연속분포 HMM을 이용한 한국어 연속 음성 인식 시스템 개발

김 도 영 박 용 규 권 오 욱 은 종 관

한국과학기술원 전기 및 전자공학과

On the Developement of a Continuous Speech Recognition System using Continuous Hidden Markov Model for Korean Language

Do Yeong Kim Yong Kyu Park Oh Wook Kwon Chong Kwan Un

Department of Electrical Engineering, KAIST

요 약

본 논문에서는 연속분포 hidden Markov 모델을 이용한 화자독립 연속 음성 인식 시스템에 관해 기술한다. 연속분포 모델은 평균과 분산 벡터로 구성되며 음성신호를 직접 모델링하여 양자화 왜곡이 없어진다. 특징벡터는 filter bank 계수 및 그 1, 2차 미분계수를 사용하여 음성신호의 동적 특성을 반영하였다. Segmental K-means 알고리즘을 이용하여 학습하였으며, 연속어 인식에서 가장 문제가 되는 조음화 현상으로 인한 인식률 저하를 막기 위해 앞뒤의 음소를 고려해 주는 triphone을 인식단위로 사용하였다. Search 알고리즘으로는 시간 면에서 효율이 좋은 one-pass search 알고리즘을 사용하였다. 성능 평가를 위한 화자 독립 인식 실험에서 문법이 없을 경우 83%, finite state network을 적용한 경우에는 94%의 인식률을 나타내었다.

I. 서 론

사회가 고도로 정보화되면서 인간과 인간 혹은 인간과 기계간의 자유로운 통신에 대한 관심이 고조되고 있다. 음성은 인간의 가장 자연스러운 의사표현수단으로 만일 컴퓨터 등의 기계에 음성을 통해 명령을 내리거나, 서로 다른 언어를 사용하는 사람들이 컴퓨터를 매개로 자유롭게 대화할 수 있다면 매우 획기적인 일이라 아니 할 수 없다. 음성인식은 기계번역, 합성 등과 함께 음성을 이용한 정보교환에 핵심적인 기술로 1950년대 초부터 연구되기 시작하였다. 초기의 음성인식 시스템은 대개 고립단어 인식을 그 목표로 하였으나 최근에는 미, 일, 독 3개국간의 자동통역 전화가 발표되는 등 화자독립 연속어 인식시스템의 개발에 연구의 촛점이 맞추어져 있다.

임의의 화자가 발음하는 임의의 음성을 인식하는 시스템은 아직까지 존재하지 않으

며 따라서 모든 음성인식시스템은 어느 정도의 제약조건을 가지게 되는데 일반적으로 임의의 화자가 발음한 자연스러운 연속어이면서 문법의 복잡도가 높은 음성을 인식하는 시스템일수록 높은 수준의 기술을 요구하게 된다. 특히 연속어의 경우는 단어내의 조음결합뿐만 아니라, 문장 내에 있는 각 단어의 경계에서 발생하는 조음현상으로 인하여 각 단어의 경계가 불명확해지고 조음현상의 정도가 심각해 지게 되어, 더욱 정확한 모델링이 필요하게 된다.

개발된 시스템은 제한된 100 개의 어휘로 구성된 시간구 연속어를 인식하도록 설계되어 있다. 인식의 기본이 되는 모델로는 hidden Markov model(HMM)을 사용하였으며, 출력확률 분포는 6개의 mixture를 갖는 Gaussian mixture density로 구성된다.

본 논문의 구성은 다음과 같다. 2장에서는 연속음성 인식 시스템의 개요 및 각 부분의 특성을 살펴보고, 3장에서는 인식에 사용된 데이타 베이스와 실험결과를 고찰한다. 또한 4장에서는 결론 및 향후 연구방향에 대해 정리한다.

II. 연속 음성 인식 시스템

연속음성 인식 시스템의 구조는 그림1과 같으며, 음성 신호의 특징 추출부분, 단어단위 정합(match) 부분, 문장단위 정합부분의 3부분으로 나뉜다. 단어단위 정합부는 입력 특징 벡터열과 시스템의 단어 모델 사이의 유사도를 측정하여 어떤 단어가 가장 유사한가를 결정하는 역할을 하며, 이때 단어 모델은 lexicon에 따른 subword model의 결합으로 얻어진다. 문장단위의 정합부는 언어 모델을 이용하여 문법에 맞는 최대 확률을 내는 단어열을 인식결과로 하는 부분이다. 그림 1에 시스템의 전체구조를 나타내었다.

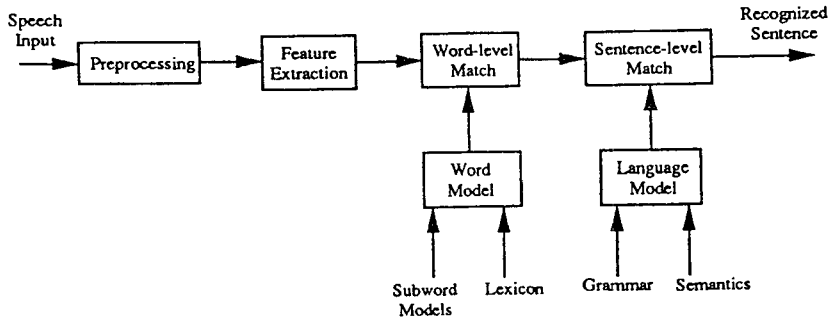


그림 1. 연속 음성 인식 시스템의 구조.

Fig. 1. Architecture of the continuous speech recognition system.

A. 음성 신호 전처리 및 특징 추출

음성신호는 16kHz, 16bit로 sampling된다. 0.975의 factor로 pre-emphasis하며, 20ms(320sample)씩 Hamming windowing하여 처리된다. 10ms씩 shift되며, 512 point FFT를 통해 16개의 sub-band별로 energy를 구한다. 또한 전체 energy를 구하여 전체 17차의 특징 벡터로 음성신호를 나타내게 된다.

음성신호의 동적 특성을 반영하기 위하여 추출된 특징벡터의 1차 및 2차 미분 계수를 구하여 사용한다. 따라서 사용되는 음성 신호의 특징벡터는 51차가 된다.

B. 음성신호의 모델링 : Hidden Markov Model

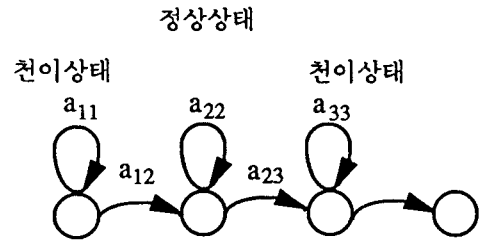
음성신호를 모델링하기 위해 사용한 방식은 continuous mixture density left-to-right HMM이다. HMM은 Baum에 의해 처음 그 이론이 소개된 이래 음성 인식 분야에 적용되어 좋은 결과를 나타내고 있으며, Carnegie Mellon 대학의 SPHINX[1], AT&T의 음성인식 시스템[2][3], BBN의 BYBLOS[7], IBM의 TANGORA[6]등의 시스템에서 널리 사용되고 있는 방법으로 관측 불가능한 상태열과 관측가능한 관측열로 구성되는 2중의 stochastic process 이다.

HMM의 파라미터는 초기상태 확률, 상태천이 확률, 출력확률로 구성되며 학습이란 이 확률 값들을 반복적으로 재추정하여 주어진 모델에서 높은 확률 값을 내도록 하는 것을 말한다. 본 시스템에서는 sub-word(triphone)단위가 기본이 되며, 각 단위는 천이, 안정, 천이의 3가지 상태로 구성된다. 따라서 단어 모델은 그림 2와 같이 이러한 sub-word 모델의 결합으로 나타낼 수 있다.

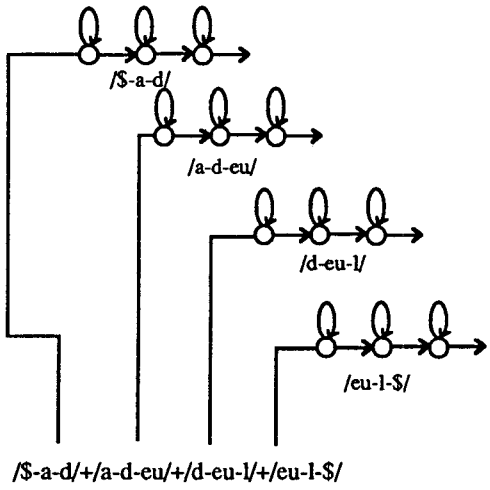
C. 인식단위 : Triphone

좋은 인식 단위는 trainability와 consistency를 모두 만족하여야 한다[1]. trainability란 학습에 충분할 정도로 빈번히 발생되어야 한다는 것이며, consistency는 일관된 독특한 성질을 가져야 함을 말한다.

초기의 시스템에서 사용된 인식 단위인 단어는 consistency면에서는 장점을 가지는 반면, 단어수가 증가함에 따라서 trainability에는 문제점이 발생하게 된다. 또한 인식 단위를 음소로 사용하게 되면 부족한 학습 데이타에서도 충분히 학습시킬 수 있을 정도로 trainability는 좋으나, 이웃하는 음소등의 영향에 민감하게 반응하여 일관성이 없어진다. Triphone은 이러한 문제를 해결하기 위한 절충형이라고 할 수 있으며, 좌우에 이웃하는 음소를 고려해 주는 방식이다. 즉, triphone은 조음화 현상으로 인한 음소의 특성 변화를 잘 모델링 해 줄수 있다. 그러나, triphone의 경우도 일부 단위는 그 발생빈도가 너무 작어서 제대로 학습이 되지 않을 경우가 많다. 이론적으로 생각해 보면 가능한 triphone의 갯수는 음소 갯수의 3승에 달하게 되므로 한국어의 경우 음소의 수를 30개 정도라 가정하면 약 30^3개의 단위가 존재하게 되는데 각각의 단위가 발생빈도가 서로 다르고, 거의 발생하지 않은 경우도 많이 있다. 이러한 문제의 해결을 위해 triphone의 갯수를 줄이는 방법들을 사용하는데 널리 알려진 방법으로는 발생빈도가 일정치 미만인 triphone들을 묶어서 한꺼번에



(a) Triphone 단위의 HMM.
(a) Triphone level HMM.



(b) 단어 단위의 HMM (아들).
(b) Word level HMM.

그림 2 HMM 의 구조.
Fig. 2. Structure of HMM.

학습시키는 unit reduction rule과 정보이론을 이용한 merging 방식이 있다. 본 연구에서는 우선 인식단위로 적당한 32개의 phone like unit(PLU)를 정의하고 이에 근거한 triphone을 사용하였다[4]. 전체 triphone의 갯수는 242개이며, 그중 학습 데이타 베이스에서 발생빈도가 10회 미만인 것만 unit reduction rule을 이용하여 합쳐주어 결과적으로 사용된 subword 모델의 갯수는 208개이다.

D. 학습 알고리즘 : Segmental K-means 알고리즘

Sub-word 단위의 모델 학습을 위하여 초기에는 음성신호를 hand-labeling하는 경우가 많았으나, 최근에 와서는 초기에 균일하게 sub-word단위로 분할한 다음 반복되는 학습의 결과에 의해 주어지는 parameter들을 이용하여 분할정보 역시 바꾸는 bootstrapping방식이 널리 사용되고 있다. 분할에는 Viterbi 알고리즘이 사용되며 대용량 연속어 인식을 위한 시스템에서는 학습 데이타가 방대하여 hand-labeling이 불가능한 만큼 이러한 방식은 필수적이라 하겠다.

본 시스템에서는 HMM 학습 알고리즘으로 segmental K-means알고리즘을 사용하였으며 다음과 같은 과정을 통해 수행된다[2][3][8].

(1) 초기화 - 모든 학습 data(음성 신호)를 인식단위와 상태별로 균일하게 나눈다.
(2) 군집화 - 전체 data에 대해 같은 인식단위, 같은 상태 s로 분할된 것들을 모아 M 개로 clustering 한다.
(3) 추정 - 상태 s의 모든 cluster에 대해 평균, 분산, mixture gain을 구한다.
(4) 분할 - 바뀐 paramter들을 이용하여 Viterbi 알고리즘을 통해 학습 data를 다시 분할한다.
(5) 반복 - 분할의 결과가 이전 단계와 같으면 중단하고 그렇지 않으면 (2)-(4)의 과정을 반복한다.

HMM 학습의 대표격이라 할 수 있는 Baum-Welch 알고리즘이 모든 경로에 대한 likelihood를 이용하는 반면 segmental K-means 알고리즘은 최적의 분할 경로만을 이용하는 방식으로 그 수렴이 수학적으로 증명되어 있다[8].

E. Search 알고리즘 : One-pass Search 알고리즘

연속 음성 인식에서의 search 알고리즘은 여러 단어가 연속적으로 발음된 입력신호와 가장 유사한 최적의 단어열을 찾는 것을 의미한다. Search 알고리즘은 계산량, 계산방식, search 공간 등에 따라 다양한 방식들이 있는데, 본 시스템에서 채택한 방식은 one-pass(OP) search 알고리즘이다[10].

OP 알고리즘의 특징으로는 두번째, 세번째 최적후보열을 찾기 어려운 대신 계산량이 작고 frame-synchronous 방식이어서 실시간 처리가 용이하며, loop-transition을 갖는 syntax의 적용이 가능하다는 것을 들 수 있다. 또한 beam-search 기법을 효율적으로 결합시

킬 수 있으므로 계산속도면에서 더욱 유리하다.

F. 문법 : Finite State Network

학습 및 인식에 사용된 database는 시간구로서 <그림3>과 같은 유한한 상태의 network으로 나타낼 수 있다.

이러한 문법의 결합은 연속어 인식에서는 필수적이라 할 수 있는데, 인식 대상의 문법적 어려움의 척도로는 perplexity 라는 정의를 사용한다[11]. Perplexity는 정성적으로 임의의 단어뒤에 올 수 있는 평균적인 후보 단어의 수라고 할 수 있으며, 문법이 적용되지 않을 경우는 당연히 전체 단어의 갯수가 perplexity가 되며 (본 시스템의 경우는 전체 단어의 갯수가 100개 이므로 perplexity도 100이 된다) 인식률도 저하된다.

본 시스템에 적용된 finite state network을 <그림 3> 에 나타내었으며 perplexity는 약 20 정도가 된다.

III. 인식 실험 및 결과

A. 데이타베이스

인식대상이 되는 문장은 날짜, 요일, 시간 등으로 구성된 시간구이다. 각 문장은 그림4의 FSN에 근거하여 작성되었는데, 예를 들면 "12월 31일 오후 12시 55분" 또는 "토요일 오전 9시 30분"등과 같은 것이 있다.

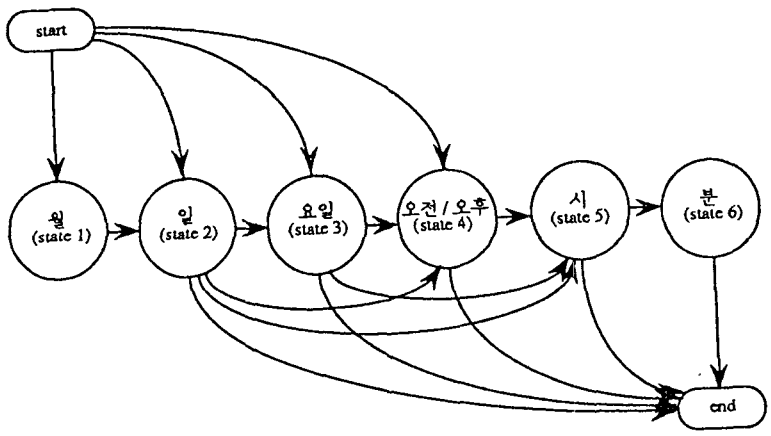


그림 3. Finite State Network을 이용한 문법.
Fig. 3. Grammar using Finite State Network.

모두 100개의 어휘로 구성되어 있으며, 서로 다른 20대 남, 녀 화자 83명으로 부터 다양한 문장을 녹음하도록 하여 학습 및 인식실험에 이용하였다. 학습에는 남, 녀 각각 30명씩의 데이타를 임의로 선정하여 사용하였으며, 인식실험에는 학습에 참가하지 않은 5명의 남성화자와 4명의 여성화자가 발음한 221개의 연속어 문장을 이용하였다. 조용한 사무실 환경에서 녹음하였으며 16bit, 16kHz로 A/D 하였다.

B. 실험 결과

연속어의 인식률은 흔히 단어단위의 인식률로 나타내게 되고 아래식과 같이 나타낼 수 있다.

$$\text{인식률} = \frac{\text{총단어수} - (\text{치환단어} + \text{삭제단어} + \text{첨가단어})}{\text{총단어수}} * 100(\%)$$

또한, 치환과 첨가가 동시에 일어나는 경우가 많이 있으므로, 첨가단어는 인식률 계산에서 제외시키기도 한다.

<표 1>에는 문법이 없을 경우와 finite state network(FSN)을 적용하였을 경우에 대한 실험 결과를 나타내었다. FSN의 문법을 적용하였을 경우 인식률이 매우 향상됨을 알 수 있다.

표 1. 연속음성인식 실험 결과

Table 1. Experimental results of continuous speech recognition

문 법	전체단어	치환단어	첨가단어	삭제단어	인식률(%)	
					첨가단어 포함	첨가단어 제외
No Grammar	772	67	55	0	83.1%	90.7%
FSN	772	40	2	2	93.6%	94.2%

문법이 적용되지 않았을 경우의 오인식 결과를 분석해 보면 연속어의 고유한 특성이라고 할 수 있는 조음화현상이 매우 심각한 문제로 작용함을 알 수 있다. 예를 들어, "일요일 십일시 오십분" 이라는 문장이 "일요일 십일시 오후 십분"으로 오인식되었는데, "오후 십분"이라는 단어열을 조금 빠른 속도로 발음해 보면 "오십분"과 매우 유사하게 들림을 알 수 있다. 이때 문법을 적용하면 시간을 나타내는 "십일시" 다음에 "오후"라는 단어가 올 수

없으므로 finite state network을 적용하였을 경우 제대로 인식하게 된다.

실험결과를 좀더 자세히 살펴보기 위하여 <표 2>와 <표 3>에는 문법이 있을 경우와 FSN을 적용한 경우에 대하여 각 화자별 인식실험 결과를 나타내었다. 일반적으로 남성화자에 비해 여성화자의 오인식률이 큰데 그 이유로는 여성화자의 경우 피치(pitch)와 제 1포만트의 구별이 여려운 점을 들 수 있다. 또한, 문법이 적용되지 않았을 경우에는 첨가에 의한 오인식이 많이 발생하며, FSN을 적용하였을 경우는 이러한 첨가단어가 급격이 줄어듦을 알 수 있다. 즉, 이로부터 첨가된 단어가 문법에 맞지는 않으나 음향학적으로 매우 모호하다는 것을 알 수 있으며 적당한 문법의 결합이 연속어 인식에서 얼마나 중요한가를 확인할 수 있다.

표. 2. 문법이 없을 때의 화자별 인식률.

Table 2. Recognition rate without grammar for each speaker.

		전체단어수	첨가	삭제	치환	인식률(%)	
						첨가포함	첨가제외
남성화자	GGHA	78	9	0	3	84.6	96.1
	SJG	81	3	0	6	88.9	92.6
	SSB	83	6	0	6	85.5	92.8
	YSJ	74	4	0	5	87.8	93.2
	GJT	82	7	0	8	81.7	90.2
여성화자	BSB	84	10	0	13	72.6	84.5
	HUT	82	4	0	8	85.3	90.2
	IGH	73	7	0	9	78.1	87.7
	JWS	85	5	0	9	83.5	89.4

표 3. Finite state network을 사용하였을 경우 화자별 인식률.

Table 3. Recognition rate with FSN for each speaker.

		전체 단어수	첨가	삭제	치환	인식률(%)	
						첨가포함	첨가제외
남성 화자	GGHA	78	0	0	2	97.4	97.4
	SJG	81	1	0	4	93.8	95.1
	SSB	83	0	0	5	94.0	94.0
	YSJ	74	0	0	3	95.9	95.9
	GJT	82	0	0	5	93.9	93.9
여성 화자	BSB	84	1	1	6	90.5	91.7
	HUT	82	0	0	5	93.9	93.9
	IGH	73	0	0	7	90.4	90.4
	JWS	85	0	1	3	95.3	95.3

IV. 결론 및 향후 연구 방향

본 논문에서는 연속분포 HMM을 이용한 한국어 연속음성 인식 시스템의 구성 및 실험결과를 정리하였다. 개발된 시스템은 100단어의 어휘를 가지는 한국어 시간구를 자연스럽게 연속적으로 발음한 음성을 인식하는 시스템으로, 화자독립 인식실험에서 문법이 적용되지 않았을 경우 83.1% , FSN을 적용하였을 경우 94.2%의 인식률을 보였다. 현재는 1,000 단어 규모의 연속어 인식이 가능하도록 시스템을 확장하는 과정에 있으며, 이에 따라 추가되어야 할 부분으로는 beam search 등의 기법을 이용한 시간 감축 알고리즘과 연속어에서 각 단어사이의 조음화 현상을 반영해 줄 수 있는 inter-word modeling, 대규모 어휘에 적당한 문법의 결합 등을 들 수 있다.

< 참 고 문 헌 >

[1] K. F. Lee, *Automatic Speech Recognition*, Kluwer Academic, 1989.

[2] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, vol. 4, pp. 127 - 165, 1990.

[3] C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg,

"Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, vol. 6, pp. 103 - 127, 1992.

[4] 은 종관 외, 연속음성인식 시스템의 개발 연구, 한국과학기술원, 1992.

[5] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of Isolated Digits using HMMs with Continuous Mixture Densities," *Bell Syst. Tech. J.*, vol. 64, pp. 1211-1234, July 1985.

[6] A. Averbuch et. al., "Experiments with the TANGORA 20,000 Word Speech Recognizer," *Proceedings of ICASSP*, pp. 701-704, April 1987.

[7] Y. L. Chow, et. al., "BYBLOS : The BBN Continuous Speech Recognition System," *Proceedings of ICASSP*, pp. 89-92, April 1987.

[8] B. H. Juang and L. R. Rabiner, "The Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 38, no. 9, pp. 1639-1641, 1990.

[9] E. Zwicker and E. Terhards, "Analytical expression for critical bandwidth as a function of frequency," *J. Acoust. Soc. America*, vol. 68, pp. 1523-1525, 1980.

[10] H. Ney, "The Use of a One-Stage Dynmic Programming Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, vol. 32, no. 2, April 1989.

[11] F. Jelinek, R. L. Mercer, and S. Roukos, "Principles of Lexical Language Modeling for Speech Recognition," In: *Advances in Speech Signal Processing*, ed. S. Furui and M. M. Sondhi, MARCEL DEKKER, NY, 1992.