연속 숫자음의 음절구간 검출

김득수**, 정현열*
* 영남대학교 정보통신공학과
** 대구공업대학 전자계산과

A Study on Determining Syllable Length of Connected Spoken Digits

Deok-Soo Kim**, Hyun-Yeol Chung*
* Department of Information and Communication Eng., Yeungnam University
* Department of Computer Science, Taegu Technical College

요 약

본 논문은 한국어 숫자를 연속적으로 또박또박 발음한 음성의 음절 구간 검출에 관한 내용이며 음절의 최소구간 및 스펙트럼에너지를 이용하여 연속 음성에서 구간 검출 알고리즘을 제안한다. 숫자음 11개를 연속으로 발성하여 음절 구간을 검출하며 결정된 구간과 수작업으로 한 음절구간을 비교한다. 음절시작점인 경우에는 수작업시단과 동일하거나 항상 전방향이며 종단인 경우에는 92% 데이터가 ±1 프레임내에 존재하며 제안된 알고리즘이 실용성이 있음을 보인다.

Ⅰ. 서 론

컴퓨터, 통신, 신호처리 기술 등의 급속한 발달과 함께 사람과 기계사이의 편리한 연결 방법인 음성인식의 필요성이 증가하면서 음성인식에 대한 연구가 꾸준히 진행되어 왔다. 음성의 구간 결정은 음절의 시작점과 끝점을 결정하는 것이다. 음성의 구간 결정 방법은 일반적으로 영교차율과 에너지 조합을 이용하며 대표적인 것으로는 Rabiner와 Sambur[1]의 에너지와 영교차율을 이용한 음성 끝점 검출 알고리즘과 최근 연구가 되고있는 웨이브렛(Wavelet) 변환을 이용한 잡음음성의 끝점검출[2] 등을 들 수 있다. 음성으로 전화번호를 인식하는 시스템을 설계하는 경우 전화번호가 7자리이면 경우의 수는 10^7개가 된다. 이 경우 입력된 음성을 인식하는데 음소로 인식하는 방법과 음절로 인식하는 방법으로 나누어 볼 수 있다. 만약 정확하게 음절 구간의 검출이 가능하다면 음소로 음성인식[3,4]을 하는 경우보다 음절로 인식하는 경우가 음성인식률이 증가된다. 따라서 본 논문은 한국어 연속 숫자음을 인식하는 시스템을 구성하기 위한 기초 작업으로 음절의 구간 결정에 음절의 최소구간 및 스펙트럼에너지를 이용한 방법을 제안한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 전체적인 데이터 처리방법과 입력된 음성신호를 분리, Ⅲ장에서는 음성구간을 추출하는 조건과 알고리즘을 제안, Ⅳ장에서는 실험 및 고찰 마지막으로 Ⅴ장에서 결론을 맺는다.

Ⅱ. 시스템 개요

2.1 시스템 개요

시스템은 그림 1과 같이 구성된다. 입력되는 음성은 사운드 카드를 이용하여 8kHz로 표본화하고 8비트로 양자화한다. 양자화된 데이터는 전달 함수가 $H(z) = 1-0.97z^{-1}$ 인 디지털 필터를 통과시켜 고주파 성분이 강조되며, 이 신호는 16ms 구간마다 FFT[5]처리 된다. FFT 처리된 신호는 2개 신호로 분리하여 음성의 시작점과 끝점을 결정한다. 재생데이터는 결정된 구간을 이용하여 디지털 필터를 통과하기전의 데이터에서 데이터를 추출하며 사운드 카드로 출력하여 정확도를 확인한다.

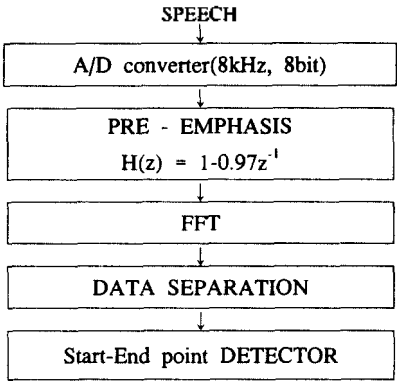

그림 1. 전체 시스템의 개략도.

2.2 데이터 분석

8kHz로 양자화된 데이터는 16ms구간으로 FFT처리되며 단위 구간은 128개 신호가 된다. 128개 음성신호가 FFT 처리되면 64개의 주파수 성분이 되며 이들 64개 주파수 성분은 2개의 주파수 성분으로 변환하며 표 1과 같다.

표 1. 2개 주파수 밴드.

FFT point	Frequency(Hz)	FFT point	Frequency(Hz)
0 ~ 25	0 ~ 1624	26 ~ 63	1625 ~ 4000

또한 이들 주파수 성분의 총합은 시간의 함수이므로 SP(t)로 표현하며 음성신호의 크기를 의미한다. 또한 이 f(i) 신호는 시간의 함수이므로 f(i,t)로 표현하며 i≦25 이하인 신호와 i≧26 인 2개의 신호로 구분하여 SPvv(t), SPcc(t) 신호로 구분하며 SPvv(t)신호는 저주파이고 SPcc(t)는 고주파 영역이 된다. 파형을 관측하면 파형의 형태는 주로 SPvv(t)신호로 결정된다.

$$SP(t) = \sum_{i=0}^{63} f(i,t) \tag{1}$$

$$SPvv(t) = \sum_{i=0}^{25} f(i,t) \tag{2}$$

$$SPcc(t) = \sum_{i=26}^{63} f(i,t) \tag{3}$$

Ⅲ. 음절구간 검출 알고리즘

음성신호는 종 모양의 분포를 가지는데 끝 부분은 시작부분에 비해 서서히 감소[6,7]하는 원리를 알고리즘 1에 이용한다. 연속음성에서 음절을 구분하기 위해서는 음성이 있는 부분과 묵음부분을 구별해야 한다. 숫자음에서 음절이 있는 부분은 표 2의 '숫자음 지속시간'을 고려하여 최소음절 지속시간을 단위시간 6이상으로 제한하며 즉 5이하는 음절이 구성될 수 없다는 의미이다. 또한 음성의 크기를 고려하여야 한다. 표 4의 '음절 평균 에너지'를 참조하면 한 음절에서 최소평균 에너지는 417이상이므로 평균 에너지를 100이상으로 제한하였다. 즉 음절 구간이 결정되면 사람이 인식할 수 있는 정도의 소리크기 인가를 조사해야 하는 의미이다. 본 논문에서 음절구간을 검출하기 위하여 음절의 최소구간 및 음성에너지의 정보를 고려한 새로운 방법을 알고리즘 1과 같이 제안한다. Ⅱ장에서의 SPvv(t)와 SPcc(t)의 신호의 크기를 다음 식 (4)-(7)과 같이 제한한다.

$$SPvv(t) < 5 \tag{4}$$

$$SPcc(t) < 5 \tag{5}$$

$$SPvv(t) < 10 \text{ and } SPcc(t) < 10 \tag{6}$$

$$SPvv(t) + SPcc(t) < 20 \tag{7}$$

$SPvv(t)$, $SPcc(t)$신호가 식 (4)-(7) 조건에 해당하면 묵음으로 처리하며 음절구간 결정 알고리즘은 다음과 같다.

알고리즘에서 음성신호가 될 수 있는 구간의 최소치는 알고리즘 1의 e-s+1>5 에서 6이상으로 제한하며 신호의 크기는 알고리즘1의 SP>100에서 제한한다. 여기서 last는 전체 입력음성의 끝, spt와 ept는 결정된 음절구간의 시작점과 끝점을 의미한다.

Algorithm1. Start-End Point Set

```
sw=0
for  i=1 to last
if not(식 (4)-(7)) then
    if sw=0 then sw=1 :s=i
endif
if (식(4)-(7)) then
    if sw=1 then
        sw=0 :e=i-1
```

```
        if e-s+1 > 5 then
            SP=sum(SP(t)) (t=s~e)
            if SP > 100 then set spt=s, ept=e
        endif
    endif
endif
next i
END Algorithm1
```

Ⅳ. 실험 및 고찰

4.1 음성 데이터 녹음

실험에 사용한 데이터는 표 2와 같이 11개 숫자음이며 녹음한 인원은 남자 5명 여자 5명이며 3번씩 11개 숫자음을 연속적으로 또박또박 발음하며 비교적 조용한 연구실에서 사운드 카드를 이용하여 녹음하였다.

표 2. 음성 데이터.

숫자음	일	이	삼	사	오	육	칠	팔	구	공	영

4.2 음절의 지속 시간

표 3. 숫자음 지속시간.

화자 / 숫자	M1	M2	M3	F1	F2	F3
일	20	41	26	30	25	25
이	33	27	23	32	27	30
삼	23	26	28	34	31	28
사	23	26	27	30	32	27
오	20	18	22	25	21	23
육	9	18	20	16	21	21
칠	23	23	23	30	25	28
팔	21	21	19	27	23	23
구	18	20	21	24	25	23
공	23	22	19	27	28	23
영	14	15	19	22	22	23

표 3에서 데이터 내용은 본 논문에서 제안된 음성 구간 검출방법으로 출력된 결과이며 표 3에서 M1, M2, M3은 남자이고 F1, F2, F3은 여자이며 M1의 '일'의 데이터는 20이므로 FFT의 단위구간이 16ms이므로 발성시간은 320ms가 된다.

4.3 발성시간

표 4. 발성 시간.

화자	총 발성시간	총 음절시간	총 묵음시간
M1	549	227	322
M2	457	257	200
M3	479	247	232
M4	689	259	430
M5	505	223	282
F1	526	297	229
F2	486	280	206
F3	457	274	183
F4	520	228	292
F5	458	240	218
평균	513	253	259

표 4에서 '발성시간'의 총 발성시간은 11개 숫자음에서 첫음절의 시작점부터 11번째 음절의 끝점까지의 시간이며 10명 평균은 513단위 구간이므로 8.2초이다. 총 음절시간은 11개 음절에서 순수 음절 발성시간이며 총 묵음시간은 음절과 음절의 묵음구간의 총합계시간이다. 총 음절시간의 평균값과 총 묵음시간의 평균값이 차이가 거의 없는 속도로 발성했음을 알 수 있다.

4.4 음절 평균 에너지

표 5는 남자 3명 여자 3명에 대한 음절이 있는 구간의 에너지이며 이 값은 결정된 음성 구간에서 시작점과 끝점까지 SP(t)의 평균값이다. 음절 '구'인 경우 다른 음절보다 비교적 작은 값임을 알 수 있다.

표 5. 음절 평균 에너지.

화자 / 숫자	M1	M2	M3	F1	F2	F3
일	4773	14015	7763	5684	5440	15090
이	3987	4834	3279	1239	2349	1875
삼	1375	4283	1732	4747	6650	7702
사	1256	1824	1274	9049	13407	8089
오	1771	1830	2174	7836	16998	9244
육	3238	2005	1113	2911	3822	4381
칠	1015	6017	5187	3271	4771	3607
팔	1277	3255	1388	5923	7939	5649
구	788	1108	622	2620	2193	1214
공	417	1985	956	3876	6338	4835
영	1473	1198	1216	2422	5729	7360

4.5 수작업구간과 비교

표 6은 음절시작점에 대한 비교내용이다. 알고리즘에 의한 시작점은 수작업보다 모두 전방향이며'일' 인 경우 최대 2프레임이며 '삼','사'는 최대 5프레임 앞서 있다. 파형을 분석한 결과 '사' 음소는 발성하는 사람의 차이 수작업할 때 위치결정의 차이 때문에 상대적으로 차이가 있었다. 시점이 +2 프레임이내에 90%의 데이터가 있음을 알 수 있다.

표 6. 음절시작점 비교.

프레임 / 숫자	+0	+1	+2	+3	+4	+5
일	40	50	10	0	0	0
이	40	46	7	7	0	0
삼	13	23	23	19	10	10
사	6	36	33	10	10	3
오	19	63	10	7	0	0
육	30	46	13	3	6	0
칠	17	60	23	0	0	0
팔	36	60	0	0	3	0
구	23	70	3	3	0	0
공	40	50	7	3	0	0
영	53	40	7	0	0	0

표 7은 음절끝점에 대한 비교내용이다. 음절끝점의 프레임은 알고리즘에 의한 구간결정과 전방향 후방향 모두 출력되었다. 종점이 ±1프레임이내에 92%의 데이터가 있음을 알 수 있다.

표 7. 음절끝점 비교.

프레임 / 숫자	0	±1	±2	±3	±4	±5
일	50	33	7	3	3	3
이	40	43	10	0	7	0
삼	70	23	0	3	0	3
사	56	23	10	0	7	3
오	59	30	3	3	0	3
육	46	53	0	0	0	0
칠	59	36	0	3	0	0
팔	79	20	0	0	0	0
구	56	40	3	0	0	0
공	63	30	3	0	3	0
영	59	40	0	0	0	0

VI. 결 론

본 논문에서는 음성의 구간 결정에 대한 새로운 방법 제안을 하였고 시작점은 +2프레임이내에 90%, 종점은 ±1프레임이내에 92% 존재하여, 이 결과는 숫자음 인식에 응용할 수 있다고 기대된다. 본 연구에서는 11개 숫자음을 평균 8.2초에 또박또박 발음하였으나 좀 더 빠르게 연속적으로 발성한 음성의 구간 결정, 음성의 크기에 관계없는 방법, 연속음성에서 음절수 결정 등이 추후 연구과제가 된다.

참 고 문 헌

1. L. R. Rabiner and Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterance," Bell Syst. Tech. J., Vol.54, No.2 pp. 297-315, Feb. 1975.
2. 석 종원, 배 건성, "Wavelet 변환을 이용한 잡음음성의 끝점 검출," 제9회 신호처리 합동 학술 대회 논문집, PP.69-72, 1996.
3. Kai-Fu Lee "Automatic Speech Recognition," Ph.D Thesis, Computer Science Department, Carnegie Mellon University, 1989.
4. 이시욱, 김득수, 정현열, "음성인식 기능을 가진 주소입력 검색시스템," 제9회 신호처리 합동 학술 대회 논문집, PP.611-614, 1996.
5. S. D. Stearns, R. A. David, "Signal Processing Algorithms," Prentice-Hall, 1988.
6. C. C. Wooters, "Lexical modeling in a speaker independent speech understanding systems," ICSI Technical Report TR-93-068, 1993.
7. 한국전자통신연구소,"자동통역전화를 위한 요소기술 개발(IV)," 1994.