제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

T2S(Text-To-Speech) 시스템 구축을 위한 분석 및 설계

김정형*, 이왕헌*, 이현창*
*한세대학교 IT학부
e-mail : {windymail,whlee,hclee}@hansei.ac.kr

Design and Analysis of T2S System Implementation

Jung-Hyung Kim*, Wang-Hun Lee, Hyun-Chang Lee*
*Division of Information Technology, Hansei University

요 약

IT 정보 기술 발달로 관련 하드웨어와 소프트웨어의 보급이 일반화 되며, 그에 따라 신체적 장애를 겪고 있는 사람들에게 IT기술을 이용하여 신체적 결함을 극복할 수 있는 정보통신 응용제품은 필수적이다. 특히, 고령화 사회로 접어들면서 신체적 기능저하들 중에서 시각 기능 저하도 대표적인 부분이다. 이러한 시각장애를 겪고 있는 사람들을 위한 정보전달 수단으로 점자책등이 존재한다. 그러나 일반 서적에 비하면 이용 및 활용을 위한 제반 기술이 상당히 부족한 현실이다. 이에 본 연구에서는 시각 장애를 겪는 사람 및 장애자들에게 일반 책을 읽을 수 있도록 오픈 소스 기반에 소형 스캐너를 부착한 웨어러블(wearable) PC를 직접 제작하여 개발 완료시점에 있는 시스템 내용에 관한 분석 및 설계에 관한 내용이다. 이를 위해 본 연구에서 일반 스캐너 내부 구성을 살펴보며, 책 혹은 정자로 주어진 문서를 실시간으로 스캐닝을 통해서 글자를 추출하고, 추출된 글자를 음성으로 들려주는 휴대용 통합시스템(T2S:text-to-speech)의 개발 진행된 연구에 관하여 살펴본다.

1. 서론

정보통신기술의 발달은 정보통신 관련 하드웨어와 소프트웨어의 보급을 일반화하였으며 보급된 정보통신기기는 의사소통과 정보획득을 위한 중요한 수단이 되었다. 사회의 정보화는 산업현장 뿐만 아니라 가정에서도 컴퓨터와 인터넷을 통해 다양한 정보를 획득할 수 있게 하고 상품과 서비스를 구입할 수 있게 해주며 생활의 여유를 즐길 수 있도록 해주는 등 사람들의 생활에 커다란 변화를 가져왔다.

이러한 서비스의 다변화 상황 속에서 신체적 결함을 극복, 보완하기 위하여 서비스 활용을 용이하게 해줄 수 있는 정보통신 제품은 장애인들에게 필수적이다. 이러한 이유로 장애인용 특수정보기기의 개발과 보급에는 정부의 정책적, 경제적 지원이 반드시 필요하다. 미국이나 유럽 등 선진국에서, 장애인 정보통신 접근이 장애인의 삶을 질적으로 향상시킬 수 있다는 인식아래, 장애인용 정보통신기기의 개발과 보급에 많은 지원을 하고 있는 것은 많은 시사점을 던져준다[1].

시각장애인들에게 정보 전달매체는 점자책이 있지만 일반 책에 비해 부족한 것이 현실이다. 이에 시각장애인이 휴대할 수 있으면서, 시간과 장소에 제약받지 아니하고 언제 어디서든 일반 책을 스캐닝 이후 글자 추출과 추출된 글자를 음성으로 읽을 수 있게 해주는 T2S (text-to-speech)시스템의 개발이 필요하게 되었다.

T2S 시스템 개발에서 데이터 추출에 꼭 필요한 스캐너의 소형화는 스캐너를 휴대해 어디서든 문서를 스캔 및 저장 할 수 있게 되었다. 또한, 음성합성 기술은 많은 발전을 거쳐 자연스러운 합성음을 생성할 수 있게 되었으며, 그 기술을 이용한 국내 제품으로 매직보이스, 보이스 텍스트, 음성마법사, 나랏소리 등이 있으며 국외 제품으로는 Festival, Emacspeak, VoiceText등의 오픈프로젝트 등의 제품이 있다. 본 연구에서는 이와 같은 T2S 시스템 구축을 위한 분석과 설계를 살펴본다.

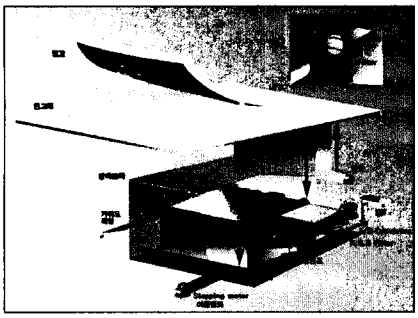
본 논문의 구성은 다음과 같다. 2장의 관련 연구에서는 스캐너와 음성합성에 대해 고찰하고, 3장에서 T2S시스템의 구성 및 기능에 대하여 분석 하였다. 제4장에서는 구축

모델에 대한 평가와, 마지막 5장에서 결론 및 향후 연구 방향에 관하여 살펴본다.

2. 관련연구

2. 1 스캐너

스캐너는 스캔할 대상에 빛을 반사시킨 다음 빛을 투과시켜 각 부분의 반사도에 대해 데이터를 수집하고 수집된 데이터를 전기 신호로 변환한 다음 디지털화하여 전송하게 된다.



(그림 1) 평판 스캐너의 동작 과정

(그림 1)은 평판 스캐너의 기본 동작 과정을 보여주고 있다. 대부분 원본에서 반사된 빛을 한줄 한줄 측정하고, 반사광을 집광 렌즈를 통해 전하 결합 소자(CCD, charge-coupled device)로 모은다. 해상도는 CCD의 광센서 개수와 스캔하는 라인 간격에 따라 결정된다. 트랜스패런시(투명지) 스캐너는 원본에서 반사되는 빛이 아니라 투과하는 빛을 탐지하여 측정하게 된다[2].

2. 2 음성합성

초기의 음성 합성 시스템은 전기적 장치를 이용한 포만트 합성기였으며 두 개의 공진회로를 버저(buzzer)에 의해 여기 되도록 만들어졌다. 이 시스템은 모음의 안정 구간을 두개의 포만트 주파수 즉 F1, F2를 공진 회로로 모델링 하였으며, 이에 따라 모음에 근사화된 합성음이 생성된다. 1939년 벨 연구소의 Dudley는 분석/합성법을 제안해 음성을 인위적으로 가공할 수 있는 방법을 열어 놓았다. 즉 Voder라는 합성 시스템은 유/무성에 따른 음원을 선정해 있으며, 피아노 건반과 같은 키보드로 10개의 대역필터로 구성된 공진회로의 출력을 조절한다. 또한 음의 높이인 기본 주파수를 페달로 조절하게 되어 있다. 이 시스템으로 연속음을 합성하기 위해 오퍼레이터를 1년 이상 훈련시켜야 했다고 한다.

1951년에는 Haskins lab에서 스펙트럼 패턴을 음성으로 변환하는 'Pattern Playback' 합성기를 개발했다. 이 시스템은 스펙트럼이 그려진 투명한 벨트를 광학적 장치에 의해 읽혀져 음성 발생 과정과 역으로 음성이 합성된다. 1960년 음성 합성은 Fant에 의해 새로운 전기를 마련하게 된다. 지금까지 합성은 주로 시간에 따라 변화해가는 음성의 스펙트럼을 그대로 이용하는 방식이었으나 Fant는 '음성이 어떻게 발생되는가'하는 음향 이론(acoustic theory)에 바탕을 둔 발생학적 모델링을 제안했다. 즉 음성의 발생은 여기 신호와 선형 필터로 모델링(source-filter theory)할 수 있는데 여기 신호는 유/무성으로, 선형 필터는 입술, 구강(oral cavity), 인두강(pharynx cavity)으로 구성된 성도의 공명 효과를 모델링 한다.

이와 같이 source-filter theory에 의한 선형 필터 즉, 성도 전달 함수는 모두 극점(pole)으로 모델링 되며 비음과 같은 음성은 영점(zero)을 첨가해 근사적으로 비강(nasal cavity)을 모델링하게 된다. 이 모델에 따라 최초의 병렬 포만트 음성 합성기인 PAT(Parametric Artificial Talker)와 직렬 포만트 음성 합성기인 OVE I(Orator Verbis Electris)이 개발됐다. PAT 음성 합성기는 세 개의 공진회로가 병렬로 연결되어 있으며 각 공진기의 출력을 단순히 더하여 합성된다. 이 시스템은 세 개의 포만트 주파수, 유/무성 크기, 기본 주파수 등 6개의 제어 파라미터를 갖는다. OVE I은 공진회로가 직렬로 연결되고 두 개의 포만트 주파수, 유/무성 크기, 기본 주파수 등의 제어 파라미터로 모음과 같은 음성만 합성한다. 이후 두 시스템은 파찰음, 비음 등을 모델링한 공진 회로를 추가함으로서 자연스러운 합성음을 생성하게 된다.

1968년 디지털 컴퓨터를 합성에 이용함으로서 전자 회로에 의한 합성기는 쇠퇴하기 시작했다. 그리고 합성 방식도 병렬 포만트 방식과 직렬 포만트 방식이 결합하게 되었고, 유성음화된 마찰음과 같은 세밀한 음성도 모델링 되었으며, 다양한 제어 파라미터도 추가됐다. 1973년에 Holmes는 기존의 병렬 포만트 방식과 음원 모델을 사용해 매우 자연스러운 음성을 생성할 수 있는데, 이 시스템은 1984년에 실시간으로 동작되는 칩으로 구현된 바 있다. 1984년 새로운 음원 모델을 사용한 Infovox SA-101 음성 합성기가 개발됐고, 1985년에는 Klatt가 수학적 모델에 의한 음원을 적용했으며, 1986년에는 Fujisaki가 영점을 첨가한 음원 모델을 제안했다.

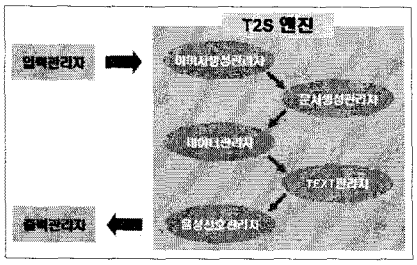
한편 성도를 튜브로 단순화해 모델링하고, 이 튜브를 여러 개의 작은 부분으로 세분화한 다음, 공기의 체적 속도나 압력의 분포 등을 전기 회로로 근사화한 조음 합성기(articulatory synthesizer)가 개발됐다. 이후 이 모델에 유/무성 회로, 비강에 해당하는 회로가 첨가되었고 1975년과 1985년에는 주파수에 따른 영향, 저주파에서 성도 벽면의 움직임이나 성문에서 time-varying impedance를 모델링해 성능을 개선시켰다. 또한 음성 신호 분석/합성에 의한 선형예측 방법이 Itakura, Atal 등에 의해 소개되어 오늘날 대부분의 분석/합성 시스템에 사용되고 있다. 이 방법을 이용해 합성단위를 미리 저장하고 합성시 이 단위를 연결하여 합성하는 방식이 무제한 음성합성기의 주된 방식이다. 최근에는 메모리에 대한 제약이 없어져 시간영역에서 합성하는 TD-PSOLA 방식이 제안되어 요즈음 널리

이용되고 있다.

국내에서는 1990년대 들어 포만트 합성법을 이용한 한국어 규칙합성시스템의 구현에 관한 연구 및 반음절 데이타베이스를 이용한 MPLPC(Multi-Pulse Linear Prediction Coder) 무제한 단어 합성기가 학계에서 개발되었고, 업계에서는 LPC를 이용한 무제한 합성기(명칭: 가라사대)가 PC 플랫폼에서 하드웨어로 개발되어 국내 최초로 상용화된 바 있다. 한편 한국전자통신연구소에서는 LSP(Line Spectral Pairs)와 반음절 데이타베이스를 이용한 합성시스템(글소리-I)을 개발, 이를 교환기의 오디오텍스에 적용한 바 있다. 특히 1992년에 기존의 분석합성법과 다른 TD-PSOLA 방식을 적용하여 매우 명료한 합성음을 생성할 수 있으며, 현재 이 시스템(글소리-II)은 Windows95, UNIX 플랫폼에서 소프트웨어만으로 구동되어 다양한 서비스 개발이 가능하게 되었다.

3. T2S 시스템 구축을 위한 분석 및 설계

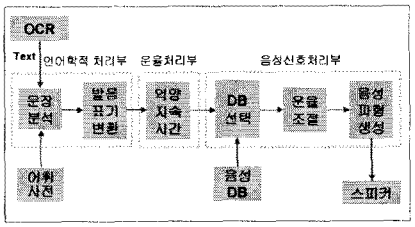
3.1 T2S 전체 시스템 구성



(그림 2) T2S엔진의 전체 구성

본 연구에서는 향후 개발하게 될 text-to-speech 시스템의 분석 및 설계 단계이며, 시스템의 전체 구성도는 (그림 2)에 T2S 엔진 전체 구성을 그림으로 도시하여 나타내고 있다. 먼저, 입력 관리자는 스캐너를 통해 들어온 신호를 받아서 이미지생성 관리자에게 신호를 보내준다. 이미지생성 관리자는 전송된 신호를 하나의 이미지 파일로 생성한 후 생성된 문서를 문서생성 관리자에게 넘겨준다. 문서생성 관리자는 이미지파일에서 문자를 추출해 text파일을 생성한다.

문서생성 관리자는 순차적으로 생성된 text파일을 데이터관리자에게 전달하고, 데이터관리자는 이어서 들어오는 2개의 text파일에서 중복된 부분을 제거하여 한 개의 완성된 text파일을 생성한다.

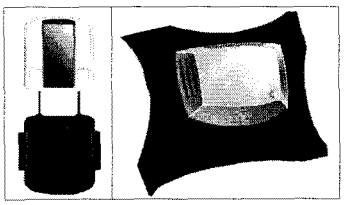


(그림 3) TTS의 구성

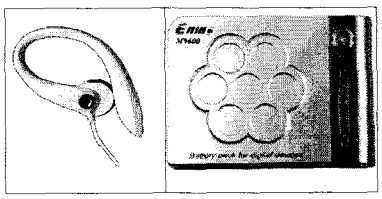
데이터 관리자로부터 생성된 완전한 text 파일은 TEXT 관리자로 이관되며, TEXT관리자는 만들어진 text 문서를 분석해 문장의 정보와 어절, 억양 등의 정보를 추출해 음성신호 관리자로 보내준다. 음성신호관리자는 추출된 정보에 맞는 음성을 합성 출력한다.

3.2 T2S 개발 환경 분석 및 설계

본 연구에서 제시하고 있는 스캐너는 손목에 탈착 가능한 휴대용 스캐너를 사용하며, 메인보드는 SBC-C677VRL(ACTAC Technology Corp)를 이용하고, 백팩 형식으로 케이스를 제작하여 등에 부착할 수 있게 설계하였다. 메인 메모리는 2기가의 플래시 하드를 사용한다. 배터리는 Enine M9600를 이용하고, 이어폰은 귀걸이형 이어폰을 연결해서 사용하며, 시각장애인의 편의를 위해 전원과 동작 2개의 버튼으로 단순화함으로서 편리성을 추구할 수 있도록 구성하고 있다. 볼륨 조절은 기존 사용방법으로 회전형 버튼을 이용한다. 운영체제는 오픈소스인 리눅스 커널 2.4.1을 이용하며 TTS 시스템은 festival-2.0, OCR 시스템은 gocr-0.40을 사용한다.



(그림 4) 스캐너(좌)와 메인보드(우)



(그림 5) 이어폰(좌)과 베터리(우)

4. 구축 모델에 대한 평가

IT기술 발전은 정보통신기기를 이용한 의사소통과 정보획득을 위한 중요한 수단이 되었다. 이러한 상황에서 신체적 장애를 보완하여 정보에 활용을 용이하게 해줄 수

있는 정보통신 제품은 장애인들에게 필수적이다. 이에 본 장에서는 향후 개발하게 될 T2S 제품의 분석 및 설계를 통해 얻을 수 있는 장단점을 살펴보고자 한다.

4.1 T2S 시스템의 평가 분석

T2S 시스템은 신체(손)에 부착된 스캐너를 이용하여 문서를 Scanning한다. 스캐닝한 결과로 생성되어진 이미지 문서에서 글자를 추출하여 각 글자에 상응하는 음성으로 시각장애인에게 실시간으로 음성 서비스를 제공할 수 있게 된다. 이와 같은 절차를 통하여 구축되는 시스템은 스캐너, 문자추출, TTS 등의 기능이 하나의 시스템에 통합된 시스템이기 때문에 각각의 기능별 장치를 사용하는 것보다 실시간이면서 빠른 응답을 얻을 수 있게 된다. 또한 개인 휴대성이 뛰어나기 때문에 백팩의 형식으로 부착이 가능하여 일상 생활 및 활동에도 지장 없이 언제 어디에서든지 활용이 가능하다.

이와 같은 장점을 지니고 있지만 스캔하는 속도, 기울기 등의 영향이 스캔율에 영향을 크게 끼치기 때문에 스캔율이 좋지 않을 경우 문자판독 또한 정확성이 떨어지게 된다. 그 외에 해결되어져야 할 사항중에 TTS의 발음이 자연어를 사용하는 인간의 발음에 크게 미치지 못하기 때문에 이에 대한 기술 개발이 필요하다.

4.2 기대효과 및 응용분야

본 연구에서 제시하고 있는 T2S 시스템이 개발될 경우 시각장애인들은 타인의 도움 없이 일반인과 유사하게 일반 문서를 읽을 수 있게 되어 더욱 많은 정보를 손쉽게 획득 할 수 있게 된다. 더욱 특징적인 부분은 일반인들이 오히려 어두운 부분에서 책을 판독하기 어렵지만 본 연구의 결과를 응용할 경우 어두운 공간에서도 문서를 파악할 수 있을 것이다. 그 외에도 인쇄물의 빠른 Text화로 eBook, VoiceBook 등을 제작할 수도 있으며, GPS와 연결하여 음성 네비게이션을 구축할 수도 있게 된다.

5. 결론 및 향후 과제

시각장애인들에게 정보 전달매체로 점자책이 있지만 일반 책에 비해 턱없이 부족한 것이 현실이다. 이에 시각장애인이 휴대할 수 있으면서, 시간과 장소에 제약받지 아니하고 언제 어디서든 일반 책을 스캐닝 이후 글자 추출과 추출된 글자를 음성으로 읽을 수 있게 해주는 T2S (text-to-speech)시스템의 개발이 필요하게 되었다.

이에 본 연구에서는 향후 개발하게 될 T2S 시스템 구축을 위한 분석과 설계를 살펴보았다.

(그림 6) T2S 예상 모형도

(그림 4)는 개발하게 될 시스템의 예상 모형도를 도시하였으며, 본 시스템의 특징으로 항상 몸에 지니고 다녀야 하기 때문에 휴대성과 디자인적인 측면도 강조되어 개발되어야 필요가 요구된다.

참고문헌

[1] http://ksc.digitalsme.com/club/club_board_view.
[2] http://blog.naver.com/pr2780?Redirect=Log&log
[3] http://www.ubiu.com
[4] http://www.actac.co.kr
[5] http://www.maxan.com
[6] http://www.actac.com.tw/Show_product.asp?id=623
[7] http://kelp.or.kr/korweblog/stories.php?story=05/10/20/4723113
[8] http://kelp.or.kr/korweblog/stories.php?story=04/04/28/7166008
[9] http://www.aleph1.co.uk/taxonomy/term/31/
[10] http://www.linux-usb.org
[11] http://www.linux-mtd.infradead.org
[12] http://blog.naver.com/come2alex/10000941689
[13] http://www.epson.co.kr
[14] 이정철, 이상호, "교육용 한국어 TTS 플랫폼 개발", 말소리 제 50호, p41-50, 2004
[15] 한국과학기술원, "한국어 텍스트에서의 문장구조 추출 도구 개발"