DSP를 이용한 자동차 소음에 강인한 음성인식기 구현

Implementation of a Robust Speech Recognizer in Noisy Car Environment Using a DSP

정 익 주*
Ik-Joo Chung

ABSTRACT

In this paper, we implemented a robust speech recognizer using the TMS320VC33 DSP. For this implementation, we had built speech and noise database suitable for the recognizer using spectral subtraction method for noise removal. The recognizer has an explicit structure in aspect that a speech signal is enhanced through spectral subtraction before endpoints detection and feature extraction. This helps make the operation of the recognizer clear and build HMM models which give minimum model-mismatch. Since the recognizer was developed for the purpose of controlling car facilities and voice dialing, it has two recognition engines, speaker independent one for controlling car facilities and speaker dependent one for voice dialing. We adopted a conventional DTW algorithm for the latter and a continuous HMM for the former. Though various off-line recognition test, we made a selection of optimal conditions of several recognition parameters for a resource-limited embedded recognizer, which led to HMM models of the three mixtures per state. The car noise added speech database is enhanced using spectral subtraction before HMM parameter estimation for reducing model-mismatch caused by nonlinear distortion from spectral subtraction. The hardware module developed includes a microcontroller for host interface which processes the protocol between the DSP and a host.

Keywords: Robust speech recognizer, noisy car environment

1. 서 론

자동차 환경에서의 음성인식은 운전자가 운전에 집중을 하면서 동시에 최소한의 노력으로 자동차의 여러 편의 시설을 조작하는 것을 가능하게 하기 때문에 음성인식 응용 분야 중에서 가장 각광받는 응용 분야 중에 하나이다[1][2]. 더구나 최근 들어, 휴대폰, 네비게이션과 같이 운전 중에 조작해야할 장치들이 많아지면서 자동차 환경에서의 음성인식 기술은 점차 필수적인 기술로 자리 잡아가고 있다. 네비게이션의 경우, 목적지를 음성으로 입력하는 대용량 음성인식 기술이 시도되고 있으며, 휴대폰의 경우, 이미 휴대폰에 내장되어 있는 음성 인식기를 핸즈프리의 형태로 사용하고 있

* 강원대학교 전기전자공학부

다. 일부 자동차 제조사에서는 자동차의 편의시설을 음성으로 제어할 수 있는 음성인식 기능을 탑재하여 판매하고 있다.

자동차용 음성 인식기에 적용되는 음성인식 기술을 분류해 보면, 자동차에 장착된 단말기에서 인식이 수행되는 임베디드형 음성인식 방식과 전화망을 통한 원격지의 서버에서 인식이 수행되는 서버형 음성인식으로 나누어진다. 임베디드형 음성인식 방식의 경우는 자동차 편의 장치의 조작이나 폰북에 저장된 인물이나 장소로의 음성 다이얼링 정도를 수행할 수 있는 반면, 서버용 음성인식의 경우에는 네비게이션의 목적지나 음성인식 교환 서비스와 같은 대용량 음성인식 기술이 이용되고 있다. 한편, 최근 들어 네비게이션의 목적지를 인식하는 대용량 임베디드 인식기술이 시도되고 있으며, 반도체 기술의 발달로 머지 않아 상용화 될 것으로 예상된다. 그러나 현재의 일반적인 임베디드 음성 인식기는 사용할 수 있는 리소스의 한계로 인하여 중소요량의 인식기술이 적용되고 있으며, 인식 단어 수가 제한되어 있기 때문에 응용에 따라서, 가변어휘 기술 또는 고정어휘 기술이 모두 적용되고 있다. 고립단어 인식 방식의 경우, 특정 인식기에 최적화된 음성 데이터베이스를 구축할 수 있기 때문에 어휘를 변경할 수 없다는 단점에도 불구하고, 높은 인식률로 인하여 명령어를 변경할 필요가 없는 응용에 흔히 사용된다.

본 연구에서는 자동차의 편이 장치를 음성으로 제어하고 음성 다이얼링을 수행하는 음성 인식기 개발이 목표이기 때문에, 편이 장치 제어의 경우는 자동차 잡음에 강인한 실용적인 음성 인식기를 구현하기 위하여 화자 독립 고정어휘 방식의 인식 기술을 적용하였다. 음성 다이얼링의 경우는 화자 종속 인식을 통해 이루어진다. 이를 위하여 개발된 인식기는 화자 독립 및 화자 종속을 위한 이중 모드(dual-mode) 인식기를 내장하고 있다. 한편, 잡음에 강인한 인식기를 구현하기 위하여 음성 분석 단계에서 주파수 차감 방식을 적용하지 않고, 전처리 단계에서 적용함으로써 잡음 제거된 신호로부터 끝점 검출을 수행하였다.

2. 자동차 잡음을 고려한 음성 데이터베이스 구축

고정 어휘 방식의 화자 독립 인식기는 융통성이 없다는 단점이 있기는 하지만, 타 방식의 인식기에 비하여 높은 인식률을 얻을 수 있는 장점이 있다. 특히, 많은 단어의 인식보다는 적은 수의 단어라도 잡음 환경 하에서 신뢰성 있는 인식기를 구현해야 하는 상용 인식기 개발에서는 비교적 선호되는 방법이다. 그럼에도 불구하고 선정된 명령어들을 위한 음성 데이터베이터를 구축하는 것은 상당히 부담스러운 과정이다. 본 개발에서는 네비게이션 명령어, 자동차 편이 장치를 위한 명령어, 단일 숫자음 등을 포함하여 167 개의 명령어를 선정하여 음성 데이터베이스를 구축하였다. 음성 데이터베이스 구축을 위한 음성 녹음은 전문 업체에 의뢰하였으며 그 내용은 다음과 같다.

- ■ 각 명령어에 대하여 남녀 각각 250 명이 두 번씩 발음
- ■ 화자의 구성은 20대 20%, 30대 40%, 40대 40%로 되어 있음
- ■ 지역별 구성은 경인, 충청 53%, 호남 29%, 영남 18%로 구성되어 있음
- ■ 발성 환경은 사무실 환경 53%, 자동차 환경 47%로 구성

자동차용 음성인식을 위한 음성 데이터베이스 구축의 경우, 실제로 달리는 자동차 안에서 다양한 소음을 발생시키면서 녹음을 하는 것이 이상적이겠지만, 현실적으로 많은 양의 데이터를 이런 방식으로 구축하는 것은 매우 어렵다. 뿐만 아니라, 화자와 소음의 조합이 고정되기 때문에 다양한 화자와 다양한 종류의 소음 조합을 만들어 내는데도 한계가 있다. 따라서 많은 경우, 실제 화자의 음성과 소음을 별도로 녹음하여 음성과 소음을 인위적으로 더함으로써 음성 데이터베이스를 구축하게 된다. 이럴 경우, 다양한 종류의 소음, 예를 들어 여러 속도에서의 자동차 소음, 도로 포장 형태에 따른 소음, 터널 소음 등을 별도로 채집하여 미리 구축된 음성에 더하여 원하는 소음이 포함된 음성 데이터베이스를 구축할 수 있다. 단, 녹음 시에 자동차의 room acoustics를 얻기 위하여 자동차의 시동을 끈 상태에서 자동차 실내에서 녹음을 하였으며, 사무실 환경에서 녹음할 때에도 자동차 실내와 유사한 공간을 만들어서 사용하였다. 자동차 소음은 여러 종류의 소음을 녹음하기 위하여 다양한 조건의 실차 환경에서 집적 채집하였다. 본 개발은 임베디드 형태의 음성 인식기를 개발하는 것이므로 개발될 하드웨어의 front-end(예를 들어 코덱, signal conditioning 회로)와 동일한 frond-end를 가지는 음성 데이터 획득 장치를 제작하여 이를 통해 음성 및 잡음을 녹음하였다. 이는 훈련용 음성 데이터와 실제 인식 시의 입력 데이터의 차이에서 발생하는 모델 불일치(mismatch) 현상을 가능한 최소화하기 위함이다[4].

녹음된 음성 데이터베이스는 수작업 끝점 검출을 포함하는 편집 과정이 포함되어 있지 않다. 따라서 녹음된 음성은 음성 양끝에 상당량의 묵음을 포함하고 있다. 음성 데이터베이스 구축은 편집 과정을 포함하는 것이 일반적이나, 본 개발에서는 실제 인식기에서 사용되는 끝점 검출 알고리즘을 이용하여 자동 편집하는 방법을 채택하였다. 잡음이 별 문제가 되지 않는 경우에는 수작업을 이용한 끝점 검출이 가장 좋은 방법일 수 있으나, 잡음 환경에서는 실제 인식기의 끝점 검출기의 결과가 수작업을 통한 끝점 검출 결과와 차이가 날수 있으므로, 훈련에 사용되는 음성 데이터베이스의 경우에도 실제 인식기에서 사용되는 동일한 끝점 검출기를 이용하여 자동 편집을 수행하였다. 즉, 일단 편집이 되지 않은 음성 데이터베이스에 별도도 채집한 다양한 자동차 소음을 더하여 소음이 섞인 음성 데이터베이스를 만든다. 그런 다음 실제 인식기에서 사용한 끝점 검출기를 오프라인 형태의 프로그램으로 변환하여 이를 이용하여 자동 편집하였다.

한편, 상당량의 음성 데이터를 구축할 경우, 일차적으로 잘못 녹음된 데이터를 걸러내게 되지만, 그럼에도 불구하고 훈련용으로 사용되기에는 적절하지 않은 데이터를 포함할 수 있다. 명확히 잘못 녹음된 것이 아니라 하더라도 정상적인 발성에서 많이 벗어나게 되면 이런 데이터는 훈련에 포함하지 않는 것이 바람직하다. 하지만, 이런 데이터들은 단순 검증 작업에서는 제대로 걸러지지 않다. 정상적인 음성 데이터의 판별 기준으로 발성의 크기와 발성 속도를 이용하였다. 각 명령어는 500명이 두 번씩 발음하였기 때문에 한 명령어 당 1000 개의 음성 데이터가 존재하는데 각각의 최대 진폭을 구하여 상위 2%와 하위 2%를 훈련에서 제외하였다. 즉, 과도하게 크게 발음되거나 작게 발음된 음성 데이터는 훈련에서 제외되었다. 발성 속도의 경우도 잡음을 섞지 않은 상태에서 끝점 검출을 하여 발성의 길이를 구하고 가장 길게 발음된 상위 2%와 가장 짧게 발음된 하위 2%를 훈련에서 제외하였다.

자동차 잡음은 아스팔트 도로와 시멘트 도로에서 훈련 및 테스트를 위하여 다양한 속도에서 채집하였으며 이들 중 10 km/h, 60 km/h, 100 km/h 속도에서 채집한 잡음을 음성 데이터에 더하여

잡음이 포함된 모델 훈련용 음성 데이터베이스를 구축하였다. 따라서 훈련용 음성 데이터베이스의 크기는 잡음이 포함되지 않은 것에 비하여 3 배로 커진다.

3. 음성인식기 구조

다음은 개발된 음성인식기의 구조를 나타내는 블록도이다.

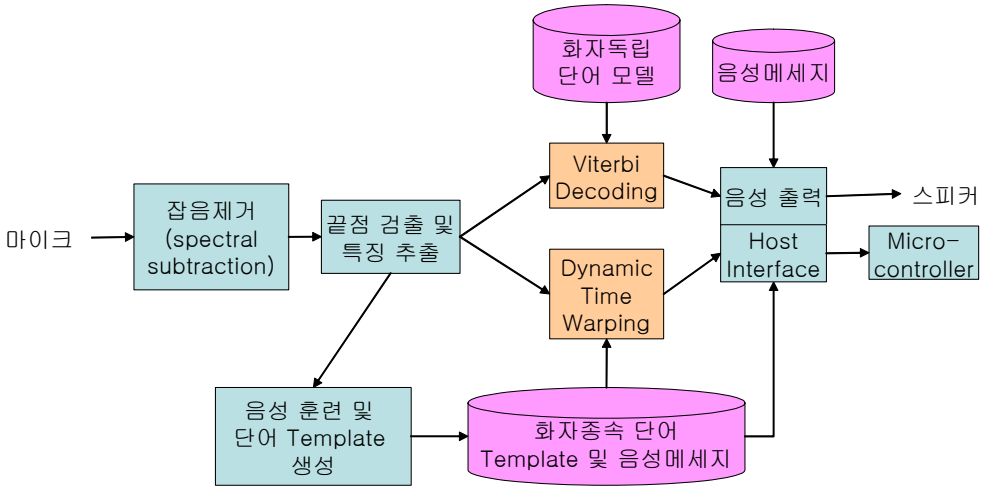

그림 1. 음성 인식기의 구조

화자 독립과 화자 종속 두 개의 음성 인식기를 포함하고 있는데, 두 인식기가 한 도메인에서 인식되는 것은 아니다. <그림 2>와 같이 기본적으로 화자 독립 모드에서 동작하다가 "전화"라는 단어를 인식하면 화자 종속 모드로 전환되고, 사용자가 미리 훈련해 놓은 단어(인명이나 상호)를 인식한 후 자동으로 화자 독립 모드로 돌아오게 된다.

음성이 입력되면 우선 주파수 차감 방식을 이용하여 잡음을 제거한다[3]. 일반적으로 FFT를 기반으로 하는 특징 벡터 추출의 경우, 주파수 차감 방식은 특징 벡터 추출 과정에서 함께 수행되는데, 본 개발에서는 <그림 3>과 같이 입력된 음성 $x(n)$을 FFT 한 후 잡음 성분을 차감하고 IFFT을 수행하여 다시 시간 영역으로 전환한 후, overlap & add 방식으로 잡음이 제거된 음성 파형 $x'(n)$을 구하고, 이를 이용하여 끝점 검출과 특징벡터를 추출한다.

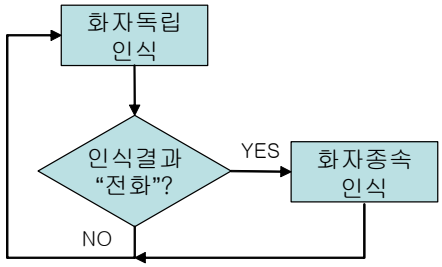

그림 2. 인식기 동작 모드

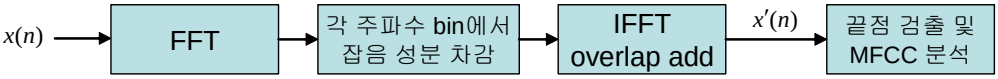

그림 3. 음성인식을 위한 전처리 단계

<그림 3>과 같이 주파수 차감 방식을 이용하여 잡음이 제거된 음성 파형을 이용하는 경우는 잡음이 제거된 상태에서 끝점을 검출하기 때문에 보다 정확한 끝점 검출이 가능하며 무엇보다도 잡음 제거와 인식 알고리즘이 독립되어 있기 때문에 그 구조가 명료하다는 장점이 있으나, 주파수 차감과 MFCC를 계산하는 과정에서 FFT→IFFT→FFT와 같이 3 번의 FFT 관련 연산을 필요로 하기 때문에 연산량이 증가된다는 단점이 있다. 특징 벡터 추출은 24 차로 분석한 후 12 개의 MFCC 계수 및 delta MFCC 계수를 추출하였다. A/D 변환을 위한 샘플링 주파수는 16 KHz를 사용하였으며, MFCC 분석 시, Mel 필터 뱅크에서 200 Hz 이하와 6 KHz 이상에 해당하는 대역의 에너지 값을 0으로 대치하였다. 이는 음성의 주파수 정보의 큰 희생 없이, 자동차의 저주파 잡음과 고속 주행 시 발행하는 바람소리와 같은 고주파 잡음의 영향을 효과적으로 제거하기 위함이다. 한편, 화자 종속 모드에서는 delta MFCC는 사용하지 않기 때문에 이를 추출하지 않는다.

주파수 차감 방식은 잡음이 많을 경우, 잡음 제거 과정에서 원음성에 비선형적인 왜곡이 발생하게 된다. 따라서 인식 단계에서만 주파수 차감 방식으로 잡음이 제거된 신호를 이용할 경우, 잡음이 제거되기는 하지만, 음성의 왜곡으로 인하여 인식률이 저조해지게 된다. 이는 훈련 시와 인식 시의 조건이 다른데서 기인하는 모델 불일치 때문인데, 이를 고려하기 위하여 음성 모델 훈련 시에도 인식 시와 동일하게 주파수 차감된 신호를 이용하여 모델을 생성하였다.

화자 종속 인식기는 통상적인 DTW 알고리즘을 이용하였고, 화자 독립 인식기는 연속분포 HMM 알고리즘을 이용하였다. 임베디드 인식기의 경우, 연산 성능이나 가용한 메모리의 제한이 있기 때문에, 오프라인 인식 실험을 통하여 적절한 인식률을 얻을 수 있는 조건을 구하였다. <표 1>은 한 개의 mixture를 사용한 base-line 인식기의 오프라인 성능이다. 실험은 녹음한 음성 데이터 중에 단독 숫자음을 포함하여 실제 개발에 사용한 84개의 명령어에 대한 인식률이다. base-line 인식기는 주파수 차감 방법을 사용하지 않기 때문에 잡음이 섞이지 않은 음성 데이터베이스를 이용하여 훈련하였다. <표 2>는 주파수 차감 방법을 이용하여 잡음을 제거한 인식기의 오프라인 성능이다.

표 1. base-line 인식기 인식률

자동차 속도	인식률
10 km/h	97.01%
30 km/h	89.32%
60 km/h	68.37%
80 km/h	41.72%
100 km/h	33.58%

표 2. 잡음 처리된 인식기의 인식 성능

자동차 속도	**mixture 1개** (10km/h, 60km/h)	**mixture 2개** (10km/h, 60km/h, 100km/h)	**mixture 3개** (10km/h, 60km/h, 100km/h)	**mixture 3개(*)** (10km/h, 60km/h, 100km/h)
10 km/h	98.80%	98.56%	98.78%	98.78%
30 km/h	96.58%	97.61%	97.79%	97.72%
60 km/h	92.71%	92.93%	93.76%	93.89%
80 km/h	88.65%	91.89%	92.51%	92.35%
100 km/h	84.21%	88.57%	89.90%	89.93%

훈련용 데이터는 10 km/h, 60 km/h, 100 km/h 세 종류의 잡음 데이터를 이용하여 훈련한 것이다. (단, **mixture 1**의 경우는 10km/h, 60km/h 두 종류의 잡음 데이터를 이용하였다.)

인식기가 다양한 소음에 대하여 좋은 성능을 보이기 위해서는 그 밖의 속도에 해당하는 소음들도 추가하는 것이 좋겠지만, 현실적으로 모든 속도의 소음을 포함할 수 없기 때문에 위의 세 속도가 저속, 중속, 고속을 대표한다고 가정하고 선택하였다. 한편, 위의 결과에서 **mixture 3(*)**의 경우는 10 km/h, 60 km/h, 100 km/h 세 종류의 잡음데이터에 대하여 각각 스테이트 당 한 개의 mitxure를 사용하여 한 명령어 당 3 개의 모델을 사용한 경우인데, 3 개의 mixture를 사용하여 하나의 모델을 사용하는 **mixture 3**과 비교할 때 10 km/h, 60 km/h, 100 km/h 잡음에 해당하는 인식률은 **mixture 3**의 경우보다 약간 좋지만, 그 외의 속도의 잡음에 해당하는 인식률의 경우는 미미한 차이이기는 하지만 **mixture 3**보다 약간 낮았다. 즉, 3 개의 mixture를 사용한 단일 모델을 사용했을 경우, 비록 10 km/h, 60 km/h, 100 km/h에 해당하는 잡음데이터를 이용하여 훈련하였지만 그 외의 속도에 해당하는 잡음에 대해서도 어느 정도 보간 하는 능력이 있다고 할 수 있다. 한편, 각각 하나의 mixture를 사용하여 한 명령어 당 3개의 모델을 사용한 경우를 고려한 이유는 이 방식이 출력 확률을 계산할 때, 전이 확률 연산이 증가되고, 메모리 사용도 증가되기는 하지만, mixture가 하나이기 때문에 log 덧셈 연산이 불필요하게 되고, 따라서 전체적인 연산량의 증가는 크지 않기 때문이다. 그러나 **mixture 3**의 경우가 전반적으로 여러 종류의 자동차 소음에 대하여 우수한 인식률을 보였고, 메모리 사용량을 고려하여 실제 구현에서는 3 개의 mixture로 구성된 단일 모델을 사용하였다.

4. DSP를 이용한 실시간 구현

실시간 구현을 위하여 Texas Instruments 사의 부동소수점 DSP인 TMS320VC33 DSP를 이용하였다[5]. 다음은 구현한 음성인식 하드웨어의 구성도이다.

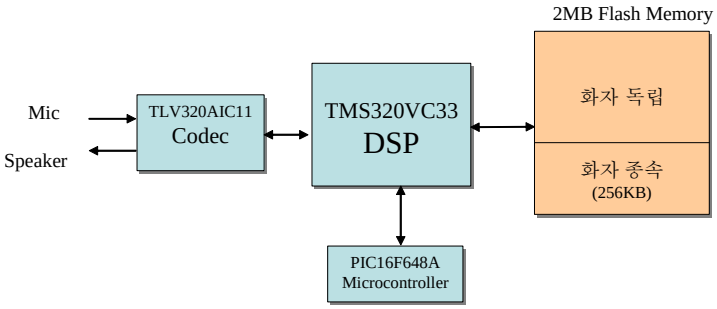

그림 4. 음성인식 하드웨어의 구성도

하드웨어는 DSP, 코덱, 플래쉬 메모리, mircocontroller로 구성된다. 2 Mbyte 용량의 플래쉬 메모리는 화자 독립 모델과 인식 종료 후, 결과를 음성으로 출력하기 위한 음성 메세지를 저장하는 ROM의 역할과 화자 종속 인식기의 사용자 훈련 데이터를 저장하기 위한 사용자 메모리 공간으로 사용된다. TMS320VC33 DSP는 134 Kbyte의 SRAM을 내장하고 있기 때문 외장 SRAM은 사용하지 않았다. 일반적으로 DSP는 연산 능력에 초점이 맞추어져 있기 때문에 제어나 통신과 같은 기능이 미비한 편이다. 자동차의 타 장치와의 통신 및 프로토콜 처리를 위하여 PIC16F648A mircocontroller을 추가로 사용하였다. 한편, 사용된 DSP는 최대 60 MHz(60 MIPS)까지 동작할 수 있는데, 코덱이 필요로 하는 적절할 시스템 클럭을 생성해 내기 위하여 DSP의 동작 주파수를 40.96 MHz로 설정하였다.

실시간 알고리즘 구현에서 주파수 차감, 끝점 검출, 특징 추출 단계까지만 실시간으로 처리된다. 여기서 실시간 처리라 함은 핑퐁버퍼를 이용하여 스트림 데이터의 형태로 처리되는 것을 의미한다. 비터비 알고리즘을 이용하는 화자 독립 인식기의 경우, frame-synchronous 방식의 비터비 알고리즘을 이용하면 실시간 비터비 연산이 가능하지만, 본 개발에서는 주파수 차감 과정의 연산량 증가로 실시간 비터비 알고리즘 연산은 현실적으로 어려웠다. 따라서 주파수 차감과 끝점검출을 포함한 특징추출 부분만 실시간으로 수행되며 <표 3>과 같이 전체적으로 DSP 성능의 52%만으로 처리가 가능하였다. 그러나 고립단어 인식의 경우, 발성 시간이 길지 않기 때문에, 인식 단어 수가 많지 않다면 오프라인 비터비 연산을 적절한 시간에 마칠 수 있다. 개발된 인식기의 경우 발성 종료 후, 발성의 길이에 따라서 0.5~0.8 초 안에 인식 결과를 얻을 수 있었다. 다음은 실시간 동작하는 알고리즘 블럭들의 연산 소요량을 보여준다.

화자 종속 인식기를 포함하는 임베디드 인식기의 경우, 호스트 인터페이스를 위한 프로토콜이 비교적 복잡해진다. 앞에서 언급한 바와 같이 DSP는 호스트와의 인터페이스를 위한 통신 관련 주변 장치가 없기 때문에 호스트와의 인터페이스를 위한 프로토콜 처리 및 호스트와의 통신은 PIC16F648A 마이컴이 담당한다.

표 3. 알고리즘 블록들의 연산 소요량

알고리즘 블럭	프로세싱 타임(percentage)
windowing & FFT (주파수 차감)	9.8%
소음 전력 연산 및 차감 (주파수 차감)	8.5%
IFFT & Overlap add (주파수 차감)	9.4%
MFCC (특징 추출)	17.8%
끝점 검출 및 기타	6.4%
idle time	48.1%

DSP와 PIC16F648A의 연결을 위해 DSP의 외부 인터럽트와 I/O 단자를 이용하여 SPI 프로토콜과 유사한 직렬통신을 구현하였다. PIC16F648A은 DSP와 호스트 사이에서 호스트의 요청이나 DSP의 요청 및 인식 결과를 적절한 프로토콜로 변환하여 연결하여 준다. PIC16F648A와 호스트는 UART 직렬 통신방식으로 연결된다. 프로토콜은 크게 COMMAND와 STATUS로 나누어지며 8bit의 패킷으로 구성되어 있다. STATUS는 DSP가 호스트 측으로 전달하는 인식 결과나 인식기의 현재 상태에 대한 정보를 의미한다. 화자 종속 모드에서는 훈련 시, 훈련 상태에 대한 정보도 포함된다. COMMAND는 인식 개시, 인식 명령어가 그룹핑 되어 있을 경우 그룹 인덱스 등을 포함한다. 화자 종속의 경우, 훈련 명령어 인덱스, 사용자 인덱스를 포함한다. 한편, 화자 종속의 경우 2 명의 화자가 등록하여 사용할 수 있도록 되어 있으며, 최대 30 개의 명령어를 등록할 수 있다.

다음은 음성인식기가 동작할 때의 state diagram이다

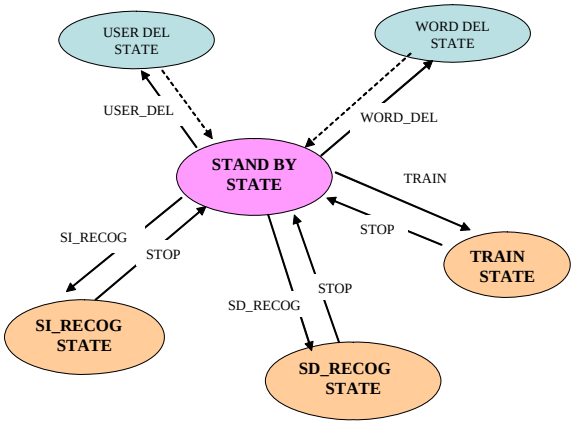

그림 5. 음성인식기 동작 state diagram

각 상태의 의미는 다음과 같다.

<그림 5>에서 화살표는 각각의 상태를 이동할 때 요구되는 COMMAND이고, 점선으로 된 화살표는 해당 상태에 들어갔을 경우, COMMAND의 요청 없이 STAND BY STATE로 돌아온다는 것을 의미한다. 대부분의 자동차용 인식기가 그러하듯이 개발된 인식기도 버튼 트리거(button trigger)

방식을 사용한다. 따라서 주파수 차감 방식에서 잡음의 주파수를 추정하기 위하여 잡음과 음성 신호를 구별할 필요가 없다. 버튼이 눌려지지 않은 상태(STAND BY STATE)에서는 입력되는 신호가 잡음이라고 가정하고 잡음의 주파수 스펙트럼을 추정하고, 인식 버튼이 눌러진 이후에는 더 이상 잡음을 추정하지 않고 그때까지 추정된 잡음의 주파수 스펙트럼을 이용하여 잡음 주파수 성분을 차감하게 된다.

표 4. 음성인식기 상태

상태	설명
SD_RECOG STATE	화자 종속 인식 상태에 들어가 있는 상태
SI_RECOG STATE	화자 독립 인식 상태에 들어가 있는 상태
TRAIN STATE	화자 종속 인식을 위한 훈련 모드에 들어가 있는 상태
WORD DEL STATE	화자 종속 인식기에서 훈련된 명령어를 삭제하는 상태
USER DEL STATE	화자 종속 인식기에서 등록된 사용자를 삭제하는 상태
STANBY STATE	대기 중인 상태. 이 상태에서는 주파수 차감에 적용할 잡음의 주파수를 추정하고 있는 상태

코덱은 A/D 변환기와 D/A 변환기를 모두 포함하고 있기 때문에 이를 이용하여 인식 결과 및 안내 메세시를 음성으로 출력하도록 하였다. 인식 결과와 안내 메시지는 ADPCM으로 압축 저장되어 있으며, 필요시 해당 ADPCM 음성 데이터를 복호화하여 출력한다. 한편, 화자 종속의 경우는 화자가 해당 명령어를 훈련할 때, 화자의 음성을 ADPCM으로 압축하여 저장하고 있다가 인식 시에 인식 결과를 음성으로 출력하는데 사용한다. 따라서 개발된 인식기는 ADPCM 부호화기와 복호화기를 모두 포함하고 있다. 인식을 위한 A/D 변환기의 샘플링 주파수로 16 KHz를 사용하기 때문에 입력된 음성을 곧 바로 ADPCM으로 변환하면 데이터 양이 증가함으로 ADPCM으로 압축하기 전에 8 KHz로 샘플링 주파수 변환을 한 후에 압축을 한다. TLV320AIC11 코덱은 A/D 변환기와 D/A 변환기의 샘플링 주파수가 동일하게 설정되기 때문에 출력 시에는 ADPCM 복화화를 한 후 다시 16 KHz로 샘플링 주파수 변환하여 출력한다.

실차 테스트의 경우 마이크는 앞쪽 실내등에 위치하기 때문에 입과 마이크의 거리가 다소 먼 편인데 이는 대부분의 자동차가 음성인식을 위한 별도의 마이크를 제공하지 않고 핸즈프리를 위한 마이크를 음성인식과 공유하기 때문에 마이크의 위치에 제약이 있다. 마이크가 실내등 위치에 위치하는 이유는 핸즈프리의 경우 운전자와 조수석에 있는 사람을 모두 고려하기 때문이다. 한편, 실차 테스트 과정에서 한 가지 예상치 못한 문제점이 발견되었는데, 개발 단계에서는 문제가 되지 않았던 자동차의 전기 계통에서 발생하는 전형적인 잡음이 상당량 유입되었다. 이는 자동차에서 사용하는 전원과 음성인식 하드웨어의 전원을 완전히 분리할 수 없기 때문에 발생하는 현상으로 별도의 하드웨어적인 조치를 취하지 않는 한 해결되기 어려운 문제였다. 이의 해결책으로 음성인식 하드웨어를 자동차에 장착한 후, 실제 인식 상황에서 A/D 변환되어 입력되는 자동차 전기적인 잡음이 포함된 신호를 녹음하여 이를 모델 훈련에 포함하여 재훈련함으로서 인식률의 저하를 최소화 할 수 있었다. 이 과정에서 실차에 인식기를 장착한 상태에서 인식기에 입력되는 신호를 직접 수집해야 했는데, 처음부터 이를 고려했던 것이 아니었기 때문에, 소음을 채집하는 과정이 매우 번거로웠으

며 따라서 훈련용 데이터만을 수집하였고, 결과적으로 오프라인 테스트를 통한 인식률은 얻을 수는 없었으며, 실차 환경에서 체감 인식률의 증가만을 확인하였다. 한편, 이 문제는 전원 부분의 하드웨어를 보강하면, 어느 정도 해결될 문제이기 때문에 더 이상의 추가적인 고려는 하지 않았다.

다음은 개발된 음성인식 하드웨어의 실물 사진이다.

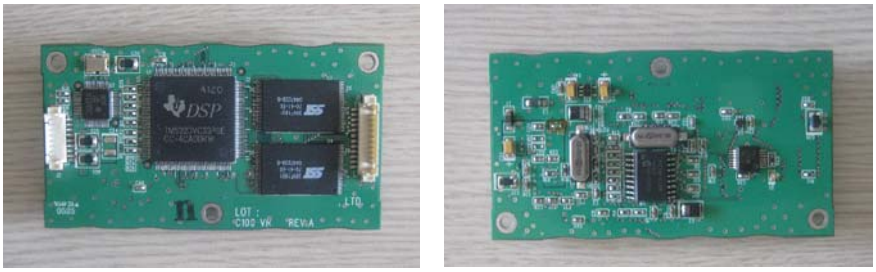

그림 6. 개발된 음성인식 하드웨어

5. 결 론

본 논문에서는 자동차 소음 환경에서 사용될 수 있는 음성 인식기를 임베디드 형태의 하드웨어 모듈로 구현하는 방법을 제시하였다. TMS320VC33 부동소수점 DSP를 이용하여 구현된 음성 인식기는 자동차의 편의 장치를 제어하기 위한 화자 독립 인식기와 음성 다이얼링을 위한 화자 종속 인식기를 모두 포함하고 있다. 화자 독립 인식기는 높은 인식률을 얻을 수 있는 고정 어휘 방식을 채택하였으며, 이를 위하여 인식기에 최적화된 음성 데이터베이스를 구축하였다. 주파수 차감 방법을 특징 추출 단계에서 적용하지 않고, 잡음을 포함한 입력 음성에 직접 적용하여 잡음이 제거된 음성을 얻은 후 끝점 검출을 수행하기 때문에 정확한 끝점 검출이 가능하였다. 또한, 주파수 차감 방법에서 기인하는 비선형 왜곡으로 인한 모델 불일치를 최소화하기 위하여 모델 훈련 시, 인식기와 동일한 주파수 차감 및 끝점 검출기를 이용하여 처리한 후 모델을 훈련시킴으로서 인식기가 동작할 때의 상황과 동일한 조건을 유지하도록 하였다.

프로세서의 성능과 가용한 메모리를 고려하여, 다양한 오프라인 인식 테스트를 통하여 적절한 인식 파라메터를 선정하였으며, 최종적으로 스테이트 당 3 개의 mixture로 구성된 모델을 사용하였다. 구현된 하드웨어는 호스트와의 인터페이스를 위하여 microcontroller를 내장하였다. 한편, 화자 독립 인식기와 화자 종속 인식기가 주어진 응용에 맞게 원활이 동작할 수 있게 하는 프로토콜을 설계하였으며, 이 프로토콜의 처리를 microcontroller가 담당하도록 하였다.

본 연구는 상용차에 장착을 목표로 진행되었으며 개발 초기 단계부터 자동차 제조사와의 상의를 통해 최종 사양을 결정하였다. 최근 대용량 인식 기술 및 연속어 인식 기술의 발달로 지능적인 인식기의 개발이 가속화되고 있지만, 인식의 신뢰성 면에서 아직 부족한 점이 많다고 할 수 있다. 이런 점에서 소용량 인식기이지만, 소음 하에서 신뢰성 있게 동작하는 것이 상용화에서는 가장 중요한 요소로 여겨지며, 이 점에 초점을 맞추어 연구 개발을 진행하였다. 한편, 최근 자동차가 지능

화되면서 음성 인식을 소프트웨어적으로 처리할 수 있는 장치들을 내장하고 있으나, 이들 장치에 소프트웨어를 사전 장착하기 위해서는 자동차 제조사의 결정과 공동 개발 작업을 포함하기 때문에 장점에도 불구하고 생각만큼 쉽지 않다. 그런 측면에서 독립적으로 개발되어 장착할 수 있는 하드웨어 모듈 방식의 음성인식기가 비용 추가에도 불구하고 개발 과정의 용이성 및 위에서 언급한 상용화 가능성 면에서 나름대로의 활용도가 있다고 사료된다.

참 고 문 헌

[1] Abut, H. 2005. *DSP for In-Vehicle and Mobile Systems*, Springer.
[2] Junqua, J. C. 2000. *Robust Speech Recognition in Embedded Systems and PC Applications*, Kluwer Academic Publishers.
[3] Boll, S. F. 1979. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. of Acoustics, Speech and Signal Processing* ASSP-27(2), 113-120.
[4] Rabiner, L. & Juang, B. 1993, *Fundamentals of Speech Recognition*, Prentice Hall.
[5] 2004. *TMS320C3x User's Guide*, Texas Instruments.

접수일자: 2008. 4. 30
게재결정: 2008. 6. 11

▲ 정익주
강원도 춘천시 효자2동 (우: 200-701)
강원대학교 전기전자 공학부
Tel: +82-33-250-6322
E-mail: ijchung@kangwon.ac.kr