발표 대회 논문집 제 20 권 제 1(s)호, pp.315-318, 2001.