DSP를 이용한 연속숫자 음성 인식기 구현

이성권°, 임영춘*, 서준배*, 정현열**
서울통신기술°, 자모바*, 영남대학교**

The Implementation of Continuous Digit Recognition Using DSP.

Seong-Kwon Lee, Young-chun Lim, Jun-Bae Seo, Hyun-youl Jung
SEOUL COMMTECH.CO,LTD, JAMOVA CLS
,joyer.lee@samsung.com, voice@jamova.com

요 약

본 논문은 TMS320C5501 16bit DSP를 적용한 실시간 화자독립 연속 숫자인식기의 구현에 관해 서술한다. 하드웨어 모듈의 구성은 TMS320C5501 300MHz DSP, 코덱으로는 TLV320AIC1103 , SDRAM, 외부장치와의 인터페이스를 위한 HPI, Uart, MIC, SPK Out 단자 로 구성되었다. 음성인식 알고리즘은 HM-Net 방식을 사용하였고 고정소수점 연산처리 방식으로 C를 이용한 최적화 작업을 수행하였으며 스트리밍 방식의 인식 방법으로 실시간 처리가 가능하도록 구현하였다. 숫자 인식에 사용한 모델은 41음소에 기반한 트라이폰을 학습하였으며, 특징 파라미터로는 LPCMEL 20차를 사용하였다. 임베디드 시스템의 실시간 음성인식 시스템 구성에 중점을 두었으며 PC상에서의 성능과 비교해 볼 때 본 DSP 상에서 500단어, 50문장의 인식을 평균 1.5초 전후로 인식하도록 하였으며 간단한 연결 단어 인식을 수행하는데 무리 없음을 보여준다. 특별히 한국어 연속숫자 부분에 중점을 두었고, 본 연구에서 구현된 연속 음성인식 시스템에 사용된 숫자 인식에서 음절 바이폰 모델에 대하여 92.92%의 인식율을 얻을 수 있었다.

1. 서론

최근 들어 음성인식 기술의 발전과 더불어 임베디드 시스템의 활용이 높아지고 있는 가운데 실생활의 응용이 점차 가까워 지고 있다. 가장 편리한 인터페이스 중의 하나가 음성이지만 다양한 환경과, 화자, 그리고 기술의 성능이 그동안 일상 생활에 적용되지 못한 이유일 것이다. 책상에 고정되어 있는 PC 상에서의 인식을 떠나 언제 어디서든지 다양한 환경에서 음성을 통한 제어나 정보를 얻을 수 있는 것이 최종적인 음성인식의 목표가 될 것이다. 이 가운데 한국어 숫자음의 인식이 가장 어렵고도 제일 먼저 난관에 걸리는 문제이다. 이 부분을 해결하고자 제한된 임베디드 시스템에서 숫자음 인식을 수행하여 보았다

2. 고정 소수점 DSP를 이용한 구현

2.1 하드웨어구성

하드웨어로는 TMS320VC5501DSP, TLV320AIC1103 코덱,SDRAM,FlashMemory,Uart, HPI,MIC, SPKOut 으로 구성하였다. TMS320VC5501 DSP 의 내부 메모리(DARAM)는 16K Word 이며, 음성인식용 모델과 인식어휘를위한메모리및음성인식시필요한동적메모리는외부메모리인 SDRAM 으로 확장성과 범용성을 고려하여 8M Word 를 사용하였다. DARAM 은 처리 속도가 가장 빠르기 때문에 실시간 처리를 해야 할 코드 위주로 할당을 하였으며 펌웨어와 실시간 처리에 영향을 주지 않는 초기화 관련루틴은 SDRAM 영역으로 할당하였다[1]. 보조기억장치로는 flashmemory2MWord 를사용하여펌웨어와 음성인식 프로그램 및 데이터를 저장하는데 사용하였다. 본 DSP 음성 모듈은다음의확장성을 가지며간략히 살펴보면다음과같다

* HPI 통신: HPI(Host Port Interface) Port를 통하여

ExternalHost와 Interface된다.

* JTAG: 통신TAG Port를 통하여 JTAG Emulator와 Interface된다.
* UART통신:UARTPort를통하여 Direct로 External 장비와 통신을 수행하며 MAX3221(RS232 RECEIVER/DRIVER)를 통하여 External 장비와도 통신을 수행한다. (Baud Rate는 115200bps까지지원한다.)
* EMIF:I2CPort를 이용하여 EEPROM의 DATA를 READ하여 H/W Version을 F/W에 알려주며 Audio Codec 과 Serial 통신을수행한다.
* CODEC: TI의TLV320AIC1103은 PCMCodec으로 사용된다. 15Bit Linear Data와 8Bit A-LAW,U-LAW Data를 지원한다. 2PORT MIC INPUT를 통해 MIC1 Port 에는 Condenser Mic Input를 MIC2 Port에는 External Line Input을 Interface한다. MAR20 Port를 통해 SPK OUTPUT을 Interface한다. SYSTEM CLOCK은2.048MHz를 사용한다. Microphone Amplifiers는 MAX35.5db를사용할수있다.
* POWER&RESET:VOICERECOGNITIONMODULE에서의 POWER는 Digital 3.3V, Analog 3.3V, DSP Core1.26V가사용된다.
* JTAG&EXTERNALCONNECTION:JTAGEMULATOR와는 14PIN (2.54 PITCH)를 이용한다. External Connector는 60PIN Connector를 사용하여 HPI, AUDIO IN/OUTPUT, External Line INPUT, External SCC, External PCM Data를 Interface한다. INPUT, External SCC, External PCMData를 Interface한다.

2.2. 소프트웨어 구현

일반적으로 음성인식 알고리즘은 실수 연산을 주로 한다. 고정 소수점 DSP에서 원할한 처리를 하기 위해서 모든 실수연산을 고정 소수점 연산으로 바꾸어야 한다. 고정 소수점 DSP를 이용하여 고속 및 고정밀도를 유지하면서 연산을 하기 위한 방법으로 일반적으로 Q Math를 사용한다. 본 모듈에서는 Q15 포맷에 실수값들을 매칭하였고 C55x에서 지원하는 고정 소수점 연산 함수들의 시뮬레이션 코드를 이용하여 PC에서 고정 소수점 연산으로 실수 연산과의 성능을 비교 검증후에 DSP 모듈로의 포팅을 하였다.

음성 인식 코드는 특징 추출 부분에 신호처리 연산이 집중되어 있고 Viterbi 스코어계산시 약간의연산을 하게 된다. DSP 컴파일러인 CCS에서는 최적화된 고점소수점 연산 함수들을 라이브러리로 제공하고 있기 때문에 라이브러리함수들을 이용하여보다 빠른연산처리를할수있다[1].

나머지 처리는 메인 인식루프에서 반복, 비교, 판단 알고리즘처리에비교적많은시간이소요된다.이 부분은 루프내부를최대한간략화하여어셈블작업없이C 코드최적화를수행하였다[2].

음성인식은가변적인요인이많기때문에동적메모리 할당을많이하는데이부분은 ANSI-C 에서제공하는메모리 할당 계열의 함수를 사용하지 않고 효과적인 메모리 관리및빠른처리를위해서메모리관리함수를구현하여 사용하였다[2].

음성인식 알고리즘의 DSP 포팅 작업은 특징 추출부분을 고정 소수점 연산으로 변환 하는 작업과 메인 인식 루프의가장하위단인 Viterbi 스코어계산하는 부분의고정 소수점 연산 처리와 스코어 누적을 Q math 포맷에 맞게 포팅을 해 주면 나머지 부분의 포팅은 보다 빠른 처리를 위한코드최적화작업을하면대부분의포팅작업은완료된다. 그외에고려해야 할 부분은음성인식 모델과인식어휘 및 문법 정보에 대한 처리이다. PC에서는파일개념이 있기때문에모델파일, 인식어휘파일, 문법정보파일을읽어 들이면 되었지만, DSP 모듈에는 하드 디스크 개념이 없기 때문에 이 부분에 대한 처리도 해야 한다. 보조기억장치로는 flash memory 가 사용되는데 본 모듈에서는 PC 상의 시뮬레이션 코드에서 읽어 들인 모델파일, 인식어휘파일, 문법정보파일들의구조체 정보를 PC 의 메모리 상태 그대로 16진수 바이너리 형태로 제 저장하여 flash memory 의 정해진 번지에 write 하였고 DSP 보드 부팅시 미리규약된구조체정보대로 flash memory 에서 읽어들이게구성하였다.

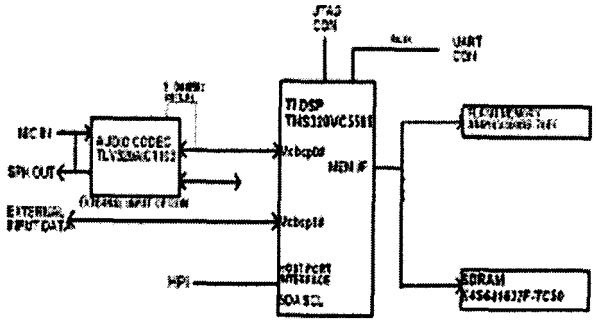

그림 1. 음성 모듈 블록 다이어그램

3. 연속 음성 인식

3.1 스트리밍 처리

보다 빠른 실시간 처리를 위해서 소리검출 후 인식을 수행하지 않고 스트림 단위로 인식을 수행하게 하여 발성하는 동안에인식을 수행하게하였다. 스트리밍처리의 관건은입력되는음성버퍼와동기를잘맞춰서특징추출해야하고 인식을수행해야 한다. 이는 특징추출시분석단위와 관계가 있다. 본 모듈에서는 25msec 의 hamming window 와 10msec 의 shift 를 하였다. 이에 녹음 블럭 버퍼의 단위를 560sample 로 하여 하나의 녹음 블럭 버퍼에데이터가가득차게되면 7 frame 의특징추출을수행할수있다.

560samples/80samples(8kHz 의 10msecshift)=7frame

즉, 7 frame 의 10차 LPCMEL 파라미터를계산하게 된다. 여기서 차분 성분을 계산하여 나머지 10차의 값을 계산하여야 최종적으로인식에 사용될수 있는데 차분성분은 Delta Window 로 5를 사용하였으므로 현재 frame 에서 앞뒤로 5개 frame 이있어야하는데현재계산된 7 frame 으로는앞뒤로 5개 frame 이확보되지않음으로첫번째와 두 번째프레임에 대해서만차분 성분이계산되고 인식에도첫번째와두번째프레임만인식을수행한다. 나머지 5개 frame은 다음 음성 블럭 버퍼가 입력될때 까지 메모리에저장만해둔다.

이후 입력된 스트림에서 7 frame 의 LPCMEL 10차가 계산되면 전체 14 frame 이 되므로 3 frame 에서 9 frame 까지 차분성분을계산할수있다. 이에디코딩도 3 frame 에서 9frame 까지수행할수있다. 이런 방식으로 입력되는 음성 버퍼에 대해서 인식을 수행하면 보다 빠른 인식처리를할수있다.

하드웨어 적으로는 DMA 와 MCBSP 를 이용하여 인식하는동안에도 DSP 에부하를최소화하면서인식과녹음을진행한다[3].

4. 음성 DataBase

4.1 음성 DB

음성 데이터베이스는 TLV320AIC1103 코덱을 사용하여 조용한 사무실 환경에서 DSP 보드로 직접 녹취하였다. 8Khz 8bit a-law 상태로 녹음을 하였고 모델링 단계에는 8Khz 16bit Linear PCM 으로 변환을 하여 사용하였다. 20대 ~ 30대 사이의 서울에 거주하는 대학생 위주로남자 200명, 여자 100명의 4연속숫자와 7~8 연속숫자음(지역번호, 휴대폰번호)을 수집하였다. 학습에는 전체 수집된 데이터의 90%를 사용하였고 10%는 테스트 데이터로사용하였다.

4.2 특징 파라미터

특징파라미터 추출에는 Q15 포맷으로 변환했을때정수 0으로 매핑되는 차수는 배제하여 LPCMEL 20차를 사용하였다. 1차 Power 에 로그를 씌우고 Normailize 한 Power 를 사용하였고, 9차의 LPCMEL과 차분 성분으로 10차를 사용하였다. 분석단위는 25msec 의 hamming window 와 10msec 의 shift 로 분석하였으며 Delta window 는 5를사용하였다.

표 1. 음성 데이터의 분석조건

주파수	8kHz
양자화	16bit
프레임 길이	25ms
프레임 주기	10ms
분석창	Hamming Window
특징파라미터	power 1차 + LPC-MEL cepstrum 9차+ delta power 1차 + + 차분성분 9차 = 20차원

4.3 음소/음절 모델

한국어 숫자음은 단음절이며 음운현상 및 연음현상이 많이 발생하여 인식 성능이 비교적 저조한 편이다[4]. 이에 본 논문에서는 기본 숫자 음소에서 반음절 모델을 '2' 와 '5' 의 숫자에 대해서 적용하고 나머지 숫자음은 음소 표기대로 사용하여 비교 실험하였고 선행과 후행 음절을 하나의 모델로 구성한 바이폰 형태의 음절 모델을 구성하여 성능을 비교 실험 하였다.

4.4 학습 알고리즘

모델링 방법으로는 HM-Net 방법을 사용하였다. HM-Net 모델링 방법은 연속적인 상태분할 방법으로 시간 및 문맥방향으로 상태를 분할하여 음향 모델의 정밀도를 높인 모델링 방법이다[5]. 학습에도 실수 연산이 대부분이다. 학습을하는것은 DSP모듈에서하는것이아니고

표 2. 기본음소와 2와 5에 대한 반음절 모델 확장

숫자음	기본음소	반음절
일	i h r	
이	ih	ih in
삼	s aa m	
사	s aa	
오	ao	ao an
육	j u g	
칠	c h i h r	
팔	p aa r	
구	g uh	
공	g ao ng	
영	jv ng	

표 3. 선행 후행 음절을고려한숫자음 '1' 음절 모델 구성의 예

숫자음 '1' 의 음절 모델 구성
/일-일/
/일-이/
/일-삼/
/일-사/
/일-오/
/일-육/
/일-칠/
/일-팔/
/일-구/
/일-공/
/일-영/

Linux 나 Unix 기반의 워크스테이션에서 대량의 음성 데이터및스크립트처리를하는것이대부분이기때문에 굳이고정소수점연산으로바꿀필요는없다. 기존에처리하는 방법대로 실수연산을 수행하여 최종적으로 모델을 만든후 Q-Format 에맞게실수값들을변환만해주면고정소수점 DSP 에서사용할수있는모델이된다. 단, 주의할사항은모델의분산값이소수점이하로너무많이떨어지는 것을막아야한다. Q15 포맷에서 소수점 이하 5자리 이하로 떨어지면 Viterbi 스코어 계산시 분모로 사용되는 분산값이 정수 영역에서 0가 되기 때문에 "Divided by zero" 에러를 초래 하게 된다. 모델의 정밀도는 떨어지게 되지만 고정소수점 프로세서를 사용하면 감수해야 하는 사항이다.

5. 결과

인식 실험은 오프라인으로 고정 소수점 형태로 시뮬레이션 한 코드로 PC에서 테스트 하였다. 사용한 데이터는 수집한 데이터의 10% 에 해당하는 남/녀 30명의 화자를 사용하였다. 표 4 는 모델 종류 별로 인식률을 비교한 것이다. 인식률은 발성 문장 전체가 맞는 경우의 인식률을 나타낸 것이다.

표 4. 모델 종류별 인식률

모델 종류	인식률
기본 음소	90.63%
기본 음소 + 반음절 확장	92.29%
음절 모델	92.92%

인식결과를 살펴 보면 기본 음소를 사용했을때는 90.63% 이었고 기본 음소에서 '2' 와 '5' 숫자음에 대해서 반음절 모델을 적용했을때 1.66% 의 인식률 향상이 있었고 음절 모델의 경우 기본 음소에 비해 2.29%의 인식률 향상이 있었다. 작성된 모델을 DSP 모듈에 올려서 온라인 테스트 한 결과 4연속 숫자음의 경우 평균 응답속도는 0.7초 정도이였으며 스트리밍 방식을 사용하였을 경우 발성이 끝나고 Pause time 이 지나면 거의 동시에 인식결과가 출력 되었다.

6. 결론

본 연구에서는 우리 일상 가운데 자리 잡은 휴대폰에서의 인식을 목적으로 먼저 DSP 상에서 음성인식을 구현하였다. 사용한 DSP는 TMS320VC5501(Fixed Point)이며 DSP 모듈을 통해 받은 데이터를 가지고 훈련시켰으며 고정 소수점 연산방식으로 음소, 음절의 모델에 대해서 높은 인식 성능이 나왔다. 41음소 트라이폰 모델로 HM-Net 인식 엔진을 사용하여 92.29%, 음절 바이폰 모델로 92.92%의 결과를 얻을 수 있다. 향후 본 시스템에서 자연스럽게 숫자의 조합을 말하더라도 인식할 수 있도록 모델 뿐 아니라 인식성능 향상을 위해 연구가 더 진행되어야 할 것이다.

참고문헌

1. "TMS320C5x User's Guide", Texas Instruments, 1997
2. "C++ Footprint and Performance Optimization", Sams Publishing, 2000
3. 정훈, 정익주, "TMS320C32 DSP를 이용한 실시간 화자종속 음성인식 하드웨어 모듈 구현", 음성통신 및 신호처리 워크샵, 15권 1호, 1998
4. 이기영, 최성호, 이호영, 배명진, "반음절 문맥종속 모델을 이용한 한국어 4 연숫자음 인식에 관한 연구", 한국음향학회지, 제22권 제3호, pp175 ~ 181, 2003
5. Se-Jin Oh, Chul-Jun Hwang, Bum-KoogKim, Hyun-Yeol Chung, and Akinori Ito, "New state clustering of hidden Markov network with Korean phonological rules for speech recognition," IEEE 4th workshop on Multimedia Signal Processing,pp.39-44,2001.