음성인식에 의한 연구센타 부서안내 시스팀 개발에 관한 연구

구 명완, 손 일현, 도 삼주, 이 종락
한국통신 연구개발단 기초기술 1 연구실

A Study on the Development of Korea Telecom Automatic Voice Recognition System

Myoung-Wan Koo, Il-Hyun Sohn, Sam-Joo Doh and Jong-Rak Lee
Basic Research Section 1, Korea Telecom Research Center

요 약

이 논문에서는 음성인식기술을 이용한 연구센타 부서안내 시스팀(KARS:Korea Telecom Automatic voice Recognition System)에 대하여 기술하였다. 이 시스팀은 기본적으로 음성응답 시스팀과 유사하지만 명령입력을 위해 푸시버튼 대신 음성을 이용한다는 점이 다르다. 사용자가 마이크로폰을 통해 음성명령을 입력하면, 이 시스팀은 사용자의 음성명령을 인식하여 연구센타내 각 부서의 간략한 소개, 전화번호 및 위치를 안내해 준다. 이 시스팀은 HMM(Hidden Markov Model)을 이용하는 화자독립 격리단어 인식시스팀으로서 116개의 부서이름과 7개의 제어용 단어로 구성되어 있는 123개 단어를 인식할 수 있다. 이 시스팀은 음소와 유사한 한국어 서브워드(subword)를 HMM의 기본단위로 사용하며 인식 실험결과 98.6%의 인식율을 얻을 수 있었다.

I. 서 론

현재 상용화되어 있는 음성정보검색 시스팀(audio response system)은 명령어를 전화기의 버튼(push-button)을 눌러서 입력하면 필요한 정보가 음성으로 나온다. 그러므로 MFC 전화기를 소유하지 않은 사람은 이런 시스팀을 이용할 수 없으며, 또한 숫자 이외의 한글 명령어를 입력하기가 매우 어렵다.

이러한 단점을 보완하기 위하여 음성을 직접명령어로 입력할 수 있는 시스팀이 출현하였다.

미국내의 Ameritech 및 NYNEX 전화회사에서는 1989년에 Voice Service Nodes를 설치하여 수신자 요금부담(collect call), 제삼자 요금부담(third-number billed call) 및 전화카드(calling card) 요금부담 등을 음성으로 직접 명령할 수 있게 하였으며[1], 일본의 NTT에서는 ANSER (Automatic Answer Network System for Electrical Request)라는 시스팀을 도입하여 전화망을 통한 음성의 인식뿐만 아니라 음성 합성기능까지도 가능하게 하였다[2]. 한편 스페인의 전화회사에서는 12단어를 인식할 수 있는 AUDIOTEX라는 시스팀을 개발해 시험중에 있다[3].

본 논문에서는 음성인식이 가능한 연구센타 부서안내 시스팀(KARS:Korea Telecom Automatic Voice Recognition System) 개발에 관하여 기술한다. 먼저 Ⅱ장에서는 KARS 시스팀의 개요에 대해서 설명하고 Ⅲ장에서는 KARS시스팀의 성능에 대해서 기술하며 Ⅳ장에서는 결론을 맺는다.

Ⅱ. 시스팀 개요

KARS 시스팀의 개요는 그림 1.에 나타나있다. 이 시스팀은 음소모델 및 음성메시지로 구성된 2종류의 데이타 베이스를 사용하고 있다. 음소모델은 HMM 훈련과정에서 얻어져서 Viterbi decoder에 의해 사용되는며, 음성메시지는 KARS와 사용자와의 음성대화를 위한 메시지로서 미리 녹음되어 저장된다. KARS에 음성이 입력되면 먼저 음성구간이 검출되고 음성특징이 추출된다. 대화관리는 KARS의 모든 제어를 관리하는 프로세서로서 입력된 음성특징을 Viterbi decoder로 전공하여 입력음성을 인식하거나, 인식된 결과에 따라 전화번호, 부서소개 및 위치안내를 음성으로 들려주는 제어를 담당한다.

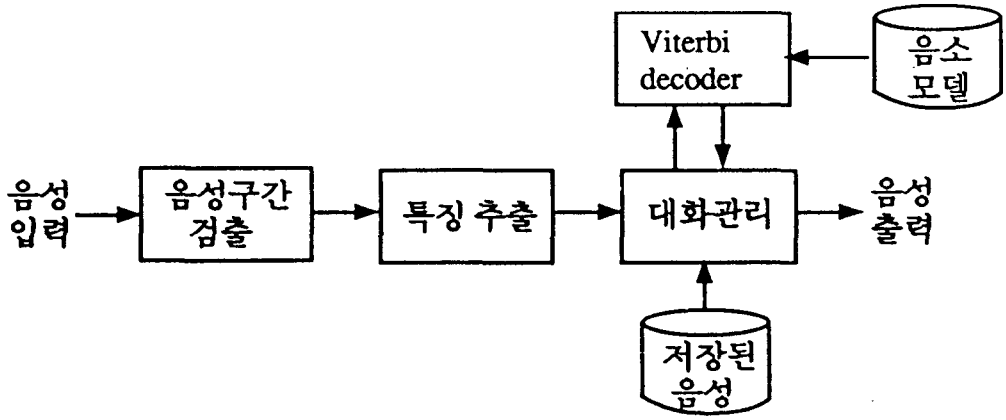

그림 1. KARS 시스팀 개요

1. 특징 추출

음성은 10kHz로 샘플링되며 매 샘플링된 음성은 10 msec 단위의 프래임으로 분할된다. 매 프레임에서는 LPC(linear predictive coding) 분석에 의한 cepstral 계수를 구한 후 mel scale로 변환한다. 현재 사용중인 특징은 12차의 mel-scaled LPC cepstral coefficients, 12차의 델타 LPC cepstral coefficient, 에너지 및 델타 에너지이다. KARS는 VQ(vector quantization)-based

HMM 인식 시스팀이므로 VQ codebook이 필요하다. 우리는 3종류의 VQ codebook을 사용하며 매 codebook은 256개의 codeword로 구성된다.

2. 대화관리

대화관리는 그림 2.와 같이 finite state diagram 으로 구성된다. 매 state 에서는 음성에 의한 대화가 가능하도록 음성인식 및 음성출력기능을 수행한다. 실제적으로 KARS 시스팀을 사용할 경우 발생되는 대화과정의 예를 그림 3.에 나타내었다.

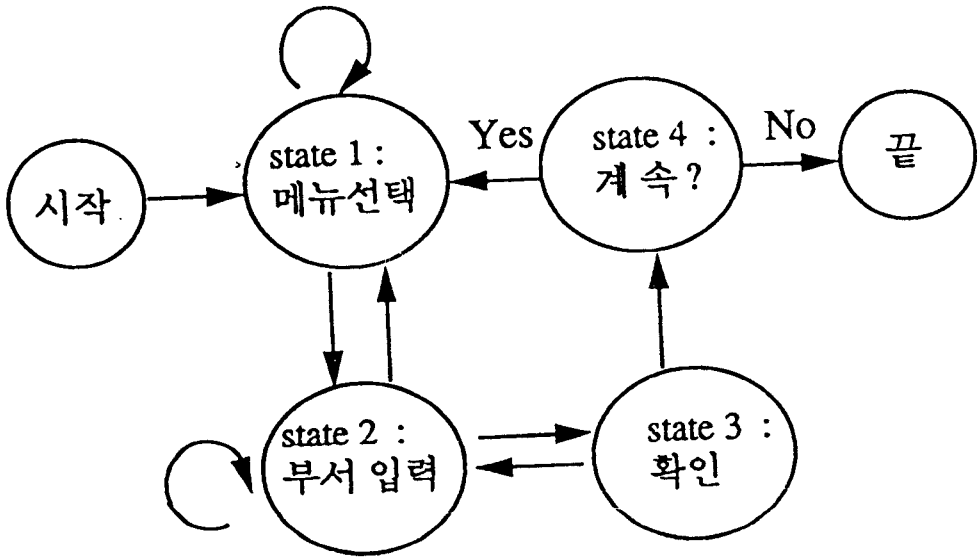

그림 2. State diagram

3. 음소모델과 데이타 베이스

KARS의 기본 유니트로 음소와 유사한(phoneme-like unit) 서브워드를 사용한다. 먼저 한국어의 음소를 44개의 context-independent 음소로 구성하였으며 인식율을 향상시키기 위하여 context-dependent 음소 모델도 사용하였다. context-dependent 음소모델을 구성하는 방식은 Lee의 방식과 유사하다[4]. Context-dependent 음소모델은 음성의 co-articalation 영향을 고려하기 위하여 문맥 변화에 따라 동일한 음소도 음가가 변환한다는 사실을 고려한 모델이다.

KARS 시스팀에는 2종류의 context-dependent 음소모델(triphone, pentaphone)을 사용하였다. Triphone 모델은 매 음소 P 를 $P \rightarrow P_L\text{-}P\text{-}P_R$ 로 표시하며 pentaphone 모델은 매 음소 P 를 $P \rightarrow P_{L2}\text{-}P_{L1}\text{-}P\text{-}P_{R1}\text{-}P_{R2}$ 로 표현한다. 여기서 P_L 은 P 의 좌측에 있는 음소이며 P_R 은 우측에 있는 음소이다. 또한 P_{L1}, P_{L2} 는 각각 P, P_{L1} 의 좌측 음소이며 P_{L1}, P_{L2} 는 각각 P, P_{R1} 의 우측에 위치하는 음소이다. 이러한 방식으로 context-dependent 음소를 만들면 음소의 양이 너무 방대하여져 한정된 양의 훈련음성 데이타베이스로는 통계적으로 적절한 모델을 만들 수 없는 때문에 음소의 양을 줄이는 것이 필요하다. 이를 위하여 본 논문에서는 unit reduction 규칙을 사용하

였다[4].

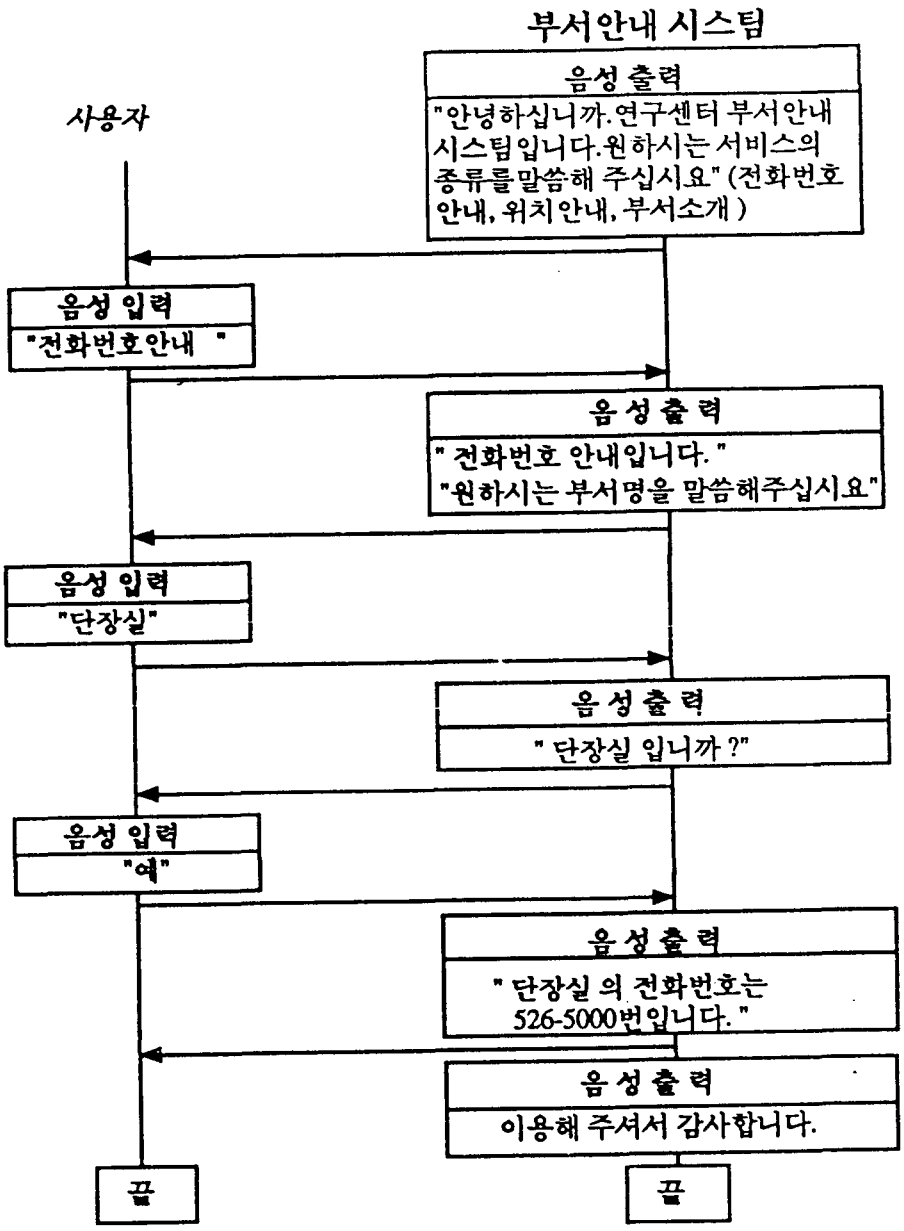

그림 3. KARS와의 대화과정 예

기본 유니트를 표현하기 위한 topology 는 그림 4.과 같이 7 state 와 12 transition 갖는 모델을 사용한다. 매 transition 은 3종류의 그룹으로 묶어 3종류의 출력확률(그림 4.의 B,M,E)로 표현한다[5].

한편 음소길이 정보를 구현하기 위해서는 state transition 길이 모델을 사용하였다. state transition 길이 모델은 훈련과정에서 매 음소 state transition 의 최대값과 최소값을 구한 후 인식과정에서 이 영역을 벗어나는 state sequence를 제외시켜 버리는 방식이다[6].

KARS 에 사용될수 있는 언어는 116 단어의 부서이름과 7 단어의 제어단어를 포함한 123 단어이다. 매 단어의 평균음절은 5.3이다. 전체 10명의 남성이 123단어를 2번 발음하였으며 그중 8

명의 음성을 훈련 데이타베이스로 사용하며 나머지 2명을 시험 데이타베이스로 이용한다.

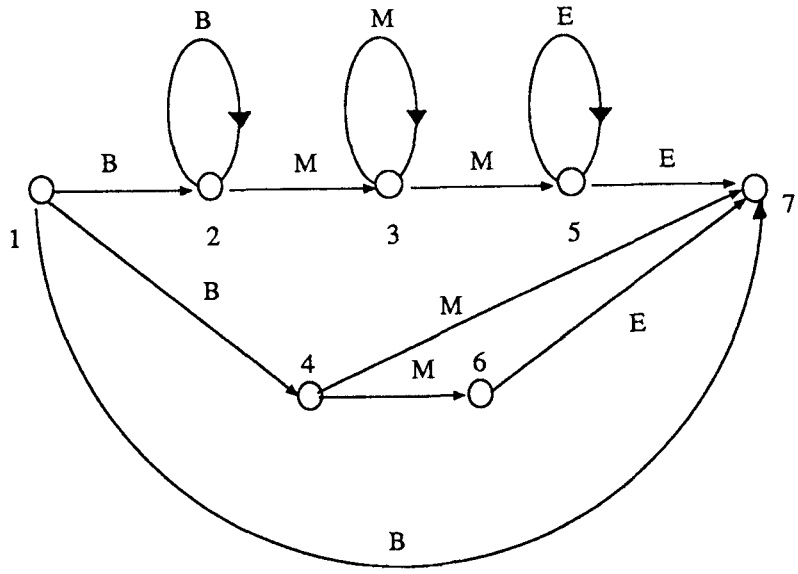

그림 4. HMM 서브워드 유니트

4. HMM 훈련과 인식

HMM 은 hidden stochastic process 와 출력심볼을 표현하여 주는 stochastic process 를 갖는 double stochastic process 이다[7]. HMM의 파라미타를 구하기 위한 훈련 알고리즘으로는 Baum-welch 알고리즘을 사용하였으며 HMM 파라미터의 성능을 향상시키기 위해 corrective training(CT) 알고리즘도 사용하였다. CT 알고리즘은 Baum-welch 알고리즘에 의해 구해진 HMM 파라미터를 초기값으로 하여 훈련 음성데이타의 자체 인식율을 높일 수 있도록 모델 파라미터를 변경시켜 주는 알고리즘이다.

HMM의 인식 알고리즘으로는 Viterbi beam 검색알고리즘을 사용하였다. 본 논문에서는 beam width 를 정하여 매 프레임에서 가장높은 likelihood 순으로 beam width 만큼만 고려하였다.

Ⅲ. 실험 결과

우선 discrete HMM 에 근거하는 화자독립 고립단어 인식 시스팀인 기본 시스팀을 개발하였다. 그 후 이 시스팀의 성능을 향상시키기 위하여 여러종류의 실험을 수행하였다. 기본 시스팀은 44개의 context-independent 서브워드 유니트를 사용하였으며 음소분할된 고립단어와 음소분할되지 않는 고립단어를 이용하여 HMM 파라미터를 구하였다.

1. Beam 검색

KARS beam 검색을 위한 조건으로 beam width를 사용하는데 beam width의 변화에 따른 인식율은 표 1에 나타나있다. Beam width가 작으면 음성 인식시간은 빨라지나 인식율이 저하되므로 인식시간과 인식율과의 trade-off가 고려되야한다. 표 1에 의해 Beam width가 10일 경우 첫번째 후보에 대한 인식율(top1)이 95.5 %로 나타났으며 첫번째 후보 및 두번째 후보를 포함한 인식율(top2)이 99.5%로 나타났다. 이 Beam width 10은 인식시간과 인식율의 trade-off를 고려한 beam width값으로 결정하였다.

표 1. Beam width 변화에 따른 인식율

	인 식 율 (%)					
Beam width	5	7	8	10	20	30
Top 1	88.0	93.3	94.1	95.5	94.9	95.5
Top 2	90.2	97.2	98.0	99.6	99.9	99.6

2. Corrective training

기본 시스팀의 인식율 향상을 위하여 CT 알고리즘을 수행하였다. CT 알고리즘의 최대 반복횟수를 2로 고정시켰으며 그때의 인식율은 표 2에 나타나있다. 표 2의 결과에 따르면 CT 반복횟수가 2이고 beam width 가 10으로 고정되어 있을 경우 96.7 %의 인식율이 나타났다.

표 2. CT를 적용할 때의 인식율
(Beam width = 10)

	인 식 율 (%)		
반복 횟수	0	1	2
Top 1	95.5	96.5	96.7
Top 2	99.6	99.9	99.6

3. Context-dependent 음소모델

음성인식율을 향상시키기 위한 또 다른 시도로 context-dependent 음소모델을 사용하였다. 2장의 unit reduction rule에 의해 triphone 모델로 51, 75 및 117개의 context-dependent 음소를 얻었고 pentaphone 모델로 131개의 음소를 얻었다. 이러한 context-dependent 음소 모델을 사용하고 Beam width를 10으로 하고 CT 알고리즘을 수행하였을 경우의 인식율은 표 3에 나타나

있다. 표 3을 보면 117개의 context-dependent 음소와 CT 반복횟수를 2로 고정하고 beam width를 10으로 정하였을 경우 97.4 %의 인식율이 나타났다.

표 3. 음소의 갯수에 따른 인식율의 변화

	인 식 율				
음소 갯수	44	51	75	117	131
CT 비사용	95.5	96.3	96.5	96.9	96.3
CT 반복횟수=1	96.5	96.3	96.7	97.4	96.9
CT 반복횟수=2	96.7	96.7	96.5	97.6	96.9

4. 음소길이 모델

음성 인식율을 높이기 위한 마지막 시도로 음소 state transition 길이정보를 이용하였다. 먼저 훈련 과정에서 매 음소의 모든 state transition 길이의 최대값과 최소값을 구한다. 인식 과정에서는 이 최대 및 최소값을 이용하여 매 state transition 길이가 이 범위에 있는 경우만 인식한다. 표 3에서 117과 135개의 context-dependent 음소를 사용하고 매 음소에 state transition 길이 정보를 구현 했을 경우 인식율이 표4에 나타나있다. 표4의 결과로 부터 131개의 context-dependent 음소에 state transition 길이 정보를 구현하고 CT 반복횟수가 2일 경우 98.6%의 인식율을 얻을 수 있었다.

표 4. 음소길이 정보을 사용했을 경우의 인식율

	인 식 율	
음소 갯수	117	131
CT 비 사용	97.4	97.4
CT 반복횟수=1	97.6	98.4
CT 반복횟수=2	97.8	98.6

Ⅳ. 결 론

본 논문에서는 음성을 인식할 수 있는 한국 통신 연구센타 부서안내 시스팀(KARS) 개발에 대하여 기술하였다. KARS는 123단어를 인식할 수 있는 화자독립 고립단어 인식 시스팀이며 음소와

유사한 서브워드를 기본 단위로 사용하는 HMM 인식 시스팀이다. 이 시스팀의 최고 성능은 131개의 context-dependent 음소와 state transition 길이 정보 및 CT 알고리즘을 사용했을 경우 98.6%로 나타났다. 현재 이 시스팀은 Sun SPARC station Ⅱ에 demonstration 시스팀으로 개발되어 있다. 앞으로 전화망과 정합시켜 성능실험을 할 예정이며 시험 결과에 따라 연구센타 부서안내 시스팀으로 활용할 예정이다.

< 참 고 문 헌 >

[1] M. Lenning, "Putting speech recognition to work in the telephone network," *IEEE computer*, vol. 23, no. 8, pp. 35-41, Aug. 1990.

[2] R. Nakatsu, "Anser: an application of speech technology to the Japanese banking industry," *IEEE computer*, vol. 23, no. 8, pp. 43-48, Aug. 1990.

[3] M. J. Poza et al., "An approach to automatic recognition of keywords in unconstrained speech using parametric models," in *Proc. 2nd European Conf. on Speech Comm. and Tech.*, Sep. 1991, pp. 471-474.

[4] C. H. Lee et al., "Acoustic modeling of subword units for speech recognition," in *Proc. 1990 IEEE Int. Conf. Acoust., Speech, Signal Processing*, April 1990, pp. 721-724.

[5] K.-F. Lee, *Automatic speech recognition: the development of the SPHINX system*. Kluwer Academic Publisher, Norwell, Mass., 1989.

[6] Hung-yan Fu, et al., "Isolated-utterance speech recognition using hidden Markov models with bounded state durations," *IEEE Trans. on Signal Processing*, vol. ASSP-39, pp. 1743-1752, Aug., 1991.

[7] L. R. Rabinar and B. H. Juang, "An introduction to hidden Markov models," IEEE *Acoustic., Speech, Signal Processing Magazine*, Jan., 1986.