《主 題》

실시간 음성인식 다이얼링 시스템 개발
Development of a Real-time Voice Recognition Dialing System

이세웅 · 최승호 · 이미숙 · 김홍국 · 오광철 · 김기철 · 이황수
(한국과학기술원 서울분원, 정보및 통신공학과)

■ 차 례 ■

*본 논문의 일부는 한국이동통신(주)의 연구비 지원으로 이루어진 결과임.

ABSTRACT

This paper describes development of a real-time voice recognition dialing system which can recognize around one hundred word vocabularies in speaker independent mode. The voice recognition algorithm is implemented on a DSP board with a telephone interface plugged in an IBM PC AT/486. In the DSP board, procedures for feature extraction, vector quantization(VQ), and end-point detection are performed simultaneously in every 10msec frame interval to satisfy real-time constraints after the word starting point detection. In addition, we optimize the VQ codebook size and the end-point detection procedure to reduce recognition time and memory requirement. The demonstration system is being displayed in MOBILAB of Korea Mobile Telecom at the Taejon EXPO'93.

Ⅰ. 서 론

음성을 이용한 사람과 기계간의 통신은 가장 자연스럽고 편리한 것으로 각종 기기에 음성인식 기능을 부가하여 사용하려는 노력이 활발히 진행되고 있다. 특히, 고도의 정보화 사회로의 발전이 가속화되고 이에 따른 정보수요의 급격한 증대로 인해 정보서비스 시스템들이 다양화되면서 보다 자연스럽고 효율적인 음성 인터페이스의 필요성이 절실해지고 있다. 이는 인건비 상승 등으로 인한 적절한 서비스 인력의 확보난을 더는 데도 유리하며 오히려 인력의 사용 보다 더욱 양질의 정보통신 서비스를 제공할 수 있다는 가능성을 갖고 있다.

본 연구에서는 이동통신의 기지국에서 다양한 정보서비스를 제공할 수 있는 음성인식 다이얼링 시스템을 개발하였다. 이를위해 먼저, 자동음성호출 교환국의 처리 시나리오에 따라 workstation상에서 일반적인 탁상용 마이크 입력을 사용하는 데모 시스템을 구

축하였으며, 이를 다시 PC 상에서 DSP 보드를 이용한 실시간 음성인식 다이얼링 시스템으로 구현하였다. 본 논문에서는 PC 상에서 구현된 실시간 음성인식 다이얼링 시스템을 위주로 서술하였다.

제 II장에서는 실시간 음성인식 다이얼링 시스템을 간략히 살펴보고, III장에서는 실시간 처리에 적합한 음성분석 및 인식 알고리즘에 대해서 설명하였다. 제 IV장에서는 데모 시스템의 인식실험 결과를 검토하고 마지막으로 V장에서 결론을 맺도록 하였다.

II. 실시간 음성인식 다이얼링 시스템 개요

2.1 처리순서

음성인식 다이얼링 시스템은 7자리의 전화번호를 차례로 발음하는 "번호", 엑스포, 이동통신, 과학원, 청와대, 시청, 방송국, 기상청, 병원 등 공통적으로 사용되는 전화번호를 미리 등록한뒤 호출할 수 있는 "기관", 각 개인별로 10개까지의 특정이름을 등록한 뒤에 호출할 수 있는 "이름"과, 가장 최근에 호출했던 전화번호를 다시 다이얼링해 주는 "다시"의 네가지 기본모드가 있으며, 그외에도 "취소", "안내"모드를 선택할 수 있도록 구성되었다.

본 시스템은 각 호출방식에 따라 발음된 번호나 이름을 인식한 뒤에 사용자에게 해당 전화번호를 다시 확인하며, 이때 사용자로부터 "예, 응, 그래, 에스, 오케이"등의 대답을 인식하게 되면 번호를 다이얼링해 주고"아니오, 아니, 노우"등의 대답을 인식하게 되면 사용자에게 호출방식을 다시 선택하게 한다. 이때 오인식이나 통화중인 경우를 대비해 반복 다이얼링 횟수를 제안하였으며, 착신번호와 연결되거나 반복횟수 제한에 걸리면 시스템은 종료 메시지와 함께 수행을 끝낸다. 다음 그림1은 본 논문에서 제안하고 있는 실시간 음성인식 다이얼링 시스템의 처리 예를 보여주고 있다.

2.2 하드웨어 구성

본 연구에서 개발된 실시간 음성인식 다이얼링 시스템의 H/W 구성은 그림2와 같다. ELF DSP platform은 TMS320C31 부동소숫점 디지탈 신호처리기와 16-bit A/D, D/A 변환기를 포함하는 DSP board이고

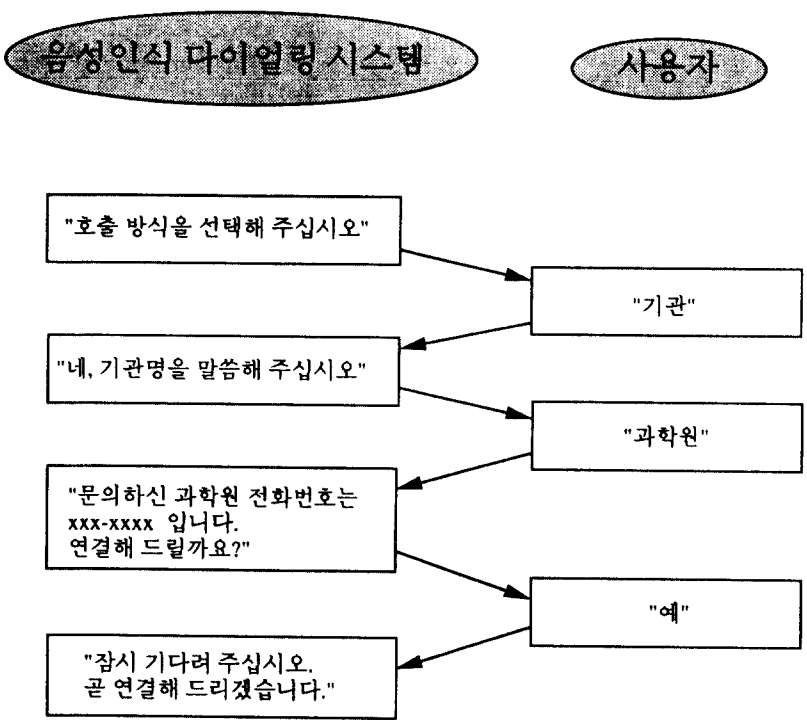

그림 1. 실시간 음성인식 다이얼링 데모 시스템의 처리 예

PC의 16-bit AT 버스 슬롯에 연결된다. 음성인식 다이얼링 시스템의 프로그램은 이 DSP 상에서 수행되어 실시간으로 인식하도록 구성하였다[1].

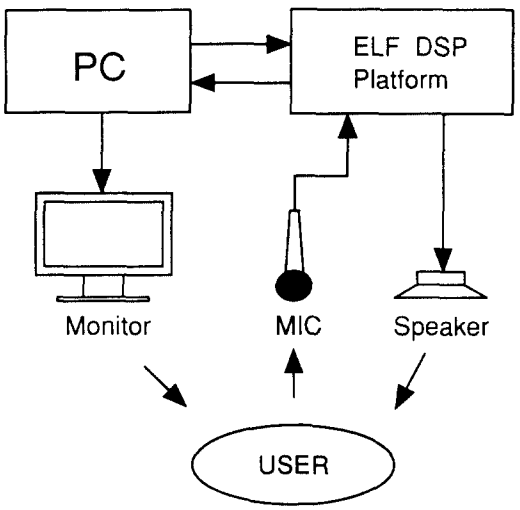

그림 2. 실시간 음성인식 다이얼링 시스템의 H/W 구성도

2.3 소프트웨어 구성

2.2절에서 설명한 데모 시스템의 하드웨어를 기본으로 이를 동작시키기 위한 소프트웨어 구성도는 그림3과 같다.

마이크를 통해 입력된 신호는 A/D변환을 거친 뒤에 음성구간을 검출하는 끝점 검출기로 들어간다. 끝점 검출기에서 나온 신호는 음성만을 포함하게 되며, 이는 특징 추출부에서 16차의 mel-cepstrum으로 변환된다[2,3]. Flow controller는 인식모드가 화자종속 또는 화자독립이냐에 따라 각각 DTW 또는 discrete HMM 알고리즘[4,5]이 수행될 수 있도록 제어신호를 보낸다. 화자독립 인식모드의 경우, 특징 벡터들은 양자화되고 다시 각 단어에 대한 HMM 모델들로부터 Viterbi score를 계산한다[6]. 이때 각 양자화 codebook과 HMM 모델들은 각각 4개의 군으로 분류되어 있으며 이들은 flow controller의 제어를 받는다[표 3참조].

화자종속 인식모드는 최대 10개 단어에 대해 사용자의 목소리로 세번씩 발음한 것을 기준패턴으로 하여, 입력 음성 패턴과의 거리(distance)를 최소로 하는 기준 패턴을 찾아내는 작업을 한다. 이상과 같은 과정으로 화자독립 또는 종속으로 입력 음성에 대한 score 또는 거리값으로부터 단어를 인식하고 그 결과를 flow controller에 보낸다. Flow controller 인식된 단어로부터 출력 메시지를 결정하여 message handler에 보내는 한편 각 블럭에 대해 적절한 제어신호를 보낸다.

Ⅲ. 실시간 음성인식 소프트웨어

3.1 실시간 끝점검출 알고리즘

음성신호는 10 msec를 한 프레임으로 하여 3프레임마다 단구간 에너지 {E(i)}와 영교차율 {Z(i)}를 식(3.1)과 식(3.2)와 같이 구한다. 여기서 LM과 UM은 하드웨어 시스템 잡음과 background noise의 영향을 배제하기 위한 margin값으로 본 연구에서는 8과 -10으로 정하였다. 3프레임의 평균에너지 m_e와 3프레임동안 {Z(i)}의 평균과 표준편차인 m_z와 σ_z를 구하여 단구간 에너지의 상한(UEL) 및 하한 임계치(LEL)와 영교차율의 임계치(TZCR)를 식(3.3)과 같이 구한다[7,8]. 각 임계치의 가중치는 0.95^{n}이고, 여기서 n은 3프레임마다 1씩 증가한다. 최초 3프레임씩 매 프레임마다 m_e, m_z, σ_z를 구하여 에너지에 따른 잠정적인 시작점 Ns를 매 프레임마다 검색한다. 같은 방식을 반복하여 시작점을 찾고 역시 최초 3프레임씩 매 프레임마다 끝점을 찾는다.

$$E(i) = \sum_{n=0}^{79} |s(80i+n)|,\ i=0,\ 1,\ 2 \tag{3.1}$$

$$Z(i) = \sum_{n=0}^{79} |sgn(s(80i+n)) - sgn(s(80i+n-1))|,\ i=0,\ 1,\ 2 \tag{3.2}$$

여기서

$$sgn(s(n)) = \begin{cases} 1, & s(n) \geq UM \\ 1, & s(n) \leq LM \end{cases}$$

$$LEL = UEL_0 = \min(3m_e,\ 500.0)$$

$$UEL = UEL_0 = 4LEL$$

$$TZCR = TZCR_0 = \min(m_z + 2\sigma_z,\ 25) \tag{3.3}$$

Pause의 검색은 20프레임동안 수행하며 20프레임 안에 새로운 시작점이 검출되지 않으면 끝점으로 정해진다. Pause는 단어를 발음하는 중에 묵음이 있는 것으로 숫자음에는 pause가 없으므로 숫자음의 끝점 검출에서는 pause 검색을 하지 않는다.

실시간 음성인식을 위한 알고리즘의 변경 이유는

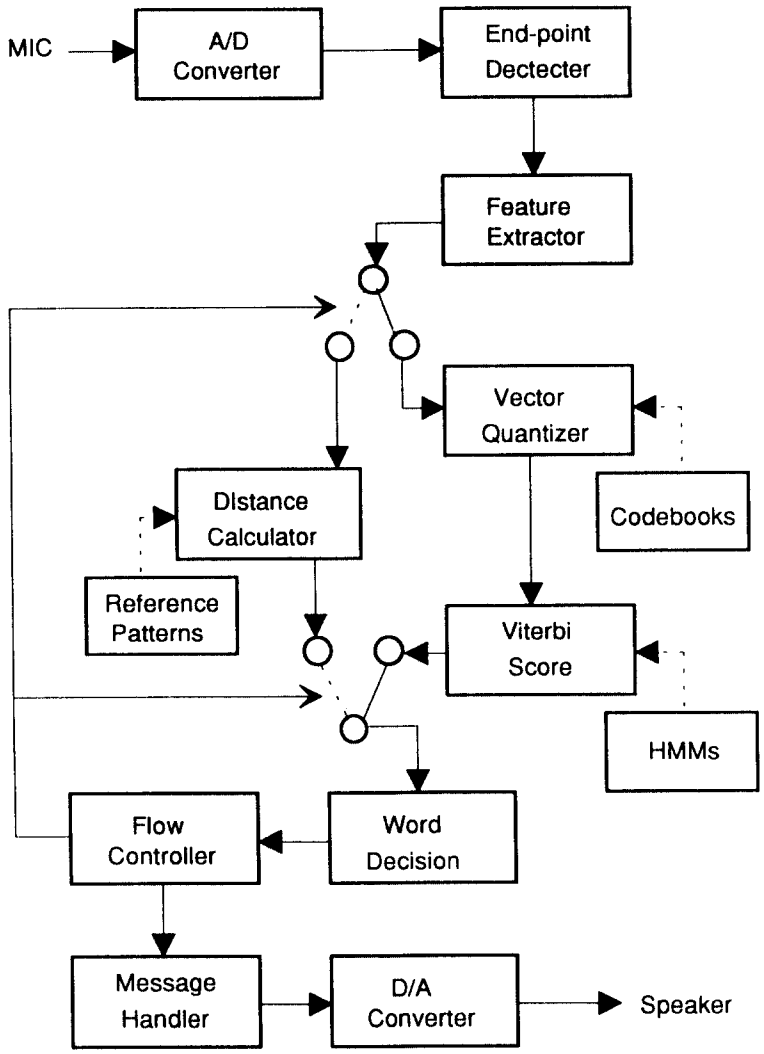

그림 3. 음성인식 다이얼링 시스템의 S/W 구성도

다음과 같다. 첫째, 에너지에 의한 시작점 검출시 그 잠정적인 시작점의 항상 현재 프레임이 될 수 있도록 프레임을 중첩시키면서 현재 프레임을 포함한 최근 3 프레임동안의 에너지를 계산하였다. 둘째, 특징추출 과정에서 3프레임을 이용하므로 프레임간의 지연없이 끝점검출과 특징추출 과정을 동시에 수행할 수 있다. 셋째, pause 검색 구간이 길면 발성이 끝나 끝점을 찾아도 pause 검색으로 인식 알고리즘의 수행을 못하므로 인식 시간에 영향을 주기때문에 최대한 짧게 20 프레임으로 하였다. 20프레임은 0.2초동안의 음성신호로서 실제 단어 사이에 0.2초를 넘는 묵음은 없다고 보았다.

3.2 VQ codebook 크기의 최적화

Discrete HMM 기반의 음성인식 시스템에서는 VQ codebook의 크기는 인식 시간과 기억 용량에 커다란 영향을 준다. Codebook의 크기가 커질수록 특징 벡터를 VQ codeword로 양자화하는 시간이 길어지며, codebook이 차지하는 기억 용량의 크기가 커지게 된다[9,10,11]. 또한 인식대상 단어의 갯수가 작고 특징

패턴이 다양하지 않을때 커다란 codebook이 반드시 인식률 향상을 가져오지는 않는다. 본 실시간 음성인식 다이얼링 시스템은 인식 어휘의 갯수가 적고 실시간 처리가 주요 목적으로, 인식률의 저하를 가져오지 않는 범위내에서 작은 크기의 codebook을 선택하려고 했다. 이를 위해 숫자음에 대한 화자 독립의 simulation을 해 보았다. 이때 DB 구성은 다음과 같다.

- 녹음 환경 : 조용한 환경의 사무실
- training 데이타 : 남5명, 여5명(10회 발성)
- test 데이타 : traninig에 참가하지 않은 남5명, 여5명(10회 발성)
- 표본화율 : 10kHz(12bit)

표1은 VQ codebook의 크기에 따른 인식률의 변화를 나타내고, 표2는 단어 수11개, 단어당 HMM state 수가 최대 18개인 경우에 VQ codebook의 크기에 따른 기억용량과 VQ encoding의 속도를 나타낸다. 본 인식 시스템은 이와 같은 simulation 결과를 참조하여 codebook의 크기를 64로 정하였다.

표 1. VQ codebook 크기에 따른 숫자음 인식률

숫자 / 크기	영	일	이	삼	사	오	육	칠	팔	구	total
64	83	74	94	96	100	96	100	92	95	94	92.4
128	77	77	88	100	85	93	100	95	100	96	91.1
256	73	78	100	100	82	91	100	99	97	97	91.7

표 2. VQ codebook 크기에 따른 기억용량 및 VQ encoding 속도

	VQ codebook의 기억용량(word)	HMM model의 기억용량(word)	VQ encoding의 속도(msec)
64	1024	16,236	0.17
128	2048	28,908	0.33
256	4096	54,252	0.65

3.3 실시간 인식을 위한 On-line 음성분석

발음된 단어의 끝점이 검출된 이후부터 특징추출 및 인식 알고리즘을 수행하는 대신, 실시간 처리를 위하여 음성의 시작점이 검출된 이후, 특징추출 및 벡터 양자화 과정을 끝점검출과 함께 수행하도록 하였다.

먼저 A/D 변환기에서 음성신호를 8 kHz로 sampling하고 10 msec마다 80개의 음성신호의 sample이 취해져 한 프레임을 구성하여 처리된다. 화자독립으로 인식할 때에는 음성신호를 16차의 mel-cepstrum (1.65 msec)으로 만든 후 그 특징 벡터를 양자화 (0.17 msec)하고 음성의 끝점인지를 검출 (0.2 msec)한다. 이 과정을 10 msec마다 반복하여 끝점이 검출될 때까지 음성 신호의 양자화된 값을 구한다. 화자의 발성이 끝남과 동시에, 만들어진 음성신호의 양자화된 데이타에 대해 각 단어의 HMM 모델들로부터 Viterbi score를 계산하여 가장 큰 값을 갖는 모델을 찾아 인식단어를 선정한다. Viterbi scoring에 걸리는 시간은 비교할 단어의 수, state의 수와 구조, 양자화 codebook의 크기 등에 따라 다르다.

화자종속으로 인식할 때에는 음성신호를 매 프레임마다 16차의 mel-cepstrum으로 만들어 저장하였다가 화자의 발성이 종료되어 끝점이 검출되면 음성신호부분의 특징 벡터패턴을 기준패턴들과 비교하여 가장 가까운 패턴의 단어로 인식한다. 기준패턴들과의 거리비교에 걸리는 시간은 비교할 단어의 수, 단어의 길이, 기준패턴의 갯수 등에 따라 달라진다.

Ⅳ. 시스템 성능평가

4.1 음성 데이타베이스

본 시스템의 인식 성능을 고찰하기 위하여 PC 상에서 ELF DSP board를 통해 받은 음성을 대상으로 실시간 인식 시스템에서 사용한 것과 같은 알고리즘으로 workstation을 사용하여 인식 simulation을 하였다. 다음은 음성 DB의 구축 환경을 나타낸다.

- 녹음 환경 : Icom HS-58
- 음성 DB 내용 : 5개군별로 6-11개의 고립 단어
- tranining data : 남17명, 여10명
- test data : 남10명, 여6명
- 발음 횟수 : 3회(남, 숫자음은 5회)
- 표본화율 : 8kHz(16bit)

음성의 모델링 단위를 단어로 하였으며, 각 단어에 할당하는 HMM state의 갯수는 단어를 구성하는 음소의 갯수에 비례하며 각 음소는 3개의 state로 구성된다. 본 연구에서는 단어에 대한 모델로서 discrete HMM을 선택하였고 workstation 상에서 training하여 HMM을 구성하였다. 우선 VQ codebook을 얻기 위하여 modified K-means 알고리즘을 사용하였고 codebook의 크기는 인식 시간, 메모리, 인식 어휘의 양 등을 고려

하여 64개로 하였으며, HMM의 파라미터를 구하기 위하여 Baum-Welch training 알고리즘을 사용하였다.

4.2 인식실험 결과

각 단어군에 대한 인식률은 표3과 같다. 여기서 응답군에 대한 인식률은 /예/ 범주와 /아니오/범주로 나뉠때 같은 범주에 소간 것들에 대한 인식률이다. 그리고 숫자음에서도 /영/과/공/은 같은 단어로 취급하여 인식률을 구했다. 표3에서와 같이 명령어군, 기관군, 이름군, 응답군 등에서는 거의 100%의 인식률을 보인다. 그러나, /일/, /칠/, /공/, /구/등의 숫자음은 인식률이 저조하였다. 표4에서는 숫자음 인식시 오인식된 것을 리스트하였다.

조용한 사무실 환경에서 탁상용 마이크 입력으로 녹음한 총 20명이 발음한 숫자음을 손으로 끝점 검출한 뒤에 화자독립으로 인식 실험한 결과들을 보면, 훈련에 포함된 화자는 약 99%, 포함되지 않은 화자는 약 92%의 인식률을 나타내었다. 즉, 명령어군, 기관군, 이름군, 응답군 등의 다음절 단어는 음성구간 검출 과정에서 약간의 오류가 포함되어도 단어 전체를 비교하는 과정에서 단어의 분류가 가능하게 숫자음의 경우 음성구간 검출에서의 작음 오류가 숫자음의 인식에 치명적인 영향을 줄 수 있다. 따라서 숫자음의 인식률을 향상시키기 위해서는 잡음환경에서의 음성구간 검출 과정의 정확도가 개선되어야 한다.

표 4. 오인식 숫자음 리스트

result / test	일	이	삼	사	오	육	칠	팔	구	공	영
일		9					15			1	2
이											
삼				2				3			
사			1								
오									2	5	
육											5
칠	15	3									
팔			1				1				
구					5					8	
공					13				2		1
영			1	1		2					

4.3 인식시간 및 기억용량

표5는 실시간 음성인식 다이얼링 시스템의 각 모듈별 처리시간과 기억용량을 나타낸다. 각 모듈의 크기는 수행하는데 필요한 header화일 및 데이타 화일을 포함하고 있으므로 실제 전체 프로그램의 크기는 각 모듈의 크기를 합한 것보다 작아지게 되어 약 74 kword 정도이다.

화자독립으로 인식할 때에는 단어의 시작점이 검출된 후, 매 10msec마다 끝점검출, 특징추출 및 벡터양자화 과정이 수행되며 끝점이 검출되자마자 Viterbi

표 3. 인식대상 단어와 인식률

단어군	항목(인식률)	총 인식률
명령어	안내(97.9), 기관(97.9), 이름(100.0), 번호(100.0), 다시(100.0), 취소(100.0)	99.3%
기관	엑스포(100.0), 이동통신(100.0), 과학원(100.0), 청와대(97.9), 시청(95.8), 방송국(100.0) 기상청(97.9), 병원(100.0)	99.0%
이름	꿈돌이(100.0), 도우미(97.9), 백원섭(100.0) 영구(100.0), 김미화(100.0), 서태지(95.8), 신신애(100.0), 최불암(100.0), 고두심(100.0), Y.S(91.67)	98.5%
응답	예(97.9), 예스(97.9), 응(85.4), 그래(95.8), 오케이(100.0), 아니다(100.0), 아니(100.0), 노우(100.0)	97.1%
숫자음	일(66.3), 이(100.0), 삼(93.8), 사(98.8), 오(91.3), 육(93.8), 칠(77.5), 팔(97.5), 구(82.5), 공(81.3), 영(95.0)	88.8%

scoring이 진행된다. 단어가 10개, 최대의 state 수가 18, codebook의 크기가 64인 경우에서 Viterbi scoring에 걸리는 시간은 360 msec로 화자가 발성한 후 짧은 시간이 경과된 후에 인식결과에 따른 응답메시지를 들을 수 있다.

화자종속으로 인식할때 패턴비교에 걸리는 시간은 단어가 10개, 단어의 최대 허용 길이가 80프레임, 같은 단어의 기준패턴 갯수가 3인 경우에 285 msec로 역시 실시간 처리가 가능하다.

표 5. 각 모듈별의 처리속도

모듈	처리속도(msec)	크기(word)
끝점검출	0.2	32,350
특징추출	1.65	24,900
벡터 양자화	0.17	25,126
Viterbi scoring	360	39,197
DTW	285	64,482

V. 결 론

본 논문에서는 이동통신 기지국에 음성인식 기능을 부여하기 위해 구현된 실시간 음성인식 다이얼링 시스템에 관해 기술하였다. 실시간 음성인식 다이얼링 시스템은 IBM PC AT/486 상에서 DSP board를 이용하여 전화번호, 이름, 기관명, 응답 등의 단어를 실시간에 인식할 수 있도록 구현되었으며, 현재 대전 EXPO'93의 한국이동통신(주) MOBILAB에서 전시되고 있다.

실시간 인식을 위해 음성구간 검출, 특징추출 및 특징벡터 양자화과정을 매 10msec마다 수행하는 한편, VQ codebook 크기를 최적화하였으며, 각 단어는 discrete HMM으로 모델링 하였다. 구현된 시스템에 대한 화자독립 인식실험 결과를 보며, 일반 사무실 잡음 환경에서 녹음된 명령어, 기관명, 이름, 응답 등의 단어에 대해서는 97%~99%, 숫자음의 경우 88.8%의 인식률을 나타낸다.

향후 잡음에 강한 음성분석 및 특징추출에 관한 연구를 통해 숫자음에 대한 인식률 향상이 기대되며, 실제 이동통신 기지국과의 인터페이스 및 이동전화 채널을 통과한 음성에 대한 왜곡 보상방안을 해결하면 음성 다이얼링 뿐 아니라 음성을 이용한 정보검색 등의 서비스가 가능해질 것이다. 또한 보다 대량의 단어를 실시간에 인식하기 위해서는 프로그램의 최적화와 함께 다중처리 등이 고려되어야 할 것이다.

참 고 문 헌

1. Atlanta Signal Processors Inc., *ELF DSP Platform Instruction Manual*, pp. 1-9, 1992.
2. S. Imai, "Cepstral analysis on the mel frequency scale," *Proc. of the ICASSP*, pp. 93-96, April 1983.
3. A. V. Oppenheim et al., "Discrete representaion of signal," *Proc. of the IEEE*, Vol. 60, No. 6, pp. 681-691, 1972.
4. X. D. Huang, Y. Akiri, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, 1990.
5. C. Myers, L. R. Rabiner and A. E. Rosenberg, "Performance trade-offs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. on ASSP*, Vol. 28, pp. 623-635, December 1980.
6. G. D. Forney, "The Viterbi algorithm," *Proc. of the IEEE*, Vol. 61, No. 3, March 1973.
7. 구명완외, "실시간 음성 끝점검출 알고리즘," 제5회 신호처리합동학술대회 논문집, 제5권1호 pp. 11-14, 1992.
8. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978.
9. Y. Linde, A. Buzo, and R.M.Gray, "An algorithm for vector quantization design," *IEEE Trans. on Communication*, Vol. 28, No. 1, pp. 84-95, January 1980.
10. R. M. Gray, "A vector-quantization-based preprocessor for speaker-independent isolated word recognition," *IEEE Trans. on ASSP*, Vol.. 33, June 1985.
11. L. R. Rabiner, J. G. Wilpon and B. H. Juang, "A segmental K-means training procedure for connected word recognition," *AT & T Technical Journal*, Vol. 65, No. 3, pp. 21-31, May 1986.

이 황 수

• 1971. 3~1975 .2 : 서울대학교 공학대학 전기공학과(공학사)
• 1976. 3~1978. 8 : 한국과학기술원 전기 및 전자공학과(공학석사)
• 1978. 9~1983. 2 : 한국과학기술원 전기 및 전자공학과(공학박사)
• 1975. 1~1975. 10 : 현대조선중공업(주) 설계부 사원
• 1983. 3~1989. 2 : 한국과학기술원 전기 및 전자공학과 조교수
• 1983. 3~1992. 1 : 한국과학기술원 전기 및 전자공학과 부교수
• 1992. 2~현재 : 한국과학기술원 서울분원 정보 및 통신공학과 부교수
• 1984. 4~1985. 5 : 미국 Stanford대학교 Information Systems Lab. Post Doc. 연구원
※ 연구분야 : 디지탈 통신, 이동통신, 신호처리(통신, 음성, 레이다)