논문 2005-42TC-11-12

텔레메틱스 단말기 내의 오디오/비디오 명령처리를 위한 임베디드용 음성인식 시스템의 구현

(Implementation of Embedded Speech Recognition System for Supporting Voice Commander to Control an Audio and a Video on Telematics Terminals)

권 오 일*, 이 흥 규**

(Ohil Kwon and Heungkyu Lee)

요 약

본 논문에서는 차량 내에서 음성인식 인터페이스를 이용한 오비오, 비디오와 같은 응용서비스 처리를 위해 임베디드형 음성인식 시스템을 구현한다. 임베디드형 음성인식 시스템은 DSP 보드로 제작 포팅된다. 이는 음성 인식률이 마이크, 음성 코덱 등의 H/W의 영향을 받기 때문이다. 또한 차량 내 잡음을 효율적으로 제거하기 위한 최적의 환경을 구축하고, 이에 따른 테스트 환경을 최적화한다. 본 논문에서 제안된 시스템은 차량 내에서의 신뢰적인 음성인식을 위해 잡음제거 및 특징보상 기술을 적용하고 임베디드 환경에서의 속도 및 성능 향상을 위한 문맥 종속 믹스처 공유 음향 모델링을 적용한다. 성능평가는 일반 실험실 환경에서의 인식률과 실제 차량 내에서의 실차 테스트를 통해 검증되었다

Abstract

In this paper, we implement the embedded speech recognition system to support various application services such as audio and video control using speech recognition interface on cars. The embedded speech recognition system is implemented and ported in a DSP board. Because MIC type and speech codecs affect the accuracy of speech recognition. And also, we optimize the simulation and test environment to effectively remove the real noises on a car. We applied a noise suppression and feature compensation algorithm to increase an accuracy of sppech recognition on a car. And we used a context dependent tied-mixture acoustic modeling. The performance evaluation showed high accuracy of proposed system in office environment and even real car environment.

Keywords: speech recognition, speech preprocessing, embedded system, voice user interface

I. 서 론

최근 음성인식 기술의 발달과 중요성 등으로 인해 다양한 범위로 응용범위가 확대되고 있다. 이러한 분야 중에 최근에 대두되고 있는 분야가 텔레메틱스 분야이다. 텔레메틱스 분야는 차량 내에서 운전자가 운전을 하는 도중에 전자 기기를 제어하기 위해 한눈을 파는 사이 운전자는 교통사고의 위험에 드러나기 때문에 안전운행을 위한 수단으로 음성인식 인터페이스의 중요성이 대두되었다.

그러나 차량 내에서 발생하는 잡음들로 인해 인식률이 현격하게 떨어지는 결과를 초래하기 때문에 상용 서비스를 제공하기에는 어려운 실정이다. 차량

내 잡음 환경 뿐만 아니라 외부에서의 차량의 소리, 오디오, 비디오 소리, 차량 내의 사람들의 말소리 등 해결해야 할 문제점들이 많다.

* 종신회원, 현대오토넷(주)
(Hyundai Autonet Corp. Navigation Team.)
** 정회원, 미디어젠(주), 음성 인식/합성 연구소
(Mediazen Corp. Speech Interactive Lab.)
접수일자: 2005년5월3일, 수정완료일: 2005년11월10일

따라서 본 논문에서는 이러한 문제점들을 해결하기 위한 방안을 모색했다. 첫째로 차량 내에 음성인식 인터페이스를 어떻게 설치 운영 할 것인가의 문제가 있다. 마이크 종류, 마이크 위치의 설정 및 음성인식 시작을 알리기 위한 버튼(push to talk)의 설치는 잡음을 효율적으로 제거하기 위한 기준이 될 뿐만 아니라 인식률 향상에 많은 기여를 한다. 본 논문에서 제안한 시스템에서는 지향성 마이크를 사용하고 마이크의 위치는 썬바이저에 장착을 한다. 그리고 음성인식 시작을 알리기 위해 별도의 버튼을 DSP 보드에 장착을 해서 사용했다.

둘째로 고려해야 할 사항은 잡음환경 하에서의 음성인식률의 향상 문제이다. 대부분의 학교나 연구단체, 기업에서의 음성인식률은 실험 환경에서는 상당히 높은 편이나 차량 내에서의 인식률은 90%이상 나오기가 힘든 실정이어서 기술상 해결해야 할 점들이 많은 실정이다. 따라서 본 논문에서는 잡음제거 기법과 특징보상기법을 적용한다. 잡음제거 기법은 스펙트럴 차감법(SS : Spectral Subtraction)을 사용한다. 특징보상기법은 잡음환경에서 인식성능을 향상시키기 위한 방법이다. 기존에 음성 모델 기반으로 한 제안됐던 기법들은 실제 환경과 유사한 조건의 음성 데이터를 이용하여 적응시킨 모델을 사용하는 방법이다. 그러나 이러한 기법들을 사용하기 위해서는 많은 음성데이터를 구축해야 하기 때문에 어려운 작업일 뿐만 아니라, 다이나믹하게 변하는 환경에 대응하기 쉽지 않다는 단점이 있다. 따라서 본 논문에서는 조용한 환경에서 수집된 데이터를 이용하여 실제 잡음환경에 맞게 변환시키는 GMM 기반의 특징보상(RATZ: multivariate Gaussian-Based Cepstral normalization) 기법을 사용한다.

셋째로 임베디드환경 하에서의 인식속도가 실시간에 가깝게 나오는 것을 요구하고 있다. 텔레메틱스 환경하에서 자주 쓰이는 단어의 목록이나 바로 가기 기능과 같은 서비스를 제공하기 위한 글로벌 명령어와 서비스의 단위에 따른 로컬 명령어의 집합을 포함해 하나의 서비스 화면에서 100단어 미만의 명령어를 요구하고 있으며, 서비스 장면의 변화에 따른 인식목록의 변화를 요구하기 때문에 가변어휘 인식기능을 요구한다. 따라서 임베디드 환경에서 음향 모델링을 위한 메모리 문제, 인식과정에 있어서의 계산량 최소화, 상업용 서비스에 적용하기 위한 높은 인식률 등을 고려해야 한다. 기존의 연속 HMM의 경우 높은 인식률을 보이지만 모델의 크기와 계산량이 커 임베디드 환경에 적합하지 않다. 또한 이산 HMM의 경우에는 벡터 양자화를 사용함으로써 발생하는 양자화 오차를 보상할 수 없는 문제 때문에 상대적으로 낮은 인식률을 가지는 단점이 있다. 따라서 본 논문에서는 이의 절충 형태로 문맥 종속 믹스쳐 공유 음향 모델링 기법을 적용한다.

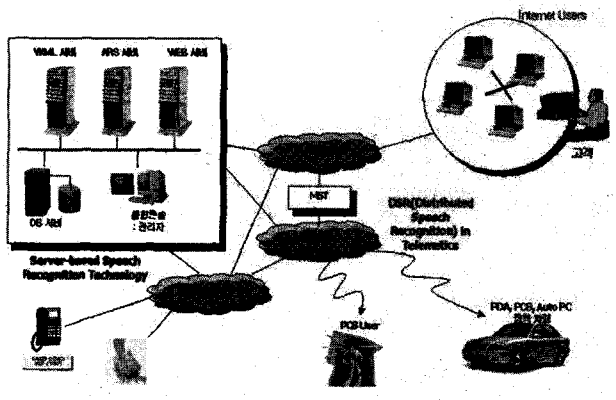


그림 1. 텔레메틱스 서비스 구조도
Fig. 1. Service architecture for Telematics service.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 차량 내에서 잡음 처리를 위한 잡음제거 기법 및 특징 보상 기법에 대해 제안하다. Ⅲ장에서는 임베디드 시스템을 위한 음향 모델링 기법을 설명한다. 그리고 Ⅳ장에서는 구현된 알고리즘을 DSP 보드를 제작하여 포팅을 한다. Ⅴ장에서는 제작 된 임베디드 시스템을 사용한 실험 결과에 대해 언급한다. 최종적으로 Ⅵ장에서는 결론을 내린다.

Ⅱ. 차량잡음 처리를 위한 전처리

잡음 환경에서의 음성 인식 성능 향상을 위해서는 입력된 잡음 환경의 음성 신호에 대한 전처리 과정이 필수적이다. 음성인식 향상을 위해 전처리 부분에서 고려되어야 하는 부분은 신뢰적인 끝점검출 기술과 오염된 음성신호로부터 잡음 데이터만을 적절히 제거하는 기법, 특징 보상 기법 들이 있다.

2.1 주파수 차감법(Spectral Subtraction)을 이용한 잡음제거

주파수 차감법은 잡음 환경의 신호를 주파수 영역에서 잡음 성분을 추정하여 차감해주는 방식이다. 음성인식을 위한 특징벡터를 추출하기 전에 신호를 깨끗한 신호로 추정하기 위한 방식으로 비음성 구간에서의 잡음에 대한 분포 추정을 한 후 적용된다.

시간 축에서의 잡음 신호는 다음식과 같이 표현된다.

$$y(m) = x(m) + n(m) \quad (1)$$

여기서 y(m), x(m), n(m)은 각각 잡음이 부과된 음성신호, 깨끗한 음성신호, 잡음 신호이다. 위 식은 주파수 영역에서 다음과 같이 표시된다.

$$Y(f) = X(f) + N(f) \quad (2)$$

각각은 해당 시간 축 신호의 퓨리에 변환 (Fourier Transform)을 한 것이다. SS를 적용하면 식 (3)과 같다.

$$\left|\hat{X}(f)\right| = \left|Y(f)\right| - \alpha\left|\hat{N}(f)\right| \quad (3)$$

α는 over-subtraction factor로 보통 1보다 큰 값을 가지며 추정된 잡음 $\left|\hat{N}(f)\right|$과 현재 입력 $|Y(f)|$의 비에 따라 그 값이 결정되어 차감의 정도가 달라진다. $\left|\hat{N}(f)\right|$는 오직 잡음만 존재하는 구간의 i번째 프레임 $|N(f)|$의 K개의 평균 잡음 스펙트럼으로 식(4)와 같이 표현된다[6][8].

$$\left|\hat{N}(f)\right| = \frac{1}{K}\sum_{i=0}^{K-1}\left|N_i(f)\right| \quad (4)$$

2.2 GMM 기반의 특징보상 기법

GMM 기반의 특징보상 기법인 RATZ는 켑스트럼 도메인에서 잡음환경의 음성 특징을 조용한 환경의 음성 특징으로 보상해 주는 기법이다. 즉 특정한 잡음 환경 음성데이터로부터 평균과 분산을 추정하여 실제 잡음 환경의 입력 신호가 들어오면 그 변화량 만큼 보상해서 깨끗한 환경의 음성신호 특징으로 추정하는 방식이다[5][9].

RATZ 알고리즘은 잡음환경과 깨끗한 환경간의 보정인자(correction factor)를 구하는 것을 요구한다. 보정인자 식은 EM 알고리즘에 의해 다음과 같이 나타나며 Likelihood 값이 크게 되는 방향으로 식 (5), (6), (7)이 반복된다.

$$\bar{\mathbf{r}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)\,|\,\mathbf{y}_t,\phi]\,\mathbf{y}_t}{\sum_{t=1}^{T} P[s_t(k)\,|\,\mathbf{y}_t,\phi]} - \boldsymbol{\mu}_{\mathbf{x},k} \quad (5)$$

$$\bar{\mathbf{R}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)\,|\,\mathbf{y}_t,\phi]\,(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{x},k} - \bar{\mathbf{r}}_k)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{x},k} - \bar{\mathbf{r}}_k)^T)}{\sum_{t=1}^{T} P[s_t(k)\,|\,\mathbf{y}_t,\phi]} - \Sigma_{\mathbf{x},k} \quad (6)$$

$$\alpha = \{r_1,\cdots,r_K,R_1,\cdots,R_K\} \quad (7)$$

여기서 α는 K개의 보정인자의 평균벡터, r과 분산벡터, R의 집합이다. $P[k/y_t,\alpha]$는 잡음 환경 신호의 t번째 특징벡터, yt가 이전 ,α에 대하여 k번째 가우시안 믹스쳐에서 발생할 사후 확률이다. $\mu_{x,k}$와 $\Sigma_{x,k}$는 각각 조용한 환경 신호의 k번째 가우시안 믹스쳐의 평균벡터와 분산벡터이다.

보정인자를 구했다면 입력 잡음 신호 특징벡터에 대해 보정 인자를 차감하여 조용한 환경의 신호 특징벡터 $\hat{x}$로 추정하며 식 (8)과 같이 MMSE estimator를 적용한다.

$$\hat{\mathbf{x}}_{MMSE} = E\{\mathbf{x}\,|\,\mathbf{y}\} = \int_{\mathbf{x}} \mathbf{x}\cdot p(\mathbf{x}\,|\,\mathbf{y})d\mathbf{x} \quad (8)$$

문제를 간단히 하기 위해 $\mathbf{x} = \mathbf{y} - \mathbf{r}(\mathbf{x})$와 같이 복구식을 가정하면, 다음과 같이 정리할 수 있다.

$$\begin{aligned}\hat{\mathbf{x}}_{MMSE} &= \mathbf{y} - \int_{\mathbf{x}} r(\mathbf{x})p(\mathbf{x}\,|\,\mathbf{y})d\mathbf{x} = \mathbf{y} - \int_{\mathbf{x}}\sum_{k=1}^{K} r(\mathbf{x})p(\mathbf{x},k\,|\,\mathbf{y})d\mathbf{x} \\ &= \mathbf{y} - \sum_{k=1}^{K} p[k\,|\,\mathbf{y}]\int_{\mathbf{x}} r(\mathbf{x})p(\mathbf{x}\,|\,k,\mathbf{y})d\mathbf{x} \\ &\cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k\, p[k\,|\,\mathbf{y}]\int_{\mathbf{x}} p(\mathbf{x}\,|\,k,\mathbf{y})d\mathbf{x} \\ &\cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k\, p[k\,|\,\mathbf{y}]\end{aligned} \quad (9)$$

Ⅲ. 문맥 종속 음향 모델링

임베디드 시스템 개발 시 메모리의 제한을 고려해야 하며 최적의 성능 및 속도를 위한 알고리즘 구현이 필요하다. 또한 실제 서비스 사용을 위한 거절 기능 또한 중요하다[11].

3.1 반연속 HMM 모델

반 연속HMM(SCHMM)은 연속 HMM의 모든 가우시안에 대한 양자화를 통해 각 코드워드(codeword)가 가우시안 풀(pool) 형태의 코드북(codebook)을 형성하며, 출력 확률 분포는 가중치(weight)가 부여된 풀 내부

가우시안들의 합이므로 상태(state)들은 그 코드 워드들의 가중치만을 가지면 된다[1].

식 (10)은 반 연속 HMM의 상태 s에서 벡터 x_t에 대한 출력확률 값을 나타낸다.

$$b_s(o_t)=\sum_{l=1}^{L} b_s(o_t)N(x_t;\mu_l,\sum_l) \quad (10)$$

여기서 L은 코드북의 크기이고, $b_s(o_t)$는 벡터 x_t의 이산 확률 값으로 가중치의 의미를 갖고 $N(\bullet)$는 가우시안 분포를 나타낸다.

3.2 믹스쳐 공유(Tied-mixture) HMM 모델

믹스쳐 공유 모델링 기법은 기존의 연속 HMM 가우시안들을 이용하여 가우시안 풀 형태의 새로운 코드북을 생성하는 방법이다[2]. 연속 HMM은 각 상태별로 가우시안 믹스쳐와 믹스쳐의 가중치들을 가지고 있다. 믹스쳐 공유 HMM 모델링에서는 존재하는 모든 연속 HMM의 가우시안들을 하나의 풀에 모아서 설정한 크기의 코드북이 되도록 한다. 이때, 존재하는 가우시안의 개수가 코드북의 크기보다 작은 경우는 분산이 가장 큰 가우시안부터 mean deviation을 통하여 부족한 개수만큼 코드북 크기를 채우며, 가우시안의 개수가 코드북의 크기보다 많으면 가중치가 작은 것부터 제거하는 방법으로 코드북을 형성한다. 이렇게 형성된 코드북을 각 HMM의상태들은 식(10)과 같은 형태로 공유하며 각 코드워드에 대한 코드북 크기만큼의 가중치 배열을 가진다.

3.3 믹스쳐 공유 HMM 계산량 최소화

본 논문에서는 믹스쳐 공유 HMM[10]의 계산량 최소화를 위해 고정 한계값 제거 기법 및 상태수준에서의 가우시안 선택 알고리즘[3][4]을 구현한다. 고정한계값 제거 기법은 일정 기준 이하의 값을 가지는 가중치는 고려하지 않는 것을 말한다. 이는 코드워드들에 대한 가중치 합으로 상태 출력 확률 분포를 계산하면 실질적인 상태 당 믹스쳐 수가 많아지게 되고, 계산 시간 또한 증가하기 때문에 이를 해결하기 위한 방식이다.

상태 수준에서의 가우시안 선택 기법은 인식과정에서 활용되는 HMM 상태에 대해서 해당상태의 최고 확률 값을 기준으로 적절한 가우시안들만을 확률값 계산에 사용하는 것을 말한다. 즉 각 상태에서의 확률값은 식 (11)과 같다.

$$b_s(x_t)=\sum_{pi>thres} p_i$$
$$p_i=b_s(o_t)N(x_t;\mu_i,\sum_i) \quad (11)$$

여기서 pi는 i번째 가우시안의 likelihood이다. 또한 많은 개수의 가우시안들이 문턱값을 넘을 경우 연산량의 증가를 피하기 위해 최대 계산 가능 가우시안 개수를 제한하도록 한다.

Ⅳ. 임베디드 시스템 제작 및 구현

본 실험에서 사용한 DSP는 TI사의 TMS320 C6711B 이며 사용한 코덱은 마찬가지로 TI 사의 TLV320AIC21 이다.

6711B의 기본 클럭은 100MHz이며 부동소수점 연산이 가능한 32비트의 범용레지스터 및 산술명령어를 지원한다. 메모리는 내부에 4Kbyte의 프로그램/데이타 메모리를 가지고 있으며 64kbyte의 사용자가 제어 가능한 L2 캐쉬를 지원하므로 인식엔진의 핵심 코드는 내부메모리에 적재하여 사용하게 된다. 그 외에 인터페이스 및 부가적인 함수들은 외부 메모리인 SDRAM을 사용하게된다. 이외의 외부 메모리는 부트를 지원하기 위한 2Mbyte 플래쉬램이 있다.

AIC21 코덱은 다채널이 지원가능하며 기본클럭은 최대 100MHz까지 지원하고 샘플링주파수는 26Khz까지 가능하다. 본 실험에서는 2개의 채널중 한개 채널만을 사용하였으며 기본클럭으로 50MHz를 사용하였고 샘플링 레이트는 11.025KHz를 사용하였다.

그림 2. 구현 된 차량용 임베디드 시스템 및 시뮬레이터

Fig. 2. Testbed simulator for car navigation system.

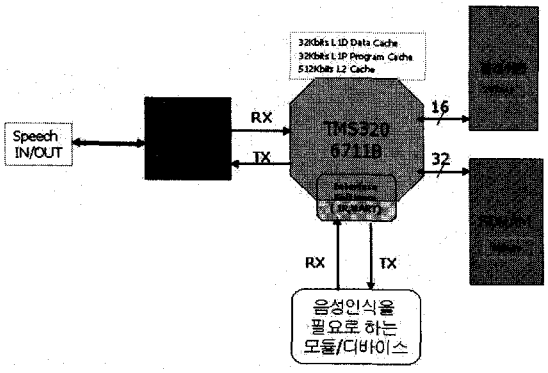


그림 3. DSP 블록도
Fig. 3. DSP blockdiagram for embedded speech recognition.

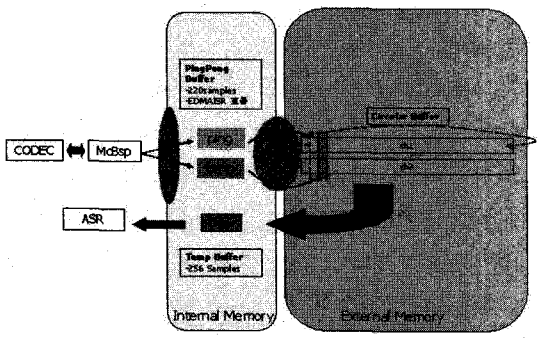


그림 4. 음성입력블록도
Fig. 4. Speech input method for time synchronous of speech recognition processing in non-operating system.

실제 하드웨어 구성은 그림 3과 같다. 코덱(TLV320AIC21)은 DSP의 McBsp(Multi Channel Buffered Serail Port)와 연결되어 코덱의 제어와 음성 입력을 주고 받는다. 본 하드웨어는 인식 결과를 인지하기 위한 방법으로 음성인식을 요하는 프로세서나 디바이스를 위한 인터페이스 구조를 가지며 통신방법으로는 UART나 IR의 기능을 프로그램하여 DSP의 McBsp나 Timer 포트를 사용하여 구현한다. 실제 ASR을 위한 음성입력 동작은 그림 4와 같다.

McBsp를 거쳐서 들어오는 음성입력은 DSP의 EDMA(Enhanced DMA)를 이용하여 Ping Pong 버퍼에 한 프레임씩 저장하게 된다. 이때 EDMAISR (EDMA Interrup Service Routine)에서는 ASR에 필요한 전처리를 수행하게 되며 한 프레임씩 데이터의 손실 없이 Ping Pong버퍼의 데이터를 저장한다. 전처리 동작을 거친 스피치는 신뢰성을 위하여 외부메모리에 저장되어지며 필요한 프레임만을 내부메모리로 옮기어 ASR을 수행하게 된다.

V. 실험 결과

본 시스템의 성능평가를 위해 일반실험환경과 실제 차량환경에서의 테스트를 수행하였다. 일반실험환경에서도 깨끗한 환경에서의 성능평가를 수행하고, 실제 차량에서 수집된 오염된 음성 데이터베이스를 수집해서 오프라인에서 테스트를 수행하였다. 그리고 최종적으로 남녀20명씩을 실차(EF쏘나타, SM5)에서 테스트를 수행했다. 다음은 실험실 환경에서 수행된 전처리 기술에 대한 성능 평가와 인식률에 대한 성능 평가를 기술한다.

표 1은 주파수 차감법을 이용한 실험 결과이다. 실험은 Aurora2 task를 이용한 음성인식율로 평가했으며 SS방식과 잡음추정을 위한 VAD를 이용한 SS방식을 비교했다. 표1에서의 개선된 SS의 경우는 계산 최소화를 수행한 작업이다. VAD를 이용한 spectral subtraction의 경우 attenuation factor를 5, floor factor를 0.05로 하였다. 결과표에서 나와 있듯이 VAD를 이용한 잡음추정방법이 효과적임을 알 수 있다.

표 2는 실제 차량 잡음 음성 샘플인 Car01에 대한 RATZ 알고리즘 적용 실험 결과를 보여주고 있다. 기본 성능이 93.36%인 음성 인식 시스템을 그대로 적용했을 경우 87.56%로 인식율이 저하되는 것을 알 수 있다. Car11 잡음 샘플 451개를 RATZ 모델 훈련에 사용하였고, 8개의 mixture로 오염 음성의 확률 분포를 추정하였다. RATZ 적용 결과 92.67%로 인식율이 상승하는 것을 확인하였으며, 이로부터 잡음 음성 모델을 기반으로 하는 특징 보상 기법이 실제의 차량 잡음 환경

표 1. Aurora2 task를 이용한 음성인식율 (Word Accuracy, %)
Table 1. Speech recognition accuracy using Aurora 2 speech database (Word Accuracy:%).

잡음추정실험방법	Explicitb			VAD	
	Baseline	SS	개선된 SS	SS	개선된 SS
clean	99.02	98.99	98.81	98.93	98.93
20dB	97.97	98.24	98.06	97.88	98.36
15dB	94.24	96.93	96.99	96.66	97.91
10dB	78.17	89.98	93.20	92.45	94.69
5dB	42.59	64.42	82.11	82.49	86.31
0dB	14.67	27.14	52.67	53.36	59.32
-5dB	9.25	10.14	18.19	18.85	21.12
Ave.	69.79	75.34	84.60	84.57	87.32

표 2. 실제 차량 잡음 오염에 대한 RATZ 적용 성능 평가
Table 2. Performance evaluation of RATZ on real car noisy speech database.

인식 데이터	Clean	Car01	Car11	Car01 +RATZ
인식율	93.36%	87.56%	84.26%	92.67%

표 3. 쌍용 자동차 포팅용 Embedded system의 음향 모델 성능
Table 3. Performance evaluation and comparison about speech recognition accuracy and speed between HTK and Workbench.

인식기		HTK	Workbench
인식률	Clean	99.81%	98.77%
	SS+RATZ	93.07%	92.47%
인식시간		T	0.57T

표 4. 실차 테스트 결과
Table 4. Test results on real cars.

	조용한 환경	시내 주행	고속 주행	평균	차종
오프라인	99.69	94.44%	82.10%	-	아반테 /씨에로
남자	-	95.4%	96%	95.7%	EF쏘나타, SM5
여자	-	89.5%	89.2%	89.3%	EF쏘나타, SM5
평균	-	92.5%	92.6%	92.5%	EF쏘나타, SM5

에서도 인식율 향상에 도움이 되는 것을 확인 할 수 있었다.

표 3은 Clean은 정차상태에 있는 차량에서 녹음한 음성에 대한 인식률이고 SS+RATZ는 시내 및 고속주행 시 녹음한 음성에 주파수 차감법과 RATZ 전처리 알고리즘을 동시에 적용하여 인식한 결과이다. 인식시간은 각 상태에서 모든 가우시안들에 가중치를 적용하고 출력확률을 계산하는 경우의 인식시간을 기준으로 한 상대적 비율이다.

다음은 남녀 20명씩 총 40명의 인원이 EF쏘나타와 SM5 차량에서 실차 테스트한 결과를 기술한다. 각 남녀 20명은 음성인식에 익숙하지 않은 일반인들을 대상으로 평가를 했다. 연령별로는 20대, 30대, 40대로 남녀 각각 7,7,6명씩 테스트를 수행했다. 단어목록은 차량 내에서 사용될 수 있는 용어를

선별한 90개의 목록을 사용했다. 테스트 환경은 차량의 조수석에 있는 썬바이져에 마이크를 설치했고 음성의 시작신호를 알리기 위한 버튼을 DSP보드에 장착을 해서, 사용자가 발음을 하기 전에 버튼을 한번 클릭 하도록 했다. 발음은 각 사람마다 시내주행 두번, 고속주행 두번, 총 4번의 발음을 했다.

표 4는 실차 테스트를 한 결과를 나타낸다. 평균 92.5%의 성능을 나타냈다. 매우 성능이 높은 것은 아니지만, 상용화 수준에 가까운 인식률로 판단된다. 남자 20명의 경우 평균 95.7%의 우수한 성능을 나타낸 반면, 여자 20명의 경우 89.3%의 만족스럽지 못한 성능을 가져왔다. 이는 마이크의 입력 게인(gain)을 설정하는 과정에서의 오차라고 볼 수 있다. 여자들의 경우 대부분 앉은 키가 남자에 비해 작고 목소리가 작아서 음성입력 파형이 작게 들어가고 해서 잡음과 음성신호 간의 변별력이 떨어졌을 것으로 판단하고 있다. 현재 마이크의 볼륨을 외부에서 조절할 수 없기 때문에 향후 이에 대한 대비책이 필요한 실정이다.

VI. Conclusion

본 논문에서 구현된 텔레메틱스 단말기용 임베디드 음성인식 시스템은 깨끗한 환경에서의 인식률 보다 차이가 다소 났지만, 인식단어 수를 줄이고 인식목록의 차별화, 서비스 기획을 통해 실제 상용서비스에 적용할 수 있는 수준이라고 판단된다. 그러나 배경 음악소리, 오디오나 비디오 소리가 큰 경우, 창문 열고 저속주행, 고속 주행 등과 같은 환경에 처했을 경우 많은 도전적인 문제에 직면하게 된다. 따라서 향후 다양한 환경의 주변 잡음 환경 내에서 잡음제거 및 외곡 된 신호로부터 음성 왜곡을 보다 정확히 보상하는 연구가 필요하다. 또한 임베디드 환경 하에서 연속어 인식기능과 같은 고급 기능 구현을 위한 연구개발도 필요한 실정이다. 이러한 기능 수행을 위해서는 음향 모델링이나 디코딩을 위한 신속한 검색기법 및 계산 최소화 기법에 대한 연구도 필요할 것으로 판단된다.

참 고 문 헌

[1] X.Huang, et al, "Semi-continuous hidden Markov models with maximum likelihood VQ", IEEE

Workshop on Speech Recognition, New York, 1988.

[2] S. J. Young, The HTK book, Cambridge University, version 3.2, 1997.

[3] J. Duchateau, et al, "Fast and accurate acoustic modeling with semi-continuous HMMs", Speech Communication 24, 1988.

[4] K.M. Knill, et al, "Use of Gaussian in large vocabulary continuous speech recognition using HMMs", Spoken Language, 1996. ICSLP 96. proceedings, Fourth International Conference on, Volume: 1, 3-6, Oct 1996.

[5] Pedro J. Moreno, "Data-driven environmental compensation for speech recognition: A unified approach", Speech Communication, vol.24, pp267-285, 1988.

[6] SAEED V. VASEGHI, "Advanced Digital Signal Processing and Noise Reduction", WILEY, Second Edition.

[7] Steven F. Boll, "A Spectral subtraction Algorithm for Suppression of Acoustic Noise in Speech", IEEE, No. 1379, pp200-203, 1979.

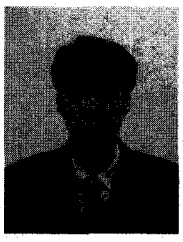
[8] Jounghoon Beh and HanseokK Ko, "Spectral Subtraction Using Spectral Harmonics for Robust Speech Recognition in Car Environments," LNCS, Vol.2660, pp.1109-1116, Jun, 2003.

[9] Wooil Kim, Sungjoo Ahn, Hanseok Ko, "Feature Compensation Scheme Based on parallel Combined Mixture Model", 8thEuropean Conference on Speech Communication and Technology September 2003.

[10] Junho Park, Hanseok Ko, "CONSTRUCTION OF DECISION TREE FROM DATA DRIVEN CLUSTERING," International Conference on Spoken Language Processing(ICSLP) 2002, Vol. 4, pp. 2657-2660, Denver, Sep, 2002.

[11] Taeyoon Kim, Hanseok Ko, "Utterance Verification Under Distributed Detection and Fusion Framework", 8th European Conference on Speech Communication and Technology September 2003.

저 자 소 개



권 오 일(종신회원)

1987년 3월~1991년 2월
고려대학교 전자공학과
학사

1991년 3월~1993년 2월
고려대학교 전자공학과
석사

1993년 3월~1996년 8월 고려대학교 전자공학과
박사

2000년 3월~현재 현대오토넷(주) 차장

1996년 8월~2000년 3월 현대전자산업(주) 차장

<주관심분야 : 음성인식,합성, 임베디드시스템>

이 홍 규(정회원)

1992년 3월~1997년 2월
서경대학교 컴퓨터과학과
학사

1997년 3월~1999년 2월
서경대학교 컴퓨터과학과
석사

1999년 3월~2005년 8월 고려대학교 전자공학과
영상정보처리협동과정 박사

2002년 1월~현재 미디어젠(주), 음성 인식/합성
연구소 개발이사

<주관심분야 : 음성인식/합성, 멀티모달 인터렉션>