論文2003-40CI-1-8

실시간 임베디드 음성 인식 시스템

(A Real-Time Embedded Speech Recognition System)

南 相 曄 *, 全 銀 姬 **, 朴 仁 政 **

(Sang-Yep Nam, Eun-Hee Jeon, and In-Jung Park)

요 약

본 연구에서는 음성인식 엔진과 데이터베이스에 필요한 메모리 규모를 최소화시킨 실시간 임베디드 음성 인식 시스템을 구현하였다. 실험을 위해 PCS 전화기에서 사용하는 40가지의 명령어와 10개의 숫자음으로 구성된 단어 목록을 만들고, 이들 단어들을 남,여 화자가 발성하여 음성 시료를 구했다. 채록된 음성을 대상으로 창크기 256표본의 단기 분석을 통해 선형 예측 계수를 구한다. 이때 고역강조를 통해 직류 성분을 제거하고 성문 등의 저역 필터효과를 제거하였다. 선형 예측 계수는 Levinson-Durbin 알고리즘을 사용해 구했고 이를 다시 켑스트럼 계수로 변환하여 인식을 위한 특징 벡터열로 구축하였다. 각 단어의 특징 벡터 열에 대해 Baum-Welch 추정법을 이용하여 HMM을 훈련시킨 다음, 기능성 계산을 통해 각 단어에 대한 인식을 수행하도록 하였다. 단어 인식을 위해 ARM CPU코어가 장착된 보드에 음성인식 엔진과 데이터 베이스를 포팅하여 실험용 임베디드 시스템을 구축하였다. 5가지 인식 계수집단에 대한 인식 실험을 실시하여 인식률이 좋은 계수 집단을 선정하였다. 전체적인 음성인식 엔진의 인식률은 95%이었고, 명령어에 대한 인식률은 96%, 숫자음에 대한 인식률은 94%로 나타났다.

Abstract

In this study, we'd implemented a real time embedded speech recognition system that requires minimum memory size for speech recognition engine and DB. The word to be recognized consist of 40 commands used in a PCS phone and 10 digits. The speech data spoken by 15 male and 15 female speakers was recorded and analyzed by short time analysis method, which window size is 256. The LPC parameters of each frame were computed through Levinson-Burbin algorithm and they were transformed to Cepstrum parameters. Before the analysis, speech data should be processed by pre-emphasis that will remove the DC component in speech and emphasize high frequency band. Baum-Welch reestimation algorithm was used for the training of HMM. In test phone, we could get a recognition rate using likelihood method. We implemented an embedded system by porting the speech recognition engine on ARM core evaluation board. The overall recognition rate of this system was 95%, while the rate on 40 commands was 96% and that 10 digits was 94%.

Keywords : 임베디드 시스템, 선형예측계수, 음성인식

* 正會員, 京文大學校 情報通信科

(Dept. of Information & Communication, Kyung Moon College)

** 正會員, 檀國大學校 電子工學科

(Dept. of Electronics Engineering Dankook Univ.)

接受日字:2002年7月2日, 수정완료일:2002年12月20日

1. 서 론

정보통신 산업의 발전에 따라 국. 내외의 임베디드 시스템의 시장도 급성장하고 있다. 특히 휴대폰같은 이

동성 기기에 휴먼 인터페이스(human computer interface) 기술인 음성인식 기술을 적용시켜야 하는 필요성이 대두되고 있다[1-3].

임베디드 시스템의 특징은 실시간 내장형 운영체제(real-time operating system)를 장착하여 사용하므로 메모리 크기의 한계 극복, 음성인식 기술의 이식성, 확장성 및 경량화에 중점을 두어야 한다[3, 4].

지금까지 자동차 내에서 운행 중 전화시의 음성인식에 관한 논문이나[3, 4], 음성인식 엔진을 PCS 전화기에 탑재해 실험한 논문이 있다[5]. 본 연구에서는 음성 데이터베이스를 확장하고 다양화하여 실험함으로써 더욱 일반적인 음성인식 장치를 구현하고자 한다.

본 연구에서 사용하는 음성 데이터 베이스는 다양한 계층의 발성 화자를 동원하여 체계적으로 음성 데이터를 수집한다. 수집된 음성 데이터의 음성 구간 설정 후 먼저 입력 음성의 DC 성분을 제거하고, 고역 성분 강조 후 중첩 이동 방식으로 256 샘플크기의 분할 데이터를 구성한다. Levinson-Durbin 알고리즘을 통해 이 데이터에서 선형 예측 계수가 구하고, 다시 켑스트럼 변환공식을 이용하여 필요한 특징 벡터열을 구한다[8]. 에너지, 영교차, 필터뱅크 등의 분석 파라미터들을 추출하여 음성구간 검출의 파라미터 셋트를 명령어, 숫자음별로 알맞은 셋트를 선정한다. 다시 검출된 음성 구간에 대한 재 분석 구간을 설정하여 추출된 인식 파라미터를 학습과정, 분류과정을 거쳐 최종적으로 인식하고자 하는 테스트 패턴이 참조패턴과 얼마나 비슷한지를 결정법칙에 의해 유사도를 부여한 후 유사도가 가장 높은 패턴을 인식된 패턴으로 결정하는 음성인식 알고리즘 과정이다.

현재 대표적인 음성인식 알고리즘으로는 비선형 시간축을 선형적으로 정규화시킨 패턴 정합방식인 DTW (dynamic time warping), 인간의 신경망을 모델링한 NN(neural network), 벡터의 계열을 양자화한 VQ (vector quantization)[6], 확률적인 방법으로 잘 알려진 HMM이 있으며 이중에서도 HMM은 현재 음성인식 기술의 적용하는 영역의 대표적인 기본 인식 모듈로 적용되고 있다[7].

본 연구에 사용할 음성인식 알고리즘은 확률적 정의를 기반으로 하는 HMM 알고리즘이다. HMM모델은 에르고딕 모델과 Bakis 모델인 Left-to-Right 모델 중 역방향 천이를 할 수 없으며 음성 신호들의 시간적 구조로 모델링하기 쉬운 Bakis 모델을 선택하여 음식 인식에 적용하고자 한다. 시뮬레이션은 기본 음성 패턴을 저장 후 각 조원의 음성을 입력하여 동작 여부를 조사하여 학습을 통하여 기본 패턴의 인식 폭을 확대시킨다. 음성인식 학습과정은 Forward-Backward인 Baum-Welch 알고리즘을 반복 사용하여 단어별 훈련을 하고 각 워드에 대한 가능성 계산을 통해 인식결과를 구한다. 마지막으로 소스의 지속적인 수정을 거쳐 알고리즘 보완 및 음성 감지율(voice detect ratio)을 높이고 인식 속도를 실시간 처리에 가깝게 한다.

실험은 임베디드 시스템에서 32비트 RISC ARM7 CPU를 많이 채택하므로 ARM 칩을 이용한 개발 툴킷에 음성인식 엔진을 포팅하였다. 입력장치는 핸드 프리용 마이크를 사용하고 처리장치는 16비트 코드도 지원하는 ARM7TDMI 코어를 사용한 KS32C50300 평가보드를 사용하고 출력장치는 음성인식 결과를 보여줄 수 있는 LCD 디스플레이를 사용하였다.

사용할 음성 데이터는 PCS 전화기에서의 메뉴 제어 명령을 남성, 여성 각각 15 명의 화자로부터 발성된 40개의 명령어와 10개의 숫자음 음성 데이터를 사용하고 7일 동안 총 2회 발성하여 주위 환경별, 지역별, 연령별로 구분하여 녹취한다. 이때 샘플링 비율은 8KHz, 데이터 크기는 16비트로 설정하고 각 발성에 대한 웨이브 화일로 구분 저장하여 사용하였다.

인식 실험은 제안하는 인식 파라미터 5가지 셋트를 명령어 단어와 숫자음으로 구분하여 실험한 결과 가장 우수한 파라미터 셋트를 설정한 다음 선정 파라미터 셋트를 이용하여 인식을 수행한다. 따라서 다양한 인식 실험 결과들을 비교, 고찰함으로써 제안하는 파라미터 및 실시간 임베디드 음성인식 시스템의 유효성과 타당성을 검증한다.

Ⅱ. 임베디드 음성 인식시스템의 구현

1. 임베디드 음성인식 시스템의 제한 조건

임베디드 시스템에 음성인식 기능을 탑재하려면 순수 음성(PCM 신호)를 가지고 처리할 경우에 메모리 크기가 문제이다. 음성특징 추출, 신호처리기법, 음성인식 기술, 잡음억제 기술 및 표준 데이터베이스 등의 기술 및 DB를 저장할 수 있는 메모리 크기의 한계성에 따라 음성인식 엔진 및 데이터베이스 크기가 우선적으로 크기에 제한을 받게된다.

임베디드 시스템에서 사용되는 메모리인 플래쉬 롬

은 보통 1개의 8M 또는 16Mbyte를 사용하는데 이 메모리는 메인 소프트 프로그램이 저장된다. 이 플래쉬 롬에는 Voice Guide Message, 사용자가 등록한 이름들 및 소스코드 등이 저장된다. SRAM은 보통 1M-2Mbyte를 사용하는데 이 메모리에는 Flag 정보, Call Processing Data, 타이머 데이터 등이 저장된다. EEPROM은 64K-128Kbyte 메모리용량을 사용하고 SN (serial number), NAM(numeric assignment), 전원레벨, 볼륨레벨, 전화번호 등이 저장된다.

실시간 임베디드 음성인식 시스템을 구현하기 위하여 가장 큰 제한 조건은 플래쉬 메모리 크기의 제한되어 있으므로 음성인식 엔진 및 데이터 베이스의 크기가 최대 1.5Mbyte이하가 되어야 한다.

2. 임베디드 음성인식 시스템의 구성

본 연구에 사용된 임베디드 음성 인식 시스템은 <그림 1>과 같다. 입력장치로 사용된 마이크는 최근 많이 사용하고 있는 핸즈 프리용 이어폰 마이크를 사용하였고, 처리장치는 32 비트 프로세서이며 16 비트 코드도 지원하는 ARM7TDMI 코어를 사용한 삼성 KS32C50300를 사용하고, 출력장치로는 음성인식 결과를 보여줄 수 있는 LCD 디스플레이를 사용하였다. CODEC를 통해 입력되는 음성데이터는 훈련 키트 상의 메모리로 저장되며, 위의 음성처리 방법에 따라 특징 벡터가 얻어지고, 플래쉬 롬에 저장되어 있는 단어 HMM세트를 이용하는 음성인식 프로그램에 의해 인식된 단어 인덱스를 문자 저장 맵에 전한다.

시스템 내의 메모리 사용은 프로그램 영역, 샘플 데이터 영역 그리고 특징 벡터영역으로 구분하여 프로그램 영역은 1Mbyte 공간을 사용하였고, 샘플 데이터 (16Kbyte) 및 특징벡터 영역(1.3Kbyte)을 할당하였다.

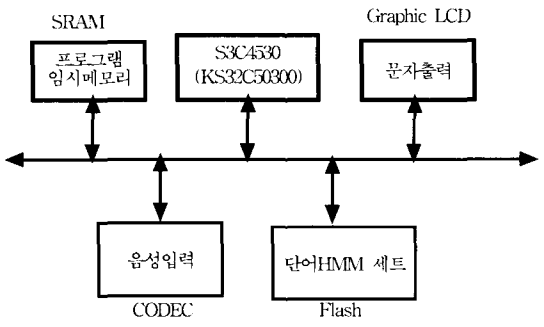

그림 1. 임베디드 음성인식 시스템 구성
Fig. 1. Overview of Embedded Speech Recognition System.

Ⅲ. 임베디드 시스템에서의 음성인식

1. 음성인식

일반적인 음성인식의 경우, 음성입력에 대한 끝점 검출이 이루어지면, 인식기의 입력패턴을 구성하기 위한 특징이 추출되며, 학습으로 만들어진 기준 모델과 패턴 비교, 거리척도 사용 및 확률적 분포 등을 이용하여, 최종적으로 가장 유사한 결과를 찾는다. 본 논문의 음성인식기는 준연속 은닉 마코프 모델 알고리즘을 사용하였고 제안된 특징 파라미터 및 HMM 적용 시에 필요한 인식 조건의 변경에 따라 인식률을 평가한다.

본 논문에서는 유사도 측정 방법으로써 패턴비교에 의한 인식 방법의 하나인 HMM법을 이용하였으며 패턴비교에 의한 음성인식의 일반적인 과정은 음성 파형 추출과정, 패턴 학습과정, 패턴분류 과정 그리고 결정논리 과정으로 나눌 수 있다.

2. 음성 검출 방법의 제안

음성 검출부는 음성인식의 성능에 커다란 영향을 미치는 부분 중에 하나이므로 잘못된 음성 검출은 어떤 음성인식 알고리즘을 사용하느냐에 따라 정도의 차이는 있지만 이후의 모든 과정에 전파되어 전체시스템 성능을 크게 저하시킨다. 따라서 정확한 음성 특징 검출이 요구된다.

본 연구에서는 주변 환경에 영향을 받지 않고 음성 신호에서 추출하기 쉬운 파라미터와 일관적인 알고리즘을 이용해 음성을 검출하였고 실시간으로 음성을 검출하였다. 대부분의 경우 영교차율(zero crossing rate)과 에너지와 같은 계산량이 적은 시간 영역의 파라미터만을 사용해 실시간으로 음성구간을 검출하고 주변 환경의 변화에 따라 음성을 검출하는 파라미터의 문턱값을 적응시키는 방법으로 주변 환경의 영향을 적게 받도록 설계하였다. 실시간 음성 검출이 아닌 경우에는 음성 신호의 주파수 영역에 특징 파라미터를 사용함으로써 좀더 정확히 음성을 검출할 수 있다. 그러나 임베디드 음성인식 시스템에서는 실제로 동작을 위해서는 최소한 음성 검출부 만은 실시간 처리가 이루어져야 한다. 음성이 입력되면, 영교차율과 에너지를 구하고 각 문턱 치에 의한 판단 후 음성 구간을 검출하게 된다. 마이크로폰으로 입력된 음성은 16비트 A/D(analog to digital)컨버터를 사용하여 16비트 디지털 신호로 변화

하였고, 마이크로폰은 현재 시중에서 판매되고 있는, 음성적용 가능한 제품으로 선택하였다.

음성특징의 추출은 마이크입력, 고역강조, 해밍윈도우, 자기상관분석, 선형예측분석의 순서로 수행하며 마이크를 통해 입력되는 음성은 입술로부터 방사 시 -6 dB/oct 정도의 경사로 기울어지고, 이것을 보상하기 위해 고역강조를 필터를 이용한다. 또한, 음성 분석을 위해 해밍창을 이용하여 음성데이터를 일정한 크기로 분할하였다. 이때, 프레임간의 음성데이터 손실을 줄이기 위해 중첩 윈도우를 사용하였으며, 그 크기는 프레임 크기의 $\frac{1}{2}$ 을 사용하였다. N 크기로 분할된 음성데이터는 인식에 사용된 특징벡터를 구하기 위해, 먼저 자기 상관계수를 구하게 된다.

자기 상관계수들은 선형 예측 계수를 구하기 위해, Levinson-Durbin 알고리즘을 사용하였다. 선형 예측 계수들은 본 연구에서 사용하기 위해, 선형 예측 계수보다 더 강인한 특징인 켑스트럼으로 변환된다.

3. 단어 음성 인식 시스템

음성인식 알고리즘은 음성의 시변적인 특성을 잘 모델링 할 수 있어야 한다. 임베디드 음성인식 시스템의 목적과 규모에 따라 여러 가지 알고리즘이 있으나 본 논문에서는 HMM 모델이 음성의 변이성을 확률모델로 이해한 것으로, 음성이 Markov 과정으로 모델링 될 수 있다는 가정 하에 학습과 인식에 사용된다. HMM 모델은 에르고딕 모델과 Bakis 모델이라는 Left-to-Right 모델이 있다. 에르고딕 모델은 모든 천이 a_{ij}는 양의 값을 갖게 되며, 순방향-역방향 모든 상태 천이가 가능하다. Bakis 모델은 역방향 천이를 할 수 없으며 음성신호들의 시간적 구조를 모델링 하기에 좋다. 따라서, 본 논문에서는 Bakis 모델을 선택하고, 음성인식에 적용한다.

Ⅳ. 실험 및 고찰

1. 실험환경 및 기준모델 선정

본 연구에서는 단어 단위 음성인식기의 임베디드 음성인식 시스템을 구현하였다. 직접 ARM 프로세서를 사용하는 훈련 키트에 적용시켜 보았고 먼저 HMM 세트를 구하기 위해 PC 상에서 프로그램을 구현하였다. 이 때, 훈련을 위한 프로그램만 구현하고, 결과적으로 얻어지는 HMM 세트는 ARM 프로세서 상에서 사용할 수 있는 형태로 만들어졌다. 프로그램의 동작은 입력 음성 데이터로부터 얻어낸 특징 벡터 영역에 저장된 특징 벡터들을 사용하여, 플래쉬 롬에 저장되어 있는 훈련된 HMM 세트와의 확률적 관계를 통해 그 결과를 얻게 된다. 최종적으로 얻어지는 인식된 단어 인덱스는 v 개의 단어와의 관계 중 가장 큰 확률 값을 나타내는 단어를 선택하였다.

본 연구에서는 최적화를 하지 않은 즉, 필요한 곳에 인라인 어셈블러코드를 넣지 않은 상태에서 프로그램을 구현하였다. 이 결과는 단지 음성인식에 관련된 부분만을 보이고 있으며, 하드웨어 프로그램에서 C 언어를 사용하고 있다.

인식을 위한 음성데이터는 남, 여 각각 15 명의 화자들로부터 수집되었으며, 주위 환경별, 지역별 및 연령별로 구분되어 녹취하였다.

주위 환경별로 보면 가정, 사무실이 30%, 거리 20%, 공공장소 20 %, 그리고 자동차가 30% 의 비율로 구성한다. 지역별로는 서울, 경기가 32 %, 경상도 20%, 전라도 16%, 및 충청도 16% 로 구성한다. 연령별로는 12~17세 ; 20%, 18~28세 ; 20%, 29~39세 ; 30% 그리고 40~60세 ; 20%로 구성한다. PCS 폰용 이어폰을 사용하여 노트북 컴퓨터로 녹음한다. 이 때, 샘플링 율은 8KHz, 데이터 크기는 16비트로 설정하고, 각 발성에 대한 웨이브 화일을 구분 기록하여 본 연구에 사용하였다.

인식 실험에 사용되는 단어 목록은 이동전화의 작동에 필요한 주 기능 및 제어단어 등, 총 40 개의 단어 목록과 10개의 숫자를 사용하여 음성데이터를 얻었다.

음성인식 시스템에서 화자독립, 화자종속의 단어 수와 데이터베이스 크기는 비례적인 함수관계이다.

2. 음성검출 파라미터 셋트

본 논문의 음성인식 시스템에서 사용한 데이터베이스인 <표 1>은 화자 독립(40단어)은 150 Kbyte, 숫자음(10단어)은 20Kbyte, 화자종속(20단어)은 20Kbyte, 상수데이터는 10Kbyte를 포함한 전체 데이터베이스 크기는 200Kbyte이고 코드 크기는 180Kbyte 이고 RAM 크기는 50Kbyte보다 적게 사용된다.

임베디드 시스템에서 요구되는 메모리 크기는 보통 최대한 플래쉬 메모리 1.5M byte이하를 요구한다.

본 논문에서는 이 조건에 만족하기 위하여 명령어 40 단어와 숫자음 10단어를 선정하였다.

표 1. 단어 선정에 따른 데이터 베이스 크기
Table 1. DB size requirements for Vocabulary Selection.

시스템	DB size
SI(40단어)	150 Kbyte
디지트(10단어)	20 Kbyte
SD(20단어)	20 Kbyte
Constant Data	10 Kbyte
Total	200 Kbyte

음성은 일정구간으로 나누어 주파수 특성을 구하고 좋은 특성을 검출하여 성능을 평가하는 방법에 여러 가지 변수가 있다. 구간 경계 부분에는 잡음이 발생하는 경우가 종종 있게 되므로 각 프레임의 양 끝점에서의 영향을 최소화하기 위하여 본 논문에서는 이러한 경우에 대하여 실험적으로 파라미터 값을 설정하였다. 파라미터 값의 선정은 기존의 파라미터의 장점은 수용하고 단점을 실험 가중치를 설정하여 그 중 좋은 파라미터를 5셋트를 선정하였다. 이 셋트는 단어, 숫자와 환경별 인식율을 실험적으로 반복하여 표준단어 및 숫자음에 대한 화자종속, 독립, 원격화자, 차량환경 등에 따라 제안 파라미터 세트가 어느 경우에 가장 인식률을 극대화하는지를 실험적으로 검증한다. 파라미터 세트의 기준점은 시작 검출 활성화, 시작점 검출 절대 경계값, 시작점 검출 비교 경계값, 끝점 검출 활성화, 끝점 검출 시간, 끝점 검출 카운터 수, 최대 시간, Nut수, 거절 오차 허용도 및 확인 오차 허용도를 선택하여 파라미터 셋트를 <표 2> 처럼 구성하였다.

제안하는 파라미터 5가지 셋트 중에서 우수한 인식 결과를 선정하기 위하여 화자 독립 명령어 인식의 경우(실험 단어 수; 32,769단어, 사용단어 40개) 에는 마이크를 사용한 인식 시스템의 명령어 인식 실험과 미지어를 사용한 화자독립 명령어의 경우의 결과에는 5가지 셋트 중 셋트 5, 셋트 1의 결과를 사용하였다. 화자 독립 숫자음 인식 실험 경우(총 실험 단어 수; 32,768 디지트, 사용 단어 수; 13개)에는 마이크를 사용한 숫자음 인식 실험과 임의 선택 숫자에 대한 결과에는 5가지 셋트 중 셋트 2 와 셋트 3의 결과를 사용하였다.

화자종속 명령어에 대한 인식 실험의 경우(총20단어를 사용)에는 사무실 및 실외 환경에 따른 인식률의 변화가 다름으로 제안하는 파라미터 5개 중 셋트4, 와 셋트 5의 사용하여 실내외 환경 명령어와 미지어를 사용한 화자종속 명령어 인식에 사용하였다.

표 2. 파라미터 셋트
Table 2. Parameter Set.

T-kit	세트1	세트2	세트3	세트4	세트5
시작점 검출 활성화	1	1	1	1	1
시작점 검출 절대 경계값	20,000	18,000	18,000	20,000	22,000
시작점 검출 비교 경계값	1,560	1,560	1,560	1,560	1,560
끝점 검출 활성화	1	1	1	1	1
끝점 검출 시간(분)	30	20	20	30	30
끝점 검출 수	10	8	5	10	5
최대시간	90	60	60	90	60
Nut 수	3	3	3	3	3
거부 오차허용도	0	1	0	1	0
확인 오차허용도	0	1	0	1	0

화자독립 원격 화자 명령어 인식 실험의 경우(총 인식 단어 수; 32,770 단어, 사용단어 수; 40개)에는 제안 파라미터 셋트 1과 세트 5를 사용하여 원격 화자 명령어 와 미지어를 사용한 원격 화자 명령어에 적용하였다.

화자 독립 원격 화자 숫자음 인식 실험의 경우(총 인식 단어 수; 32,768 단어, 사용 단어 수; 13)에는 제안 파라미터 셋트 2, 3을 사용하여 원격화자 숫자음 인식과 임의 선택 숫자음 인식을 실험하였다.

<표 2>에 제안한 파라미터를 사용하여 여러 가지 경우의 실험을 한 결과 명령어의 경우에는 제안하는 파라미터 셋트 1을 사용한 결과에서 가장 우수한 인식 결과를 알 수 있고 숫자음의 경우에는 제안하는 파라미터 셋트 3을 사용한 결과에서 가장 우수한 인식 결과를 나타내었다.

3. 제안하는 파라미터 선정을 위한 인식 실험

<표 3>의 최적의 인식률을 위한 제안 파라미터 선정 결과에 의하여 화자 독립 명령어 와 화자 독립 원격 명령어의 경우에는 파라미터 셋트 1에서 좋은 인식 결과가 나왔다. 그리고 화자 독립 숫자음과 화자 독립 원격 숫자음의 경우에는 파라미터 셋트 3의 경우에 가장 좋은 인식 결과가 나왔다. 마지막으로 실 내외 및 차량 환경에서의 화자 종속 명령어에서는 파라미터 셋트 4

표 3. 최적 인식률을 위한 파라미터 선정 결과
Table 3. Parameter Sets for Optimized Speech Recognition Rates.

인식항목	구 분 항 목		최 적 인식율 실험	선정 파라 메터
화자독립 명령어	마이크 사용의 경우		LAB 6	Set 1
	미지어 사용		LAB 6	Set 1
화자독립 숫자음	마이크사용		LAB 6	Set 3
	임의선택 숫자음		LAB 6	Set 3
화자종속 명령어	환경	사무실 환경	SD-Digit 4	Set 4
		실외 환경	SD-Digit 8	Set 4
	미지어사용	사무실 환경	SD-Digit 12	Set 4
		실외환경	SD-Digit 16	Set 4
화자독립 원격명령어	원격화자 명령어		LAB 25	Set 1
	미지어 사용		LAB 25	Set 1
화자독립 원격 숫자음	원격화자 숫자음		LAB 25	Set 3
	임의선택 숫자		LAB 25	Set 3
화자종속 차량환경 명령어	차량 화자		SD-4	Set 4
	미지어사용 차량화자		SD-4	Set 4

의 경우가 가장 좋은 인식 결과가 나왔다.

명령어에 대한 최적의 인식률을 나타낸 파라미터 셋트 1을 이용하여 40개 단어를 명령어 인식기의 인식률은 96%, 숫자음에 대한 최적의 인식률을 나타낸 파라미터 셋트 3을 이용한 13개 단어의 숫자음 인식기의 인식률은 94%을 나타내고 음성인식엔진의 인식률은 95%로 나타났다.

(1) 선정 파라미터 셋트 1을 이용한 명령어 인식결과

<표 4>는 화자독립 명령어 및 화자독립 원격 명령어에 대한 최적의 인식률을 나타낸 파라미터 셋트 1을 이용하고 사용 단어 수(32,770)와 사용 명령어(40)에 대한 인식 결과이다. 이 40단어의 전체적인 인식률은 96%가 나왔고 3단어(주소록, 교통정보, 계산기)는 90% 이하의 결과가 나왔다. '주소록'은 거부 비율와 대체에러 비율가 높기 때문에 인식률이 떨어짐을 알 수 있다. '교통정보' 와 '계산기'에선 대체에러 비율이 높기 때문에 인식률이 떨어짐을 알 수 있다. <표 5>는 40단어 전체에 대한 인식률 및 오차율을 숫자로 나타낸 것이다.

표 4. 파라미터 셋트 1을 이 명령어 음성인식 결과
Table 4. Speech Recognition Rate for command using Parameter Set1.

단어 ID	단어 이름	인식율
0	전 화	91.45
1	지 움	100.00
2	통 화	94.29
3	메 뉴	94.29
4	다 음	100.00
5	앞 으 로	100.00
6	확 인	100.00
7	선 택	94.12
8	취 소	97.06
9	예	100.00
10	아 니 오	97.14
11	종 료	91.43
12	수 신 함	91.43
13	음성사서함	100.00
14	편지쓰기	97.14
15	등 록	91.43
16	저 장	94.29
17	주 소 록	88.57
18	전화번호부	97.14
19	음성다이얼	100.00
20	벨 소 리	91.43
21	진 동	100.00
22	알 림	100.00
23	사 전	100.00
24	계 산 기	88.57
25	할 일	100.00
26	메 모	97.14
27	인 터 넷	91.43
28	책 갈 피	97.14
29	촬 영	100.00
30	내 위 치	97.14
32	뉴 스	91.43
33	교통정보	88.57
34	날 씨	100.00
35	읽 기	94.12
36	이 메 일	97.14
38	재 발 신	90.00
39	매너모드	95.00
40	스 케 줄	90.00
41	주 식	94.74

표 5. 파라미터 셋트3을 이용한 숫자음 인식 결과
Table 5. Numeral Speech Recognition Rate for Using Parameter Set.

단어 ID	단어이름	인식율
0	영	94.29
1	공	91.43
2	일	91.43
3	이	93.94
4	삼	97.14
5	사	97.14
6	오	88.57
7	육	91.18
8	칠	91.43
9	팔	94.29
10	구	94.32
22	지움	94.12
23	확인	100.00

(2) 선정 파라미터 셋트 3을 이용한 숫자음 인식결과

화자독립 숫자음 및 화자독립 원격 숫자음에 대한 최적의 인식률을 나타낸 파라미터 셋트 3을 이용하고 사용 단어 수(32,768)와 사용 명령어수(13)에 대한 인식 결과이다. 13단어의 전체적인 인식률은 94%가 나왔고 한 단어(숫자음 ; 오)는 90%이하의 결과가 나왔다. 숫자음 "오" 는 대체에러 비율이 높기 때문에 인식률이 떨어짐을 알 수 있다. <표 5>는 파라미터 셋트 3의 숫자음의 인식율에 가장 좋은 파라미터 결과가 나타남에 따라 파라미터 셋트 3을 이용한 인식 결과이다.

V. 결 론

본 논문은 한국어 음성인식의 특징추출의 파라미터 셋트를 제안하고 실시간으로 사용하는 임베디드 음성 인식 시스템의 설계 및 구현에 관한 연구이다.

본 연구 개발의 내용 및 범위는 우선 단어 기반의 임베디드 시스템용 한국어 음성 데이터베이스를 구축하고 임베디드용 소규모 음성인식 엔진개발 및 임베디드 ARM 7 평가보드에 장착하여 시스템을 구현한 것이다.

본 연구의 DB 결과는 단어 기반의 임베디드 시스템용 한국어 음성 데이터베이스를 32,770단어(사용명령어; 40단어), 32768 독립 단일 디지트(사용 숫자음 ; 10숫자) 분량의 음성 데이터베이스를 구축하였다.

인식실험 항목은 화자 독립, 환경별 화자 종속, 화자 독립 원격, 화자 종속 차량 환경을 알맞은 파라미터 5가지 파라미터 셋트를 실험하여 결과를 얻었다. 파라미터 5가지 셋트 중에서 명령어의 경우는 파라미터 셋트 1 과 숫자음의 경우는 파라미터 셋트 3의 선정 파라미터를 실험에 의하여 가장 인식율이 좋은 파라미터를 결정했다.

종래의 음성 인식 엔진을 임베디드 시스템에 적용시킬 때 시스템 메모리 크기에 제한되므로 인식률, 엔진 및 DB의 경량화 등의 과제가 선행되어야 한다. 본 연구에서는 코드 북, 코드크기 등을 포함하여 음성인식 프로그램의 메모리 요구량이 1.331 Kbyte 이고 응답시간은 500ms가 되므로 실시간 임베디드 음성인식 시스템을 적용하여 구현하기에 최적이다.

임베디드용 실시간 음성인식 엔진의 인식 결과는 인식 엔진의 인식률은 95%, 음성 명령어 인식기의 인식율은 96%, 음성 숫자음 인식기는 94%의 인식 성능을 갖는다.

본 연구 개발의 기대효과는 문서작성, 정보검색, 예약, 호출 및 자동 메뉴 다이얼링 등 편리성을 고급형 임베디드 시스템 제품을 개발할 수 있다. 한국어 임베디드용 데이터베이스 개발로 임베디드 단어기반 DB을 활용할 수 있고 엔진의 라이브러리화를 통하여 임베디드 시스템에 적용시 이식성을 좋게 할 수 있다.

참 고 문 헌

[1] 김순협 "음성인식 기술 현황 및 연구 동향", 2000년도 한국음향학회 학술발표대회 논문집, Vol. 19, No. 2(s), pp. 25~28, 2000.

[2] J. Mariani, "Recent advances in speech processing," Proc. of ICASSP, pp. 429~440, 1989.

[3] Rabiner, L. R. "Application of Voice Processing to Telecommunications". Proceeding of the IEEE, Vol. 82, No. 2, pp. 199~228, 1994.

[4] D.van Compernolle, "Speech Recognition in the Car From Phone Dialing to Car Navigation", Proceedings of EURO SPEECH'97. vol.5, pp. 2431~2433, 1997.

[5] 남상엽, 이상원, 박인정 "임베디드 시스템을 위한

소형 음성인식 시스템 구현에 관한 연구". 대한전자공학회 논문집 Vol. 37-TE-6-9, No. 2, pp. 152~158, 2001.

[6] Allen Gersho, Robert M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publisher, 1992.

[7] X. D. Huang, Y.Ariki, M. A. Jack, "Hidden Markov Models for Speech Recognition", Edinburgh Information Technology Series, 1990.

[8] J. A. Haigh, J. S. Mason, "Robust Voice Detection using Cepstral Features". IEEE TENCON-93, pp. 321~324, 1993.

저 자 소 개

南 相 曄(正會員)

1982년 2월 : 단국대학교 공과대학 전자공학과(학사). 1984년 2월 : 단국대학교 대학원 전자공학과(석사). 1984년 1월~1992년 1월 : 삼성종합기술원 정보시스템연구소 주임연구원. 1992년 1월~1998년 3월 : 모토로라반도체통신㈜ 기술연구소 차장. 1998년 3월~2002년 현재 : 경문대 정보통신과 교수. 2002년 2월 : 단국대학교 대학원 전자공학과(박사). <주관심분야 : 멀티미디어 신호처리, 컴퓨터통신, 마이크로프로세서, DVI, CD-R/W, D-TV, STB, 멀티미디어콘텐츠, 전자상거래, 3D, DSP, 영상/음성인식, 인터넷방송, 디지털TV>

全 銀 姬(正會員)

1987년 4월~1998년 8월 : ㈜삼성전자 주임. 1998년 8월~1999년 8월 : ㈜멀티미디어라인 과장. 1999년 8월~2002년 3월 : ㈜DRI 부장. 2002년 2월 : 한국방송통신대학 경영학과(학사). 2002년 12월 : 단국대학교 대학원 경영학과 재학(석사) (MIS 전공). <주관심분야 : 멀티미디어 신호처리, 컴퓨터통신, 마이크로프로세서, 이동통신, 멀티미디어 콘텐츠, 전자상거래, 영상/음성인식, 인터넷방송, 디지털TV, 경영정보(MIS), 모바일 비즈니스>

朴 仁 政(正會員) 第26卷 第7號 電子工學會誌 參照