무제한 음성합성기를 위한 음성분석장치

°김 재 인*, 김 진 영**, 이 종 락***
*한국통신 연구개발원 소프트웨어연구소 음성언어연구팀
**전남대학교 공과대학 전자공학과
***한국통신 연구개발원 통신시스템개발센터

Speech Analysis Tools for Text-to-Speech Synthesizer

°Jae In Kim*, Jin Young Kim**, Jong Rak Lee***
* Spoken Language Research Team, Software Research Laboratory, Korea Telecom Research Laboratories (02)526-6571
** Dept. of Electronic., Chonnam National Univ.
*** Systems Development Center, Korea Telecom Research Laboratories

요 약

본 논문에서는 무제한 음성합성기를 구현하기 위하여 꼭 필요한 음성분석장치의 개발에 대하여 논하였다. 이 분석장치는 신호처리 보드를 사용하여 PC에서 사용할 수 있도록 되어 있으며, 음성의 A/D, D/A 및 spectrogram display는 물론 pitch pulse위치를 Glottal Instant Closure(GCI)에 맞추어 삽입할 수 있어 Linear Prediction(LP) base의 무제한 합성기에서 필요한 음성 Data Base(DB)를 구축하기 용이하도록 개발하였다. 또한 음성인식을 위한 음성DB나 현재 사용중인 ARS에서와 같이 단어나 문장단위 음성을 저장하였다 재생하는 방식을 사용하는 합성기에 필요한 많은 양의 음성DB를 구축하고자 할 때에도 적은 노력과 시간이 소요되도록 하였다.

I. 서론

무제한 음성합성기의 개발 시에는 합성에 사용될 합성단위에 대한 연구가 선행되어야 한다. 선행연구시에는 경우에 따라서 수많은 스펙트로그램들을 실시간으로 변화시 관측하여야 하므로 한 화면을 그리는데 1초미만의 시간이 소요되는 장치를 이용하여야 효율적이다. 이러한 선행 연구가 끝나고 결정된 합성단위들을 녹음된 단어에서 분리해내는 작업은 여러 사람이 동시에 할 수 있으므로 가격이 저렴하고 사용이 편리한 음성분석장치가 필요하다. 이러한 장치들은 외국의 몇몇 회사들에 의하여 개발되어 상용화된 것들이 있으나 워크스테이션에서 동작되는 것들은 가격이 비싸서 여러 개를 한꺼번에 이용하기 곤란하여 짧은 시간 내에 DB구축작업을 끝내기 어렵고, 독립적으로 동작되는 기기 역시 비슷한 사정이다. 그리고 가격이 저렴하게 PC에서 사용할 수 있게 개발된 것들도 있기는 하지만 드물고 있는 것도 editing기능이 미흡하다. 이제품 워크스테이션용이나 PC용이나 상용 S/W를 그대로 이용하기가 어려운 실정이다. 그래서 가격적으로 저렴하고 기능 향상이 항상 가능하도록 PC상에서 동작되는 음성분석 S/W를 개발하였다.

II. H/W 환경

A/D, D/A 및 FFT를 하는데는 PC에서 동작되는 Digital Signal Processor(DSP)가 탑재된 보드를 사용하였다. DSP로는 TI사의 TMS320C31이 사용하였고 A/D 및 D/A를 위해서는 TLC32044C를 이용하였다[1]. 사용된 보드의 블럭도는 그림 1과 같다.

A/D와 D/A를 위하여 사용된 TLC32044C(Voice-band Analog Interface Circuits) 칩에는 14 bit resolution의 A/D, D/A convertor와 pre와 post 필터 링을 위한 switched capacitor filter가 각각 내장 되어있으며, 원하는 sampling rate에 따라 내부 register(RA, RB)의 값을 S/W로 조정하여 사용할 수 있게 되어 있다. 그러나 8Khz, 12Khz, 16Khz로 정확히 A/D를 하기 위해서는 관련된 계산 식에 의하여 master clock을 6.912MHz를 사용하여야 하나 사정상 사용한 보드에서는 master clock으로 6MHz를 사용하였기 때문에 이에 따른 실제 conversion clock과 SCF(Switched Capacitor Filter)용 clock주파수는 <표1>과 같아된다.

III. 음성분석용 S/W

한국통신에서 개발한 무제한 음성합성기인 "한소리"는 반음소를 합성단위로 사용하고 있기 때문에 spectrogram과 음성파형이 시간적으로 정확히 일치가 되어야 하며, sample단위까지 정확하게 분리할 수 있어야 한다[2][3]. 또 유성음 부분에서 pitch pulse가 위치가 GCI와 일치하도록 삽입이 가능하여야 하며 무성음과 유성음의 위치를 정확하게 설정할 수 있도록 하는 기능이 있어야 한다. 본 논문에서는 새로운 알고리즘등에 대한 설명보다는 구현된 결과에 대하여 설명하기로 한다.

3.1 화면 구성

그래픽은 VGA표준인 640x480의 해상도와 16color를 기준으로 만들었다. 화면의 구성(그림2 참조)에 대하여 알아보면 화면의 최상단에는 메뉴판이 위치하여 있고 그 밑에는 현재 active되어 있는 bar를 기준으로 전후 320sample의 그림이 샘플단위로 그려져 있다.

음성편집용 tool에서는 좌우 두개의 bar를 사용하고 있으며 active 상태의 bar의 색깔은 붉은 색으로 나머지는 녹색으로 나타나도록 하였다. Active bar의 선택은 'space bar'를 누르거나 마우스의 좌, 우측 버튼을 누르면 바뀐다. 다음에는 configuration에서 정의된 분석구간의 파형의 에너지, zcr(zero crossing rate), 에너지, spectral tilt와 Er/Es 즉 residual 신호의 에너지와 파형의 에너지의 비율이 표시되어 있는데 Er/Es는 LP분석이 된 후에만 표시 된다. 그 아래에는 설정되어 있는 block을 기준으로 축소된 파형이 그려져 있으며 다음으로 그려진 파형과 분석결과에 대한 수치가 차례로 표시되어 있고 현재의 위치와 bar사이의 시간 폭이 나타나 있다. 그 아래에는 spectrogram이 그려져 있으며 화면의 제일 하단에는 프로그램의 수행결과를 나타내거나 labelling되어 있는 결과가 표시된다.

3.2 Data 분석

음성을 분석하는 필요한 파라미터들로서 spectrogram과 영교차율, 파형의 에너지, 잔차신호의 에너지와 파형 에너지비, 그리고 주파수에너지의 비인 spectral tilt등을 구하였다.

3.2.1 FFT 분석

입력된 음성에 대한 주파수 성분은 DSP보드에서 512 points FFT를 수행시켜 구하였다. DSP로 전송된 data들은 먼저 pre-emphasis를 한 뒤 hamming windowing을 한다. 분석구간이 512 sample 보다 적으면 나머지는 zero을 padding하였다. 수행된 결과인 256개의 data중 화면의 크기에 때문에 전부 나타내지 못하고 192개만을 선택하였다. 선택된 data들은 13단계의 gray level로 변환되어 화면에 그려졌다. 그리고 spectral tilt를 계산하기 위하여 1000Hz이하의 에너지와 그 이상의 주파수에 대한 에너지는 각각 보드에서 계산하도록 하였다.

음성편집용 tool의 화면 해상도는 표준VGA인 640x480을 가지고 있기 때문에 그림이 그려진 부분에 글씨를 쓰게 되면 그림보다 상대적으로 커 보이기 때문에 화면을 가리게 된다. 그래서 spectrogram을 나타내는 화면에 가로 세로에 scale에 대한 표시를 하지 못하였다. 화면에서 spectrogram을 그리는 부분에 가로축은 시간(ms)을 나타내고 세로축은 주파수를 나타낸다. 그래서 그 대안으로 화면 중간에 좌우bar사이의 시간과 처음부터 해당 bar까지의 시간을 ms로 표시하고 있으며 마우스를 사용하여 click한 부분에 대하여 작은 가로bar를 그려서 표시하고 그곳에 주파수 값을 화면 중앙 우측에 표시하였다.

3.2.2 Er/Es

이것은 선형예측(Linear Prediction) 분석후 생성되는 residual data들의 energy와 파형의 에너지와 비를 나타내는 것으로 비음구간을 구분하는 데 사용할 수 있다.

3.2.3 Spectral Tilt

Spectral tilt는 FFT를 하여 구한 Spectrum에서 1 Khz이하의 성분에 대한 energy와 그 이상의 주파수 성분에 대한 energy의 비를 계산한 것으로 각음의 경우 발음시 시작에서 보이는 click을 구분하는 데 사용할 수 있다.

3.3 편집기능

편집기능이란 음성 파형을 자르고(Cut), 복사하고(Copy), 붙이고(Paste), 더함(Add) 할 수 있는 기능들을 말한다. 이들 기능 이외에 bar사이의 data들을 뒤집거나(inversion) bar사이의 data들의 gain을 조절할 수 있는 기능도 있다.

Editing하는 방법은 마우스나 키를 사용하여 분리해 내고 싶은 부분에 bar를 위치시킨 후, 화면의 상단으로 가서 최종적으로 분리하고 싶은 포인트를 지적한다. bar 왼쪽을 사용 중 일 때 왼쪽 마우스 버튼을 click하면 그 점에서 점선이 새로로 표시되고, space bar를 눌러 오른쪽 bar가 움직이도록 한 후 오른쪽 분할부분을 마우스의 오른쪽 버튼으로 누르면 왼쪽의 경우와 마찬가지로 점선으로 된 bar가 새로로 그려진다. 이렇게 하면 sample 단위로 정확하게 분리해내고 싶은 곳 지정할 수 있다.

3.4 피치펄스위치 표시 기능

피치펄스위치는 GCI와 일치하여야 하기 때문에 pitch 펄스 위치 편집기능에 들어가면 자동적으로 LP분석후 생성된 residual 신호를 이용하여 기준이 되는 신호를 만들어 화면에 표시한다. 수동으로 삽입시 원 파형과 residual 신호 그리고 만들어진 기준신호를 참조하여 pitch 펄스 위치를 입력하면 된다. 하지만 녹음된 음성 중에서 필요한 부분에 대한 것은 이미 spectrogram을 참조하여 표시하였기 때문에 표시된 구간 중에서 유성음 부분에만 pitch 펄스 위치를 삽입하고 무성음 부분은 따로 text로 표시 할 수 있다. 표시된 유, 무성음 구간과 pitch pulse 위치 그리고 선형예측 분석한 결과들을 가지고 합성에 필요한 정보들을 모두 가지고 있는 합성유니트가 최종적으로 생성된다.

3.5 화면이동

화면 이동시에는 한 페이지(한 화면) 단위로 움직이거나, 반 페이지 단위로 움직일 수 있으며 왼쪽이나 오른쪽 bar를 기준으로 그 다음 한 페이지를 볼 수 있다. 또 한 페이지 또는 반페이

지싹 이동하는 기능만으로 화일 내에서 임의의 원하는 위치로 옮기고자 할 때에 여러 번의 명령을 반복적으로 사용해야하는 단점이 있었다. 이러한 기능을 보완하기 위하여 검사하고 싶은 화일의 위치를 ms단위로 입력하면 한번에 이동 할 수 있는 jump기능도 구현하였다.

3.6 재생기능

bar 구간 내를 D/A로 출력하는 기능은 물론 3 button mouse를 사용하는 경우는 가운데 button은 눌러도 같은 동작을 하도록 하였다. 그 이외에 두 bar 사이를 반복하여 재생하는 기능, display된 화면만을 재생하는 기능 여러 개로 분리된 음성화일을 연결하여 하나의 file로 만들면서 재생하는 기능 등이 있다.

3.7 대규모 음성 DB 구축

잘라낸 data들에 화일이름을 음성DB구축을 원하는 file이름들이 적혀있는 file을 참조하여 순서적으로 붙여서 저장하는 기능을 말한다. 이때 원 file은 configure file의 working directory에서 읽어오며, 분리된 file은 saving directory에 저장된다. 이렇게 분리한 이유는 많은 양의 file을 분리할 때에는 한번에 많은 양을 ramdisk에서 A/D한 후 이 file에서 원하는 정보를 분리하면 편리하기 때문이다. 분리를 원하는 file 이름들을 save 시마다 입력하지 않고 하나의 file로 만들어 놓고 save 할 filename을 'page up' 또는 'page down' key 사용하여 선택한 후 enter를 입력하면 그 이름으로 저장된다. 그 다음에 save 할 때는 다음 file name이 화면에 표시된다. 이때 이 file은 working directory에 있어야 한다.

3.8 음성분류 기능(수동)

음성의 voiced, unvoiced, nasal, click, silenced의 구간 정보를 저장할 필요가 있을 때 사용되는 기능으로 생성된 ZCR, 에너지, spectral tilt, 혹은 LP 분석을 한 경우는 Er/Es의 값을 보고서 mouse로 그 구간을 표시한 후 'v'나 'u' key를 입력하면 종류를 묻게 되고 이때 원하는 종류를 입력하면 spectrogram display 바로 위에 종류에 따라 각기 다른 색으로 표시된다.

3.9 한글사용

녹음된 file에서 필요한 데이터를 분리하여 filename을 붙일 때는 영문으로 붙일 수도 있으나 한국어 음성데이터는 한글로 붙여야 편리하기 때문에 한글이 꼭 지원되어야만 된다. 사용되고 있는 한글code는 조합형과 완성형이 있는데 조합형으로 filename을 부치게 되면 DOS 상에서 filename을 읽을 수 없는 경우가 생기기 때문에 꼭 완성형을 사용하여야 한다. 개발된 음성편집용 tool 에서는 DOS상에서의 한글 filename을 지원하기 위하여 프로그램 내부에서는 조합형을 사용하고 작업 중에 만들어지는 결과는 완성형으로 작성되도록 하였다.

3.10 Labelling

녹음된 file들은 분리하지 않은 상태 내에서 label을 붙이고 나서 이 결과 file을 이용하여 피치나 에너지 등을 구할 수도 있으며 인식을 위한 학습데이터로도 사용될 수 있다. 구간별로 label를 붙이게 되는데 시작점과 끝점 그리고 구간의 길이가 기록된다. 다음은 labeling의 예를 보여준다.

(Labeling 예)

가을은 9480 44160 1445
pause 44160 54600 435
참 54600 115560 2540
pause 115560 126480 455
아상한 126480 202800 3180
pause 202800 218520 655
계절이다 218520 247920 1225

첫번째 항목은 label이고 두번째는 원래 file에서의 구간시작 file point 세번째는 구간끝 file point 그리고 마지막으로 구간의 길이이다. 단위는 [ms]이다. 위의 예는 어절별로 label을 한 예를 보이고 있는데, 물론 labeling은 음소별, 음절별의 어떤 단위라도 가능하다.

3.11 스펙트로그램 인쇄

스펙트로그램(spectrogram)은 단일 색상으로 나타내어지는 것이 아니라 gray level로 그려지기 때문에 이를 print할 때는 이들 각 화소의 gray level를 4x4의 dot로 풀어서 그려주어야 된다. Printer로 data를 보내는 부분은 화면 전체를 보내는 것이 일반적이나 사용상 불편하기 때문에 원하는 부분을 좌우 bar로 선택하게 하였다. 또 논문에 붙이기 위하여는 작은 크기의 그림이 필요하기 때문에 축소하여 출력하는 기능을 내장하였다. 축소시 text는 printer에 내장되어 있는 font를 축소비에 따라 다르게 사용하였는데 Qnix printer와 같이 다양한 크기의 font가 내장되어 있는 경우에는 문제가 되지 않지만 널리 사용되고 있는 HP 모드에서는 다양한 font를 default로 지원하지 않기 때문에 이를 해결하기 위한 방법을 검토중에 있다. [4][5][6] 그리고 스펙트로그램을 인쇄 할 때에는 가로축과 세로 축에 시간과 주파수에 대한 눈금이 자동적으로 그려 지게 되어 있다. 그림 3은 Qnix 모드로 0.7배로 축소되어 출력된 spectrogram이다.

IV. 결론

지금까지 음성합성 및 인식 등에 이용될 수 있는 음성분석 tool의 기능에 대하여 개략적으로 논하였다. 우리 나라에 DSP(Digital Signal Processor)가 들어와서 사용된지 10여년이

지났지만 우리의 손으로 개발된 음성분석tool이 아직 없었기 때문에 그 동안은 이 방면의 연구를 위하여 불편을 감수하면서 외국산 S/W를 사용하거나 나름대로의 필요한 부분만을 만들어 사용하였다. 그런 의미에서 개발된 분석장치는 H/W 부터 S/W까지 전부 국산이라는 데서 의의를 갖는다고 본다. 처음부터 상용화를 위하여 개발하였다기 보다는 개발중인 음성합성장치를 위하여 만들어졌기 때문에 그래픽이나 사용자 인터페이스에 불편한 점이 많다고 생각되기 때문에 이러한 점의 보완과 함께 필요하다고 생각되는 기능을 계속해서 추가해 나갈 예정이다.

참고문헌

[1] TI, TLC3204C 매뉴얼.

[2] 이종락, " 반음소 : 새로운 음성합성 및 인식단위" 제10회 음성통신 및 신호처리 워크샵 논문집, pp.208-212, 1993

[3] 김옹인외, "한소리 : 무제한음성합성시스템", 제11회 음성통신 및 신호처리 워크샵 논문집, pp.342-345, 1994.

[4] Qnix, QLBP3000 매뉴얼

[5] HP, HP Laser Jet series II 매뉴얼

[6] Addison Wesley, Programmer's Guide to the EGA and VGA cards, second edition.

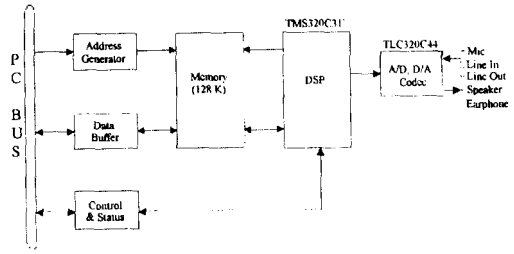
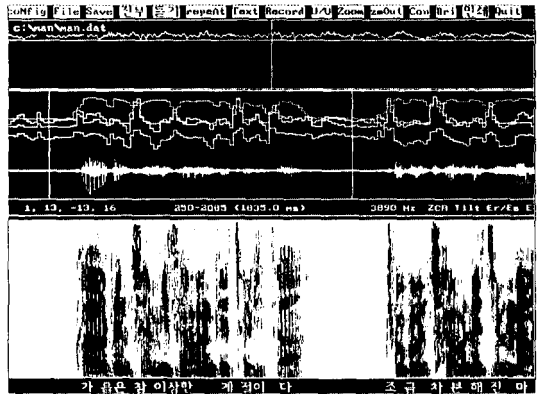


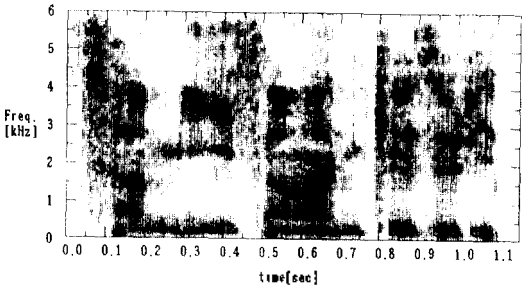
그림 1. DSP보드의 블럭도



(그림 2) 개발된 tool의 화면

<표 1.> Sampling 주파수와 실제 sampling 주파수 대비표

conversion clock	RA 값	RB 값	LPF cutoff 주파수	실제 conversion clock
8KHz	10	37	4.16KHz	8.108KHz
10KHz	8	37	5.2 KHz	10.135KHz
12KHz	7	36	5.95KHz	11.9KHz
15KHz	6	33	6.94KHz	15.15KHz
16KHz	5	37	8.33KHz	16.21KHz



(그림 3) 축소되어 프린트된 spectrogam