차세대 PC 환경에서의 음성합성 시스템 구현

박 혜미, 신 정훈, 홍 광석
성균관대학교 정보통신공학부

An implementation of Speech Synthesis system based on the next generation PC

Hye-Mee Park, Jeong-Hoon Shin, Kwang-Seok Hong
School of Information and Communication Engineering, Sungkyunkwan University
cheryred83@naver.com, only4you@chol.com, kshong@skku.ac.kr

요약

유비쿼터스 컴퓨팅 환경에서의 차세대 PC 는 다양한 입출력 장치를 이용하여 사용자에게 효과적으로 실제와 같은 정보를 제공하며, 사용자들의 편의를 고려해 웨어러블 형태의 플랫폼으로 발전하고 있다. 이러한 사용자 편의를 고려한 기술개발 동향(소형화, 경량화, 착용화)에 발맞추어 웨어러블 컴퓨팅 환경에서의 HCI 방안으로 음성 인식과 합성은 주요한 자리매김을 하고 있다.

본 논문에서는, 현재 정부에서 국가적인 차원으로 연구 개발 중인 차세대 PC 플랫폼 기반에서 음성합성 엔진을 구현하며, 구현상의 문제점 파악 및 개선사항에 대해 제안한다. 또한, 실질적인 구현 결과를 토대로 사용자 편의성 및 S/W 개발 환경을 고려한 차세대 PC 플랫폼의 개선사항에 대해 제안을 한다.

1. 서론

최근 컴퓨팅 기술과 통신, 가전기기 등의 융합화 현상은 컴퓨터 산업의 "범용 컴퓨터로부터 차세대 PC 와 같은 특화된 정보단말로의 전이" 형상을 보이고 있다. 이에 따라 국내 기업들은 PC 및 가전, 정보통신 분야 등에서 이미 상당한 수준의 제조 역량을 쌓으며 차세대 PC 를 유망품목으로 집중 투자를 하고 있다. 그 결과 차세대 PC 산업은 향후 국내 IT 산업에 주도적 역할을 할 것으로 예상되며 2010 년경에는 일반 소비자 시장을 겨냥한 제품 개발이 활성화되어 지금의 휴대폰처럼 정보생활 필수품이 될 것으로 전망된다.

차세대 PC 란 문서작성, 인터넷 검색, 데이터 관리 등에서 사용되었던 종래의 PC 와는 달리, 인간의 특성에 맞추어 정보이용 환경과 사용자 목적에 따라 특화된 기능과 형태를 가지는 네트워크 기반의 인간 중심 차세대 컴퓨터 디지털 정보기기를 총칭한다.[1]

이러한 차세대 PC 는 네트워크를 통한 자연스럽고 편리한 서비스 제공을 목적으로 PC 가 제공하는 웹, 전자메일, DB 검색, 멀티미디어 재생 등의 컴퓨터 처리능력이나 "성능 중심에서 사용자의 편의성에 초점"을 맞춘 웹패드, 웹폰, PDA, 웨어러블 플랫폼으로 발전하고 있다.[2]

또한 차세대 PC 는 모든 형태의 단말기가 네트워크에 접속돼 있어 누구든지 시간과 장소에 대한 제약 없이 다양한 서비스를 이용할 수 있는 유비쿼터스 환경을 지원하며 소형화, 착용화, 실감화, 지능화 추세로 나아감에 따라 플랫폼과 사용자간의 인터페이스 기술이 중요하게 부각되고 있다. 종래의 키보드, 마우스, 화면 중심에서 음성, 동작, 표정 등을 이용한 인간 친화적이고 지능화, 소형화된 다양한 인터페이스를 목표로 한다. 이에 본 논문에서는, 현재 정부에서

국가적인 차원으로 연구 개발 중인 차세대 PC 플랫폼 기반에서의 HCI 방안으로 TTS(Text-To-Speech : 음성 합성 시스템)을 구현하였으며, 추후 플랫폼 상에서 제공될 어플리케이션들과의 연동을 위한 인터페이스 방안을 제공하였다. 또한 시스템 구현 시 발생했던 문제점의 분석 및 해결 방안을 제시하였으며, 실질적인 구현결과를 토대로 사용자 편의성 및 S/W 개발 환경을 고려한 차세대 PC 플랫폼의 개선사항에 대해 제안하고자 한다.

2. 차세대 PC 플랫폼

현재 정부에서 국가적인 차원으로 연구 개발 중인 차세대 PC 플랫폼은 일상생활에서 언제 어디서나 사용할 수 있는 액세서리 형태의 손목시계형 웨어러블 PC 플랫폼으로 구현하기 위하여, 기술융합화에 의한 소형화, 경량화, 지능화를 목표로 한다. 또한, 차세대 PC 플랫폼은 보다 편리한 사용자 인터페이스를 위하여 오감인식 및 표현 기술이 가능하다는 점에 주목할 만하다. 그러나, 현재 이 플랫폼은 연구 개발 중이며, 완성된 형태의 플랫폼으로 제공되고 있지 않으므로 차세대 PC 플랫폼 상에서의 직접 TTS 의 구현은 어려운 실정이다. 이에 따라 본 논문에서는 차세대 PC 플랫폼과 사양이 비슷한 Motorola 의 i.MX21 Application Development System 기반의 TTS 를 구현하여, 차세대 PC 플랫폼의 개발 완료 후, 즉각적인 이식이 가능하도록 하였다.

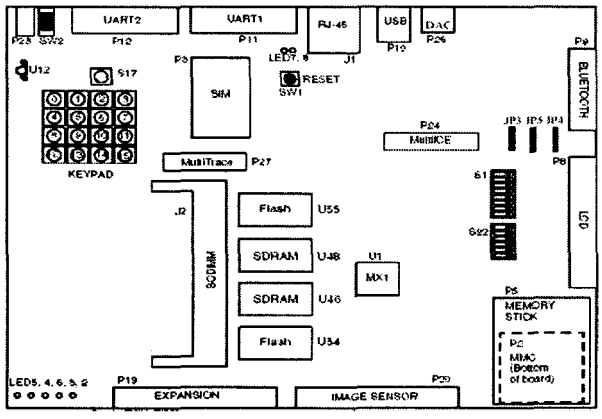

그림 1. i.MX21 Application Development System

그림 1 은 본 논문에서 구현된 TTS 의 기반이 되는 Motorola 사의 i.MX21 ADS Board 의 시스템 블록도이며 표 1 은 상세 기능 및 사양을 나타낸다.

표 1). i.MX21 ADS Board 의 상세 기능 및 사양

	참조모델 시스템
CPU	i.MX21
주메모리	SDRAM 64MB
시스템 설치	NOR FLASH 32MB
보조 저장 장치	NAND FLASH 32MB
CMOS Camera	VGA (30 만 화소)
LCD	26K Color TFT-LCD 640 × 480
무선 통신	Bluetooth using UART
Console	UART1(115200bps,8,N,1)
Sound	Microphone/Speaker
VGA Port	NTSC RCA Jack

3. 차세대 PC 기반 음성합성 시스템 구현

음성은 인간의 기본적인 의사소통 수단으로서 인간의 문화 발달에 필수 불가결한 것이다. 이러한 필수성과 편리성으로 인해, 음성을 이용한 HCI 구현 방안에 대한 연구는 지속되어 왔다. 본 논문에서는 기존의 PC 환경 음성합성 엔진을 착용형 기반 차세대 PC 로 재 구현하였으며, 실질적인 구현을 통해 보다 편리하고 효율적인 음성 합성 엔진으로 개선하였다.

음성 합성 기술은 합성음 크기에 따라 크게 두 가지로 나눌 수 있다. 자동 음성 응답 시스템(ARS : Automatic Response System)에 이용되는 제한된 어휘와 문장만 합성 가능한 제한 어휘 합성과 문자-음성 변환(TTS : Text-To-Speech) 시스템에 이용되는 임의의 단어, 문장을 입력받아 합성하는 무제한 어휘 합성으로 구분할 수 있다. 또한, 조음방식에 따른 구분으로, 인간의 발성기관을 모델링하여 음성을 합성하는 조음 합성 방식과 음원-필터 이론에 의하여 자연음의 포먼트와 같은 개수의 성도 공명 필터를 이용하는 포먼트 합성법, LPC 계열의 파라미터를 이용한 합성 방식, 미리 저장된 음성 데이터베이스를 이용하여

합성음을 만드는 연결 합성 방식 등으로 구분할 수 있다.[3][4]

본 논문에는 문자-음성 변환에 의한 무제한 TTS 를 구현하여, 차세대 PC 기반 플랫폼에 구현 되어진 다른 응용프로그램들과의 연동이 가능하도록 구현 하였다. 또한, 합성 방식으로는, 합성 단위 DB 의 구축후, 음성 분석을 통한 파라미터 DB 의 생성 후, 합성하고자 하는 문장의 합성 단위열을 생성, 미리 구축한 합성 단위 음성을 연결하여 합성음을 출력하는 연결 합성 방식을 적용하였다. 본 논문에서 구현한 합성 시스템의 합성음 생성 방식은, 실제 파형을 그대로 이용하여 음질을 높여줄 수 있으며, 음성 파형을 단순히 시간적으로 재구성함으로써 합성음 생성에 소요되는 시간을 단축할 수 있다.

3.1 차세대 PC 기반 음성합성 시스템의 구성도

본 논문에서 구현한 TTS 는 리눅스 기반 PC 환경에서 구현 후, arm 계열의 Cross Compiler 를 이용, i.MX21 ADS Board 상으로 이식 하였다. 그림 2 는 i.MX21 ADS Board 상의 TTS 및, 다른 편리한 UI(User Interface) 제공을 위해 구현 되어진 다른 응용프로그램들과의 상관도를 나타낸다.

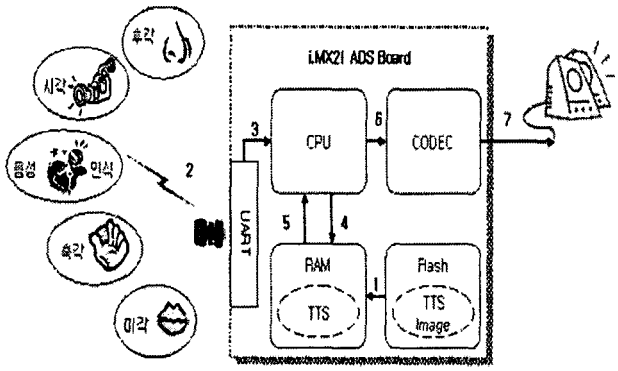

그림 2. i.MX21 ADS Board 기반의 TTS

또한, 그림 2 상에 구현되어진 TTS 는 다른 감각 정보를 표현하는 응용프로그램들과 연동을 고려하여 구현하였다. 각 감각별 제어를 위한 디바이스들과 플랫폼간의 통신은 Serial to Bluetooth 방식을 이용, 구현 하였다.

그림 2 에 나타난 바와 같이, TTS 의 Cross Compile 이 완료되면 Flash 메모리상에 TTS Image 파일을 다운로드 한 후, 저장 되어지며, 프로그램 실행 시 TTS Image 파일은 메모리로 Load 되어 CPU 의 제어를 받게 된다. 응용프로그램과의 통신을 통해 플랫폼이 메시지를 받게 되면, CPU 는 수신 되어진 제어 메시지를 이용하여 메모리에 상주되어 있는 TTS 를 제어하게 된다. TTS 를 통해 처리된 결과는 CODEC 을 거쳐 스피커로 출력되게 된다. 예를 들면, 얼굴 인식에 의해 상대방의 데이터를 화면에 표시하고, 이를 음성합성음으로 알려준다거나, 촉각 인식기를 통해 수화 등의 손동작을 입력 받아 음성합성음으로 알려주는 방식의 음성합성 시나리오 전개가 가능할 것이다.

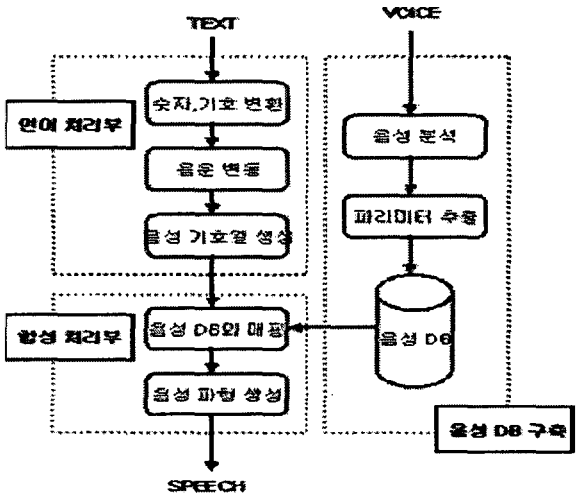

그림 3. TTS 의 블록별 구성도

본 논문에서 구현한 TTS 는 그림 3 과 같이 음성 DB 구축부 및 합성음 생성부로 크게 나눌 수 있다. 또한, 합성음 생성부는 다시 전처리에 해당하는 언어처리부와 실제 음성파형을 합성하는 합성처리부로 구분할 수 있다.

3.2 합성음 DB 구축

본 논문에서 구축한 합성음 DB 는 반음절 단위로 구성 되어 있다. 한국어의 기본 음절은 초성, 중성, 종성으로 구성되며, 이를 반음절 단위로 재 배치시, 초성+중성, 중성+종성의 두 가지 형태로 생성 가능하다. 이러한 방식의 반음절 DB 구축은 이론적으로 무제한 합성이 가능하다. 한국어 기본 음절은 초성 19 개, 중성 21 개, 종성 27 개로 구분 가능하지만, 실질적인

발성을 고려시 초성 19 개, 중성 17 개, 종성 7 개로 간략화 할 수 있다.

3.3 언어처리 및 합성처리

언어처리부에서는 한글 이외의, 숫자나 특수기호가 입력될시 해당하는 한글로 변환시켜 주며, 변환된 문장은 한국어 표준 음운 규칙에 따라 실제 발음상의 음운으로 다시 변환 하였다. 변환 후 생성된 음성 기호열은 합성처리부에서 실질적인 음성 파형으로 합성된다. 이때 DB 와 매핑된 각각의 반음절 파형의 합성방식은 TD-PSOLA 방식을 이용하였다.

TD-PSOLA 방식은 음성합성 시 음성을 피치주기단위로 분석한 각 프레임을 피치 주기에 동기 시켜 중첩하여 더하는 합성방식이다.[5] 이 방식의 장점으로는, 프레임을 중첩시키고 더하는 간격을 조절함으로써 피치주기를 제어할 수 있으며, 시간 영역에서 연산을 수행 하므로 실시간으로 동작 가능 하다.

4. 문제점 및 해결방안

본 논문에서 구현한 착용형 PC 기반 플랫폼인 i.MX21 ADS Board 의 메모리 용량은 64MB 로, 기존의 범용 PC 에 비해 많은 제약을 가진다. 본 논문에서는 이러한 메모리 제약을 고려하여, TTS 의 용량을 최소화 하였다. 이를 위하여, 음질의 저하를 감수하고 DB 용량을 줄일 수 있는 반음절 단위의 합성 방식을 이용하였으며, 이에 대한 보완책으로 TD-PSOLA 합성 방식을 적용, 음질 저하를 개선하고자 했다.

또한, TTS 에 입력된 문장은 완성형 코드체계를 사용하고 있으나, 반음절 단위의 합성을 위해서는 초성,중성,종성의 분리가 가능한 조합형 코드 체계를 사용해야 한다. 따라서 완성형 코드의 조합형 변환 과정이 필요했다. 이의 해결을 위하여, 리눅스에서 지원하는 라이브러리를 이용할 수 있으나, 이를 이용하면 플랫폼으로의 이식 과정에서 완성형과 조합형 코드의 Character Set 이 모두 플랫폼에 이식되어야 하며, 이 방법은 메모리 용량과 관련된 또 다른 문제를 야기 하였다. 따라서, 본 논문에서는 기본제공 라이브러리를 사용하는 대신 자체 제작한 알고리즘을 이용하여 완성형에서 조합형으로 변환 후, 초/중/종성으로 나누는 작업을 수행 하였다.

차세대 PC 기반 플랫폼상의 TTS 는 오감의 연동, 융합 및 재현을 위하여 다른 응용프로그램과의 연동을 고려하여야 한다. 이에 본 논문에서는 TTS 를 Server/Client 방식을 이용, 연동방안을 제공하였다. TTS Server 는 다른 응용프로그램으로부터의 메시지를 받을 때까지 대기상태를 유지하며, 메시지 수신시 TTS Client 에 수신 메시지를 전송하여준다. 메시지를 수신한 Client 는 수신 메시지를 처리 후, 생성된 음성파형을 서버로 재전송하며, 이를 Server 가 출력하게 하는 방식을 이용하였고, 출력 파형을 파일로도 저장하여 필요시 다른 응용프로그램에 전달할 수 있도록 구현하였다.

5. 결론

본 논문에서 구현한 TTS 는 적은 용량으로 보다 나은 음질의 합성이 가능한 것으로, 추후 실생활에 널리 사용될 차세대 PC 플랫폼 상에서의 HCI 방안으로 사용자의 편의성을 도모할 수 있다. 또한, 다른 감각과의 연동방안을 제공하고 있다.

향후에는 용량을 최소화하면서도 고품질의 음질을 얻을 수 있는 새로운 DB 구축 방법에 관한 연구가 추가적으로 필요하며, Server/Client 방식의 확장 구현도 추가적으로 연구되어야 할 것이다.

참고문헌

1. 정보통신진흥원, IT 차세대 성장 동력 기획보고서(차세대 PC), 2004. 4.
2. 조일연, 박준석, 한동원, "웨어러블 컴퓨팅을 위한 서비스 인프라 구조", 인간공학회, 2004. 4.
3. J. Allen, M.S. Hunnicutt, "From text to speech : The MITalk system", 1987
4. F.Fallside, W.A Woods, "Couputer Speech Processing", Prentice Hall, 1985
5. Jon R. W. Yi, "Time-Domain PSOLA Concatenative Speech Synthesis Using Diphones"