한국어 연결숫자인식을 위한 숫자 모델링에 관한 연구

김기성*, 김승희**, 김형순**, 지민제***
*국방과학연구소, **부산대학교 전자공학과, ***(주)세종스피치텍

A Study on Digit Modeling for Korean Connected Digit Recognition

Gi Sung Kim*, Seunghi Kim**, Hyung Soon Kim**, Minje Zhi***
*Agency for Defense Development, **Dept. of Electronics Eng., Pusan National University,
***Sejong Speech Tech, Inc.
E-mail : speech@sunam.kreonet.re.kr*, {cygnus, kimhs}@hyowon.cc.pusan.ac.kr**

요 약

본 논문에서는 전화망에서의 연결 숫자 인식 시스템의 개발에 대한 내용을 다루며, 이 시스템에서 다양한 숫자 모델링 방법들을 구현하고 비교하였다. Word 모델의 경우 문맥독립 whole-word 모델과 문맥종속 whole-word 모델을 구현하였으며, sub-word 모델로는 triphone 모델과 불파음화 자음을 모음에 포함시킨 modified triphone 모델을 구현하였다. 그리고 tree-based clustering (TBC) 방법을 sub-word 모델과 문맥종속 whole-word 모델에 적용하였다.

이와 같은 숫자 모델들에 대해 연속 HMM을 이용하여 화자독립 연결숫자 인식 실험을 수행한 결과, 문맥종속 단어 모델이 문맥독립 단어 모델보다 우수한 성능을 나타냈으며, triphone 모델과 modified triphone 모델은 유사한 성능을 나타냈다. 특히 tree-based clustering 방법을 적용한 문맥종속 단어 모델이 4연 숫자열에 대해 99.8%의 단어 인식률 및 99.1%의 숫자열 인식률로서 가장 우수한 성능을 나타내었다.

1. 서 론

음성인식 기술은 음성합성 기술과 더불어 human-computer interface의 핵심기술로서 정보화의 진전과 더불어 그 필요성이 증대되고 있다. 그 한 분야로 음성입력에 의한 숫자인식은 실생활의 많은 부분에서 활용될 수 있다.

숫자인식은 고립숫자인식과 연결숫자인식의 두 부류로 나눌 수 있다. 고립숫자인식은 하나씩 구분되어 발음되는 숫자를 인식하는 기술이고, 연결숫자인식은 자연스럽게 연속적으로 발음되는 숫자열에 대해 그 개수와 각각의 숫자를 인식하는 기술을 의미한다. 연결숫자인식이 사용자의 입장에서는 발음상의 자연스러움이나 효율성을 위해서 바람직하나 다음과 같은 몇 가지 문제점을 가진다[1]. 첫째로, 연속적으로 발음되는 숫자열에서는 각 숫자들의 경계가 모호하다. 둘째로, 인접 숫자들 사이의 상호조음현상에 의해 각 숫자들의 고유발음이 변하게 된다. 셋째로, 하나의 숫자열에 몇 개의 숫자가 발음되었는지를 모르는 경우가 많다. 이러한 여러 가지 문제점들을 해결할 수 있다면 연결숫자인식은 숫자가 의미를 가지는 많은 응용분야, 예를 들면, 음성 다이얼링, 은행업무 자동화, 신용카드 번호입력등에 효과적으로 활용될 수 있다[2].

본 논문에서는 HMM을 이용하여 연결숫자인식 시스템을 구성하였으며, 그 중에서도 연속 HMM에 의한 인식 방법을 채택하였다. HMM에 기반을 둔 연결숫자인식에서는 각각의 숫자에 대한 기본 모델링 단위를 정의해야 하며, 본 논문에서는 문맥독립 whole-word 숫자 모델, 문맥종속 whole-word 숫자모델[3] 및 triphone 모델[4]을 인식 실험에 사용하였으며, trphone 모델에서 불파

음화 자음을 모음에 포함시킨 modified triphone 모델을 검토하였다. 그리고 sub-word 모델 및 문맥종속 whole-word 숫자 모델에 대해서는 tree-based clustering[5][6]을 적용함으로써 인식률을 향상시킬 수 있었다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 절에서 본 연결숫자인식 시스템의 구조 및 인식 network 에 대해 설명하고, 3 절에서 숫자 모델링 방법에 관하여 기술한다. 4 절에서는 실험에 사용된 데이터 베이스 및 실험 결과에 대해서 언급하고, 마지막으로 5 절에서 결론을 맺는다.

2. 연결숫자인식 시스템

본 논문에서는 연속 HMM을 사용한 연결숫자인식 시스템을 그림 1 과 같이 구성하였다. 먼저 훈련용 데이터로부터 연결숫자인식에 사용될 숫자모델과 배경잡음을 표현하는 silence 모델을 구성한다. 인식 단계에서는 입력음성으로부터 특징 파라메터를 추출하고, 그 특징 파라메터를 HMM network 상에서 숫자 모델들과 silence 모델과 비교하여 연결 숫자를 인식한다. 일반적으로 한 숫자열 내에는 임의의 개수의 숫자가 포함될 수 있으므로 null grammar 형태가 가능하지만, 인식 대상 연결 숫자의 길이를 알고 있다면 HMM network 에서 제한을 가하는 것이 바람직하다. 본 논문에서는 인식 대상 연결 숫자의 길이를 연결 숫자 4 자리 또는 연결 숫자 3 자리의 끝에 '에'를 붙인 것, 예를 들어 '사이칠구', '오일육에' 등을 인식대상으로 하는 network 을 구성하였다.

3. 숫자 모델링 방법

HMM 에 기반을 둔 연결숫자인식에서는 각각의 숫자에 대한 기본 모델링 단위를 정의해야 하며 whole-word(단어)나 sub-word unit(음소, 반음절 등)을 기본 모델링 단위로 사용할 수 있다.

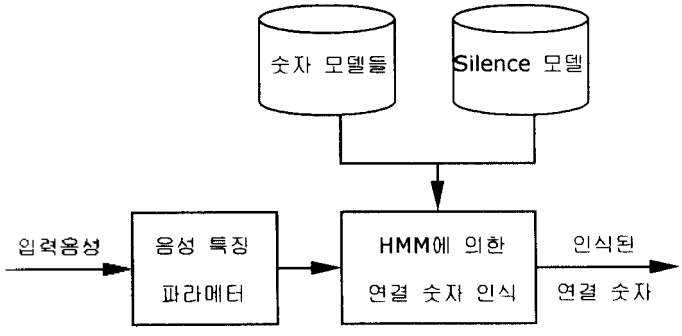

그림 1. 연결숫자인식 시스템

3.1 Whole-word 숫자 모델

단어를 기본 모델링 단위로 하는 것에는 문맥독립 whole-word 숫자 모델과 문맥종속 whole-word 숫자 모델이 있다. 문맥독립 whole-word 숫자 모델은 문맥을 고려하지 않고 각각의 숫자를 모델링한 것인데, 본 논문의 연결숫자인식 실험에서는 '영', '공', '일', '이', '삼', '사', '오', '육', '칠', '팔', '구' 등의 11 개 숫자와 '에'라는 조사, 그리고 silence 모델을 합하여 총 13 개의 모델을 사용하였다.

문맥종속 whole-word 숫자 모델은 문맥을 고려하여 앞뒤에 오는 단어, 즉 앞 뒤에 오는 숫자에 따라 그 숫자를 다르게 모델링한 것인데 앞 뒤에 silence 와 11 개의 숫자, 그리고 '에'라는 단어가 올 수 있다면 각 숫자는 169(=13×13)개의 문맥종속 whole-word 숫자 모델을 가질 수 있다[3].

3.2 Sub-word 숫자 모델

Sub-word unit 중에서 숫자에서 발생할 수 있는 음소를 기준으로 triphone 을 인식단위로 사용한다[4].

본 논문에서는 triphone 모델 중에서 불파음화 자음을 모음에 포함시킨 modified triphone 모델도 연결숫자인식 실험에 사용하였다. 즉, '육'이라는 숫자에 'ㅜ'와 'ㄱ'의 두가지 문맥독립형 음소가 있는데 그 중 'ㄱ'이 불파음화 자음이 되기 때문에 'ㄱ'을 'ㅜ'에 포함시켜 '육'을 하나의 음소로 간주하고 모델링하였다. 그리고, 이들 sub-word 모델에 대해서는 tree-based clustering 을 적용하였다.

3.3 Tree-based Clustering(TBC)[5][6]

만약 훈련용 데이터가 무제한적으로 충분하다면, 문맥종속적 문제는 가능한 모든 문맥에 대해 별도의 모델을 구성함으로써 해결할 수 있다. 그러나, 실제는 훈련용 데이터는 제한될 수 밖에 없으며, 만약 훈련용 데이터가 무제한적으로 충분하다고 하더라도 모든 다른 모델에 대한 저장을 위한 공간이 문제가 된다. 따라서, 제한적 데이터와 저장 공간에 대한 문제로 인해 문맥들을 결합하여 비슷한 특성을 나타내는 모델들을 하나의

class 로 나타내고 각 class 에 대해 하나의 모델을 만드는 것이 필요하다. 이러한 class 를 구성하는 효과적인 방법으로 binary decision tree 를 이용하는 방법이 있다.

Binary decision tree 는 먼저 음성의 기본단위(단어, 음소 등)에 해당하는 특정 단위, 예를 들어 단어인 경우는 /일/, 음소인 경우는 /ㅗ/에 해당하는 데이터들을 모은 후 이를 두 개의 부분집합으로 나누고, 그 각각의 부분집합을 다시 두 개의 부분집합으로 나누어 가는 일련의 과정을 통해 구성된다. 그림 2 에 3 개의 상태를 가지는 음소 /ㅗ/에 대한 모델들에 대해 중앙에 위치한 상태를 tree-based clustering 에 의해 5 개의 class 로 나타내는 예를 나타내었다.

4. 데이터 베이스 및 실험결과

본 논문의 연결숫자인식 실험에서는 연속 HMM 을 이용하여 앞 장에서 설명한 숫자 모델링 방법들을 비교하였다. 실험에 사용한 모든 음성 데이터 베이스는 8kHz 로 sampling 되었으며, 음성 특징 파라메터로는 12 차 MFCC, log 에너지, 그리고 각각의 delta 파라메터를 사용하여 총 26 차의 파라메터를 사용하였다.

4.1 음성 데이터 베이스

숫자 모델을 위한 훈련용 데이터 베이스로는 지역적 균형을 맞추기 위하여 4 개의 방언군 (1.서울, 경기 2. 충청 3.전라 4.경상) 에서 각각 남성 21 명씩 총 84 명이 자연스럽게 발음한 연결 숫자 음성을 사용하였으며, 각각의 숫자열은 국번(3 자리 숫자) + '에', 4 자리 숫자 등의 연결 숫자로 이루어져 있다. 그리고 이들 숫자열들은 서로 환경이 다른 2 개의 유선전화 채널 및 1 개의 무선전화 채널을 통해 3 번씩 발음되었다.

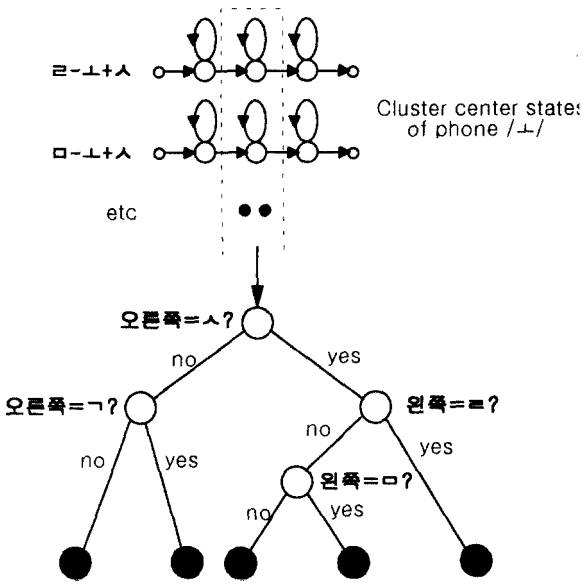

그림 2. 음소 /ㅗ/에 대한 decision tree 기반의 상태 tying

인식용 데이터 베이스로는 4 개의 방언군에서 각각 훈련에 참가하지 않은 남성 4 명씩 총 16 명이 자연스럽게 발음한 연결 숫자 음성을 사용하였다.

4.2 Whole-word 숫자 모델

문맥독립 whole-word 숫자 모델과 문맥종속 whole-word 숫자 모델에 대한 상태수를 6 개에서 14 개까지, 각 상태에서의 Gaussian mixture 의 수를 1 개에서 11 개까지 변화시키면서 인식 실험을 수행하였다. 인식 결과를 살펴보면 두 경우 다 mixture 수의 변화에 상관없이 상태의 개수가 10 개인 경우에 인식률이 가장 높았다. 이는 상태수가 10 개인 경우에 비해 상태수가 6 개인 경우는 숫자를 잘 모델링하기에 상태수가 부족하며, 상태수가 14 개인 경우는 훈련용 데이터로 사용한 음성 데이터의 양이 상태수 14 개를 모델링하기에는 부족하기 때문으로 판단된다. 그리고 각 상태당 Gaussian mixture 의 수를 1 개에서 11 개로 증가시킬수록 인식률도 증가하는 경향을 보이는데 이는 mixture 수가 많아질수록 숫자를 잘 모델링하는 것으로 판단된다. 그러나 mixture 수가 어느 이상 많아지면 인식률의 증가가 미미해지면서 인식률이 수렴하는 경향을 보인다. 문맥종속 whole-word 모델의 상태수 10 개에 대한 실험결과는 표 1 에 나타내었다. 표에서 Corr.(Percent Correction)는 전체 인식 대상 단어 중에서 삭제 오류와 대체 오류를 뺀, 즉 정확하게 인식한 단어에 대한 인식률을 나타내며, Acc.(Percent Accuracy)는 Corr.에서 삽입 오류를 고려한 인식률이다[7].

문맥독립 모델과 문맥종속 모델에 대한 인식결과를 비교해 보면, 앞 뒤 숫자를 고려하지 않은 문맥독립 whole-word 숫자 모델에 비해 문맥종속 whole-word 숫자 모델의 성능이 우수함을 알 수 있다. 이는 연속적으로 발음되는 숫자들은 앞 뒤 숫자에 의해 영향을 받아 그 특성이 변화하며, 앞 뒤 숫자에 따라 모델을 다르게 두는 것이 그 특성을 나타내는데 적합함을 나타낸다. 그림 3 에 두 모델의 인식 결과를 비교하기 위해 상태

수가 10 개인 경우 숫자열에 대한 인식률을 나타내었다. 그림에서 CI 는 문맥독립(Context-Independent)을, CD 는 문맥종속(Context-Dependent)을 의미한다.

표 1. 문맥종속 whole-word 숫자 모델의 인식률 (%)

상태수	Mixture 개수	숫자열 인식률	단어 인식률	
			Corr.	Acc.
10	1	88.1	96.9	96.7
	3	92.9	98.2	98.1
	5	94.3	98.5	98.5
	7	95.1	98.8	98.7
	9	95.0	98.7	98.7
	11	95.5	98.8	98.8

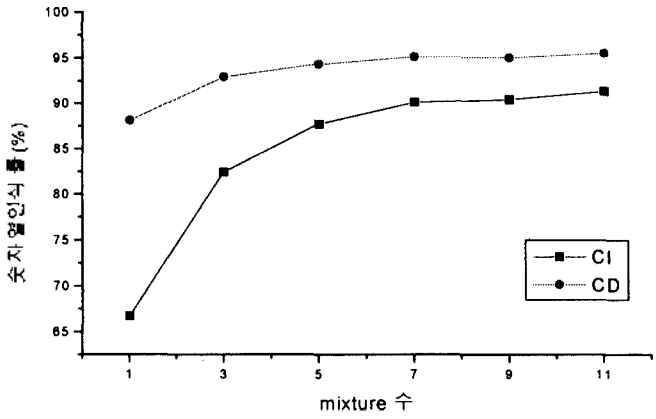

그림 3. Whole-word digit 모델의 숫자열 인식률
(상태수 10 개인 경우)

4.3 Sub-word 숫자 모델

Sub-word 숫자 모델에서는 먼저 triphone 모델에 대해 각 모델이 상태수가 3 개 및 5 개인 경우를 비교하였다. 그리고 각 상태에서의 Gaussian mixture 의 수를 변화시키면서 인식 실험을 수행하였으며, tree-based clustering 을 적용하였다. 실험결과 triphone 모델에서는 숫자열과 단어 모두에서 상태수가 5 개인 경우가 상태수 3 개인 경우보다 인식률이 높게 나타났으며, 이 결과를 근거로 불파음화 자음을 모음에 포함시키는 modified trihpne 모델에서는 상태수가 5 개인 경우에만 mixture 수를 변화시키면서 연결숫자인식 실험에 사용하였다. 표 2 과 3 에 상태수 5 개에 대한 결과를 나타내었다.

표에 나타난 인식결과를 살펴보면 mixture 개수가 적은 경우에는 triphone 모델의 인식률이 modified triphone 모델의 인식률에 비해 높게 나오지만, mixture 개수가 늘어날수록 modified triphone 모델의 인식률이 더 높음을 알 수 있다. 이는 불파음화 자음을 모음에 포함시켜 함께 모델링하는 것이 불파음화 자음을 따로 모델링하는 것보다 mixture 개수가 늘어날수록 모델의 robustness 측면에서 유리하다는 것을 나타낸다. 그러나, 전반적으로 볼 때 modified triphone 모델을 도입함으로써 얻어지는 성능향상은 기대수준에 미치지 못하는 것으로 보인다.

표 2. Triphone 모델의 인식률 (%)

상태수	Mixture 개수	숫자열 인식률	단어 인식률	
			Corr.	Acc.
5	1	93.4	98.1	98.0
	3	96.2	99.0	98.9
	5	96.9	99.2	99.1
	7	97.7	99.3	99.4
	9	97.5	99.4	99.3
	11	97.8	99.4	99.4

표 3. Modified triphone 모델의 인식률 (%)

상태수	Mixture 개수	숫자열 인식률	단어 인식률	
			Corr.	Acc.
5	1	93.4	98.8	98.1
	3	95.6	99.2	98.8
	5	96.7	99.4	99.1
	7	97.8	99.6	99.4
	9	98.3	99.7	99.5
	11	98.3	99.7	99.5

4.4 Tree-based clustering(TBC)을 이용한 whole-word 숫자 모델

4.2 절에서 인식 실험에 사용한 문맥종속 whole-word 숫자 모델중에서 인식률이 가장 높았던 상태수 10 개에 대해서만 tree-based clustering 을 적용하였으며 mixture 수를 1 개에서 11 개까지 증가시키면서 연결숫자인식 실험을 하였다. 실험에 대한 결과는 표 4 에 나타내었다.

TBC 를 이용한 whole-word 숫자 모델의 인식결과를 TBC 를 사용하지 않은 whole-word 숫자 모델과 sub-

word 숫자 모델의 인식결과를 비교하기 위해 그림 4에 그 결과를 함께 나타내었다. 그림 4를 살펴보면 TBC를 이용한 whole-word 숫자 모델이 가장 우수한 성능을 나타냄을 알 수 있다. 특히 tree-based clustering을 사용하지 않은 whole-word 숫자 모델과는 인식률에서 차이가 많이 나는데, 이는 tree-based clustering을 통하여 비슷한 특성을 나타내는 상태들이 tying 되면서 동일한 훈련용 데이터로부터 좀 더 개선된 모델을 얻을 수 있었던 것에 기인한다고 판단된다.

표 4. TBC를 이용한 whole-word 숫자 모델의 인식률(%)

상태수	Mixture 개수	숫자열 인식률	단어 인식률	
			Corr.	Acc.
10	1	97.9	99.2	99.1
	3	97.6	99.4	99.4
	5	98.3	99.6	99.5
	7	**99.1**	**99.8**	**99.8**
	9	98.8	99.7	99.7
	11	99.0	99.8	99.7

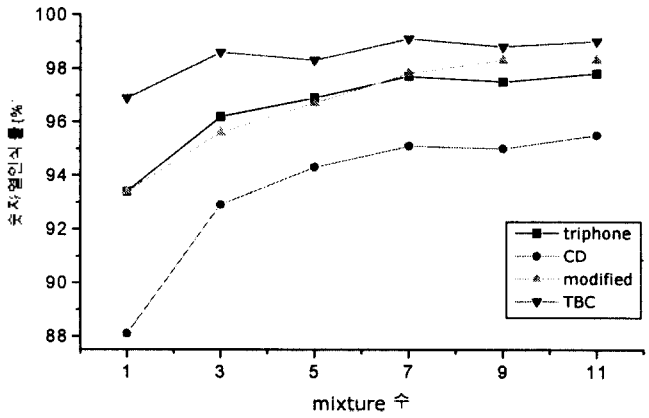

그림 4. 숫자모델들의 숫자열 인식률 비교

5. 결 론

본 논문에서는 연결숫자인식 시스템에서의 숫자 모델링 방법들에 대해 연속확률분포 HMM을 이용하여 성능을 비교하였다. 본 논문에서 검토한 여러 가지 숫자 모델링 방법들 중에서 whole-word 숫자 모델에 tree-based clustering을 적용한 모델링 방법이 가장 우수한 성능을 나타냈다. 특히 tree-based clustering을 사용하지 않은 whole-word 숫자 모델과는 인식률에서 차이가 많이 나는데, 이는 tree-based clustering을 통하여 비슷한 특성을 나타내는 상태들이 tying 되면서 동일한 훈련용 데이터로부터 보다 개선된 모델을 얻을 수 있었던 것 때문으로 판단된다.

본 논문에서는 연결숫자인식을 위한 숫자 모델링 방법에 초점을 맞추었으며, 전화망을 통과하는 과정에서의 채널왜곡에 대해서는 별도의 고려를 하지 않았다. 전화망을 통한 연결숫자인식의 성능을 보다 향상시키기 위한 효과적인 채널왜곡 보상 방법에 대한 연구가 현재 진행중이다.

참 고 문 헌

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall International, Inc., 1993.

[2] Yu-Hung Kao and Lorin Netsch, "Inter-digit HMM connected digit recognition using the macrophone corpus," in Proc. IEEE ICASSP, pp.1739-1742, 1997.

[3] S. K. Gupta, Frank Soong and Raziel Haimi-Cohen, "High-accuracy conneted digit recognition for mobile applications," in Proc. IEEE ICASSP, pp.57-60, 1996.

[4] Y. Normandin, R. Cardin and R. D. Mori, "High-performance connected digit recognition using maximum mutual information estimation," IEEE Trans. on Speech and Audio Processing, vol.2, no.2, pp.299-311, 1994.

[5] L. R. Bahl, P. V. de Souza, *et al.*, "Decision trees for phonological rules in continuous speech," in Proc. ICASSP, pp.185-188, 1991.

[6] H. J. Nock, M. J. F. Gales, S. J. Young, "A comparative study of methods for phonetic decision-tree state clustering," Proc. EUROSPEECH, vol.1, pp.111-114, 1997.

[7] S. Young, *HTK: Hidden Markov Model toolkit V2.0*, Eng. Dept., Speech Group, Cambridge, Univ., Cambridge UK, Tech, Rep.,1992.