1993년도 한국음향학회 학술논문발표회 논문집(제 12권 1(s)호)

TMS320C30을 이용한 실시간 음성부 검출 알고리즘 구현

이°항섭, 서 영주, 한 민수
한국전자통신연구소 음성응용연구실

(Implementation of A Real-time Endpoint Detection Algorithm Using TMS320C30)

HangSeop Lee, YoungJoo Seo, Minsoo Hahn
Electronics and Telecommunications Research Institute

ABSTRACT

이 논문은 최근에 개발된 실시간 음성부 검출 알고리즘[1]을 TMS320C30 System board와 IBM PC486을 이용한 implementation에 관한 논문이다. 음성부 검출 알고리즘은 Enegy와 LCR(Level Crossing Rate)를 이용하여 각 frame을 음성/묵음으로 분류하는 방법을 사용하였고 DSP 보드를 사용하여 한 frame이 입력되면 다음 frame이 입력되기 전에 그 frame에 대한 음성/묵음 분류를 하여 음성입력이 끝남과 동시에 음성이라고 판단되는 부분만을 DPS memory상에 저장하므로 불필요한 memory의 낭비를 줄이고 다음 단계의 음성처리를 위한 시간을 절약하였다.

이 알고리즘의 성능 평가를 위하여 Rabiner와 Sambur의 알고리즘과 한민수의 알고리즘과를 전문가가 수작업으로 찾아낸 결과와 비교 평가하였다. 알고리즘의 오차는 평균 남성4.925ms, 여성 5.85ms로 1 frame이내의 오차를 보였다.

1. 서 론

음성인식, 합성 및 분석등 음성공학의 거의 모든 분야에서 음성신호의 시작점 및 끝점을 주변잡음(Environmental Noise)과 분리하여 정확하게 알아내는 일은 매우 중요하다. 최근에 multimedia통신이나 지능망이라는 새로운 개념이 대두되면서 이 분야에 대한 관심이 점점 커지고 있는 실정이다. 특히 음성신호의 경계점 또는 끝점(endpoint)의 검출은 고립단어 인식시스템의 개발에는 반드시 선결되어야만 하는 과제이며 음성부 검출기의 성능은 고립단어 인식시스템의 최종 인식률에 직접적인 영향을 주게된다. 또한 믿을 수 있는 음성부 검출 알고리즘이 존재한다면 불필요한 묵음을 사전에 제거함으로서 단어 인식에 소요되는 시간을 줄일 수도 있는 것이다. 한편 음성부 검출 알고리즘은 대용량 음성 데이타베이스의 효율적인 구축에도 크게 기여할 수 있다. 즉 미리 녹음되어있는 음성데이타를 각 단어 또는 문장단위로 잘라내어 음성 데이타베이스를 구축하는 과정은 현재는 주로 전문가가 직접 음성신호를 컴퓨터 화면에 display해서 음성의 경계점을 수동으로 판정하여 필요한 부분만 잘라서 컴퓨터의 기억장치에 저장하는 지루하고도 많은 시간이 소요되는 절차를 거치고 있다. 따라서 만일 신뢰할 수 있는 음성부 검출기가 실현된다면 약간의 손쉬운 변형만으로도 이 지루한 과정들이 자동화 될 수 있는 것이다.

신호대 잡음비가 상당히 큰 경우를 제외하고는 고품질의 음성부 검출기를 실현한다는 것은 쉽지않은 일이라는 것은 널리 알려져 있는 사실이다. 즉 신호대 잡음비가 충분히 큰 경우는 가장 작은 에너지 레벨을 갖는 음성신호라 할지라도 주변잡음보다는 큰 에너지값을 가지므로 에너지함수만 이용하여도 성능이 좋은 음성부 검출기를 쉽게 구현할 수 있는 것이다. 일반적으로 신호대잡음비가 30dB를 넘는 경우에는 에너지와 영교차율(Zero Crossing Rate)을 이용하여 간단하게 음성부 검출 알고리즘을 실현할 수 있다고 알려져 있다. 이 경우 가장 작은 음성신호라 할지라도 그 에너지가 주변잡음 에너지보다 커지므로 손쉽게 음성부 검출 알고리즘을 실현할 수 있는 것이다[2]. 그러나 이렇게 이상적인 조건이 현실적으로 실현되기 어렵기 때문에 지금까지도 많은 연구가 이루어지고 있는 것이다. 음성부 검출기에 관하여 지금까지 보고된 주요 연구결과들은 다음과 같다. Rabiner와 Sambur는 에너지와 영교차율을 이용한 음성부 검출기[3]를 구현하였으며 Wilpon은 HMM을 이용한 음성부 검출 알고리즘을 제안하였다[4]. Lamel은 레벨 이퀄라이저(Level Equalizer)를 통과한 음성신호의 에너지펄스를 이용한 음성부검출 방법을 실현하였으며[5], Larar는 speech와 EGG information을 이용한 two-channel 방법을 제안하였다[6]. Neuberg는 음성신호의 저주파영역 에너지를 특징변수로 이용하여 음성부 검출기의 성능을 향상시키는 방법을 제안하여 그 유용성을 인정받았다[7]. 그리고, 한민수는 original과 preemphasized siginal에 대한 Energy, Zcr 그리고 Modified-zcr이라는 새로운 변수를 사용하여 신뢰성있는 음성부 검출 알고리즘을 구현하였다[8].

본 절에서는 위에서 언급한 여러 음성부 검출 알고리즘이 알고리즘의 특성상 실시간 구현에는 많은 어려움이 있는 이유로 실시간을 필요로하는 상용화 시스템의 구현에 적합한

TMS320C30을 이용한 실시간 음성부 검출 알고리즘 구현

실시간 음성부 검출 알고리즘을 구현하기 위하여 Enegy와 LCR(Level Crossing Rate)를 이용하여 각 frame을 음성/묵음으로 분류[9]하는 실시간 음성부 검출 알고리즘을 개발하여 TMS320C30 System Board를 이용하여 IBM 486PC상에서 구현하였다.

2. 음성 자료

본 음성부 검출 알고리즘은 음성인식 데모시스템인 샘돌이I[10]을 만들기 위해서 구상되었기 때문에 알고리즘의 개발에 사용된 데이타는 표 1.과 같이 숫자음 20개, 사칙연산 4개, 확인용 단어 2개 등 모두 37개의 단어로 구성되었다.

숫자음 : 영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구 공, 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉
사칙연산 : 더하기, 빼기, 곱하기, 나누기
계산순서 : 괄호열고, 괄호닫고
Equal발성 : 은, 는, 이면, 은?, 는?
확인 : 예, 아니오
시작, 취소 : 시작, 다시, 취소, 모두취소

표 1. 음성부 검출기 개발용 음성데이타

5명의 남성화자가 각 단어를 3번씩 발성한 555개의 단어 중 본 알고리즘의 training을 위하여 185개(제1발성)의 데이타를 사용하였고, 알고리즘 성능평가를 위해서는 남성화자가 발성한 나머지 370개(제2,3발성)의 데이타를 사용하였고, 여성화자에 대한 실험에서는 training에 전혀 참여하지 않은 여성화자 10명이 숫자음 20개의 단어를 각 1번씩 발성한 200개의 데이타를 사용하였다.

사용된 모든 데이타는 300Hz - 4.5KHz의 BPF를 거쳐, 10KHz로 샘플링 하여 16Bit 양자화 하였다.

3. 파라메타 추출

데모시스템이 사용되는 환경에 따라 음성발성환경이 수시로 변하므로 이러한 환경하에서 음성부 검출을 위한 threshold 값들을 일정한 값으로 고정시킨다면 정확한 음성부 검출알고리즘의 결과를 기대할 수 없을것이다. 그러므로 본 알고리즘은 한 단어의 음성이 발성되는 동안은(약 1.3초 내외) 음성발성환경이 갑작스럽게 변하지 않는다는 가정하에 매번 입력신호의 앞부분에서 음성부 검출을 위한 threshold값들을 구한다. 그림 1.에서 볼 수 있듯이 입력신호는 10ms 단위로 frame화 되어지고 처음의 3 frame(1~3frame)에서 입력신호의 직류성분을 계산하여 그 다음 신호부터는 자동적으로 입력신호에서 직류성분을 제거한다. 다음의 3 frame(4~6frame)에서는 energy threshold (enth)를 계산하고 이때 Level Crossing Rate를 계산하기위한 Average plus 값과 Average minus 값을 구하여 다음의 3 frame(7~9frame)에서는 lcr의 threshold(izct)값을 구한다. 즉, 본 알고리즘에서는 초기입력에서부터 실제의 음성이 입력되기 까지는 적어도 100ms의 묵음구간이 형성된다는 가정하에서 묵음구간인 90ms동안에 모든 threshol값을 계산한 이후 계속되는 입력신호의 각 frame에서 에너지(en)과 영교차율(lcr)을 구하고 이를 미리 구해진 threshold 값들과 비교를 하여 각 frame에 대한 음성/묵음 분류를 통해 음성부 검출을 행하였다.

음성 시작점 검출을 위해서는 5 frame 이상이 계속하여 음성으로 분류되어지는 시점을 음성의 시작점으로 정하였고, 끝점은 연속되는 15개의 frame중 13개의 frame이 묵음으로 분류되어지는 시점을 음성의 끝점으로 정하였다.

알고리즘의 flow chart는 그림 1.과 같고, 사용된 parameter는 다음과 같다.

- 아날로그 필터 : 300 Hz - 4.5 KHz
- A/D 변환 : 10 KHz 샘플링, 16 bit A/D 변환기
- 프레임 설정 : 프레임내 샘플수 (10ms, 160 samples)
  프레임 이동간격 (10ms, 160 samples)
- 에너지 함수 :

$$E = \sum_{m=n-N+1}^{n} |X(m)|$$

- Level crossing rate, Z :

  Z increases itself by 1 when

```
if( x(n-1) > avr_plus && x(n) < avr_minus ||
    x(n-1) < avr_minus && x(n) > avr_plus )
```

4. 알고리즘 구현

C30 board memory의 효율적인 사용을 위하여 입력신호중 음성이라고 판단되는 부분만을 memory에 저장하는 방법을 사용하였고 실시간 처리를 위하여 A/D conversion의 timer interrupt 기간 동안에 1 frame에 대한 음성/묵음 분류를 결정하는 방법을 사용하였다. 즉, memory의 시작점을 30001번지라고 가정한다면 dc_offset 값 계산을 위하여 30300번지까지 3 frame의 입력신호를 받아들인 후 interrupt 기간 동안 dc_offset 계산을 끝내고 memory번지를 30001번지로 되돌린다. 같은 방법으로 energy와 lcr의 threshold값을 계산한 이후 시작점 검출을 위하여 일정한 길이의 memory buffer를 두어 1 frame분량(100 sample)의 데이타가 입력되면 다음 sample이 입력되기전에 - 즉 interrupt time 동안 한 frame에 대한 음성/묵음 분류를 결정하여 만일 음성으로 판정되어지면 memory에 저장하고 그렇지 않으면 memory에 저장하지 않는 방법을 사용하였다. 즉, 20초의 입력신호중 음성이 2초를 차지한다면 음성을 저장하기 위하여 2초간의 memory만 있으면 되는것이다.

5. System Configuration

본 알고리즘은 TMS320C30 DSP Board[11]를 이용하여 IBM 486PC상에서 구현하였다.

PC에서는 전체적인 동작을 제어한다. 즉, DSP board를 초기화시키고, TMS320C30 machine code를 download하고, download된 program을 수행시키고 program이 종료되면 입력된 data를 upload하여 온다. DSP board에서는 실질적인 음성부 검출 알고리즘이 수행된다. 우선 자체 내장되어있는 A/D channel을 이용하여 timer interrupt 기법을 사용해 음성 데이타를 받아들이면서 동시에 음성부 검출을 수행하고 음성이라고 검출된 데이타를 일정한 번지에 저장한후 PC에 program의 종료를 알린다. TMS320C30 Board에서 사용한 memory configuration은 표 2.와 같다.

표 2. Memory Configuration

```
0x000000h - 0x00000ch : Interrupt vectors area
0x000100h - 0x0002d1h : main program area
0x030110h - 0x030155h : program data area
0x038000h - 0x03f300h : input data area
```

전체적인 system의 구성도는 그림 2.에 보이고 있다.

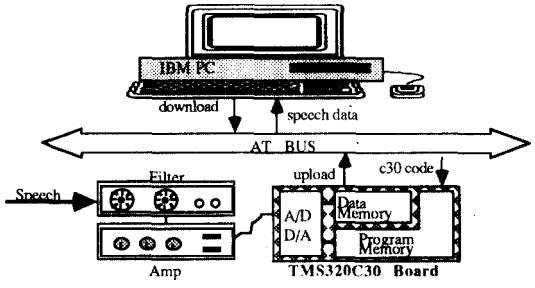

그림 2. 음성부 검출 system 구성도

6. 실험결과 및 고찰

개발한 알고리즘의 성능평가를 위하여 원래의 음성신호와 전처리된 음성신호를 보고 전문가가 수작업으로 찾아낸 음성부 검출 결과를 기준으로 한민수가 개발한 알고리즘의 수행결과와 Rabiner와 Sambur의 알고리즘의 수행결과 그리고 새로 개발한 알고리즘의 수행 결과를 서로 비교,평가하여 남성화자에 대한 실험결과는 표 3.에, 그리고 여성화자에 대한 실험결과는 표 4.에 각 화자별로 start point와 end point로 나누어 검출 오차를 ms로 나타내었다. 세 알고리즘 모두 남성과 여성화자 모두에 대해 끝점 검출시 시작점보다 많은 오차를 나타내었고, 전체적인 수행 오차는 개발된 알고리즘이 남성과 여성화자 모두 1 frame이내의오차를 보였고 한민수의 알고리즘과 Rabiner & Sambur의 알고리즘은 남성과 여성 모두 끝점에서 1 frame을 약간 초과하는 오차를 나타내었다. 그림 3.에서는 세 알고리즘의 음성부 검출 결과의 한 예를 보이고 있다.

표 3. Endpoint Detection Error - 남성화자

(Unit : msec)

Algorithm	Start	End	Average
OUR	3.06	6.79	4.925
Hahn	3.46	11.73	7.595
R&S	6.86	14.9	10.88

표 4. Endpoint Detection Error - 여성화자

(Unit : msec)

Algorithm	Start	End	Average
OUR	4.15	7.05	5.60
Hahn	4.85	13.45	9.15
R&S	9.25	15.75	12.5

그림 3. 알고리즘 수행 결과의 예 : M, R&S, Hahn, N은 각각 Handsegmentation, Rabiner&Sambur, Hanh 그리고 New 알고리즘의 결과를 나타낸다.

본 알고리즘을 개발하면서 서로 다른 세가지의 알고리즘의 결과를 비교한 결과 다음과 같은 사실을 알 수 있었다.

1) 개발된 알고리즘은 마찰음의 검출에 있어서 다른 두개의 알고리즘보다 좋은 검출 결과를 보였다.
2) 개발된 알고리즘은 다른 두개의 알고리즘보다 voice-offset 구간에서 매우 향상된 결과를 보였다.
3) 개발된 알고리즘은 모음으로 끝나는 단어의 경우 자음으로 끝나는 경우보다 voice-offset 구간에서 많은 error를 보였다.
4) 여성화자의 경우 마찰음으로 시작되는 단어의 voice-onset 구간의 검출이 남성화자에 비해 좋지 않은 결과를 보였다.

향후 본 알고리즘을 당 연구실에서 수행중인 음성인식 데모 시스템 구성에 적용할 것이고 계속해서 연결음성 검출에 적용하기 위한 개발을 계획하고 있다.

7. 결론

본 논문은 음성을 이용한 상용화 시스템 구현에 적합한 간단한 실시간 음성부 검출 알고리즘을 개발하고 이의 신뢰성을 평가한 후 TMS320C30 system board를 사용하여 구현하였다. 전체적인 수행 오차는 남성 4.925ms, 여성 5.6ms로 1 frame이내로서 시스템에 적용시 무리가 없다고 본다. 알고리즘 평가를 위하여 사용한 한민수의 알고리즘이 더 큰 오차를 보이고 있는데 이는 실제로 알고리즘의 구성상 더 좋은 결과를 보여야 하는데 그렇지 못한것은 알고리즘에 필요한 최적의 파라메타 threshold값들을 찾아주지 못한 결과인 듯 하다. 이렇듯 대부분의 알고리즘이 적용되는 데이타에 따라 즉, 음성 발성시의 환경에 따라 최적의 threshold값들을 변경해야하는 반면 새로 개발된 알고리즘은 필요한 threshold값들이 음성이 발성되는 상황에 따라 자동적으로 변화하기 때문에 사용환경에 크게 구애받지 말아야할 상용화 시스템에 적합한 알고리즘이라 생각된다.

현 알고리즘의 효용성 평가와 성능향상을 위해서는 여러 가지 환경에서 녹음된 보다 많은 데이타를 가지고 실험해 보아야 한 것이다.

감사의 글

이 연구는 한국통신 출연과제인 "통신처리 시스템의 고급 기능 개발"의 일부입니다. 연구에 도움을 주신 윤병남 부장님과 음성응용연구실 실장님 이하 모든 연구원들께 감사드립니다.

참고문헌

1. Hangseop Lee, Minsoo Hahn, "Development of a Real-Time Endpoint Detection Algorithm," ICSPAT'93, pp.1547-1552, Sep.28-Oct.1, 1993
2. L.R.Rabiner and L.W.Schafer, Digital Processing of speech signals, Englewood Cliffs, N.J., Prentice-Hall, 1978.
3. L.R.Rabiner, M.R.Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," THE BELL SYSTEM Technical Journal, Vol.54, No.2, pp 297-315, Feb.1975
4. J.G.Wilpon and L.R.Rabiner, "Application of hidden Markov models to automatic speech endpoint detection," Computer Speech and Language, Vol.2, pp.321-341, 1987.
5. L.F. Lamel et. al., "An improved endpoint detector for isolated word recognition," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-29, No.4, pp.777-785, 1981.
6. J.N.Larar, Towards speaker-independent isolated word recognition for large lexxicons: A two-channel, two-pass approach, Ph.D.Dissertation, Univ.of Florida, 1985.
7. E.P. Neuberg, "Automatic thresholding for voicing detection algorithms," Proc. IEEE International Conference on Acousics, Speech, and Signal Processing, pp.576-759, 1975
8. M.Hahn, C.Park, "An Improved Speech Detection Algorithm for Isolated KoreanUtterances," ICASSP'92, pp.1-529, March 1992.
9. 이정철, 박찬경, 이의택, "범용 디지탈 신호처리 프로세서를 이용한 실시간 음성검출 S/W의 구현", TM47IKE00701, 한국전자통신연구소, 1986.3
10. 자동통역전화를 위한 요소기술 개발(II), page 11-15, 한국전자통신연구소, 1992.12
11. TMS320C30 PC SYSTEM BOARD USER MANUAL, Loughborough Sound Images Ltd, The Technology Centre, England, Ver. 1.0, Jan, 1990

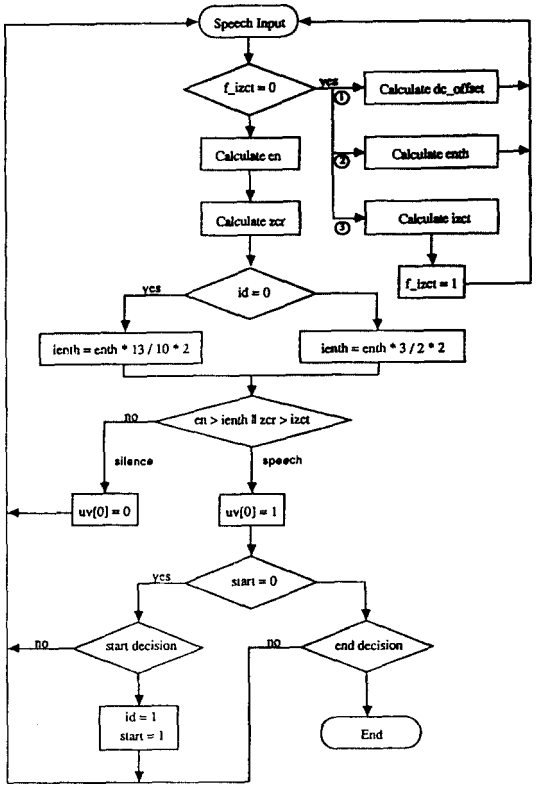

그림 1. 음성부 검출 알고리즘 Flow chart