한국음향학회지 제20권 제3호 pp. 70~75 (2001)

우리말 연속음성의 음절 분할법

A Syllabic Segmentation Method for the Korean Continuous Speech

한 학 용*, 고 시 영**, 허 강 인*
(Hag Yong Han*, Si Young Koh**, Kang In Hur*)

*동아대학교 전자공학과, **경일대학교 전자정보공학과

(접수일자: 2001년 2월 13일; 채택일자: 2001년 2월 23일)

본 논문은 우리말 연속음성에 대한 음절단위 분할법을 제안한다. 이 방법은 다음 3단계로 이루어진다: (1) 음성의 시간영역 분할 파라메터인 피치, 에너지, ZCR, PVR을 이용하여 음성데이터를 자음, 자음, 묵음 단위로 라벨링하여 토큰 (Token)을 형성, (2) 형성된 토큰을 유한상태오토마타를 이용하여 한국어 음절구조로 파서 (Parser)를 설계하여 스캐닝 (Scanning), (3) 의사 음절핵 정보를 이용하여 두개 혹은 여러 개의 음절을 가지는 음성부분에 대한 재분할을 통하여 음절단위 분할 완성. 제안된 방법에 대한 성능 평가를 위해서 문장과 단어단위 연속음성에 대한 분할 실험결과 각각 73.7%와 85.9%의 분할률을 얻었다.

핵심용어: 음절분할, 세그멘테이션, 연속음성인식
투고분야: 음성처리 분야 (2.4)

This paper proposes a syllabic segmentation method for the korean continuous speech. This method are formed three major steps as follows: (1) labeling the vowel, consonants, silence units and forming the Token the sequence of speech data using the segmental parameter in the time domain, pitch, energy, ZCR and PVR, (2) scanning the Token in the structure of korean syllable using the parser designed by the finite state automata, and (3) re-segmenting the syllable parts witch have two or more syllables using the pseudo-syllable nucleus information. Experimental results for the capability evaluation toward the proposed method regarding to the continuous words and sentence units are 73.5%, 85.9%, respectively.

Keywords: *Syllabic segmentation, Segmentation, Continuous speech recognition*
Ask subject classification: *Speech signal processing (2.4)*

I. 서 론

연속된 음성신호를 동일한 음운특성을 갖는 소구간으로 나누는 것을 세그멘테이션 (이하 "분할")이라 하며 음성신호처리의 주요한 과제 중의 하나이다. 그러나, 실제 음성 분할은 음소에 대한 정확한 정보와 지식이 필요하며, 발화자의 발음습관 혹은 심리상태 등과 같은 발화자간에 존재하는 개인성 때문에 각 음소들에 존재하는 공통적인 음성정보와 조음결합 등을 고려하여 정확한 음소의 경계점을 찾는다는 것은 어려운 작업이다. 특히, 대어휘 불특정화자 연속음성인식 시스템을 구성하기 위해서는 음성신호로부터 음운 경계를 검출하여 음소나 음절단위로 분할하는 과정은 시스템의 계산량을 감소시키는 대표적인 전처리에 속한다.

음성분할에의 접근은 확률모델, Fuzzy[1], Neural Network, HMM 등의 패턴매칭 방법으로 음소인식을 통하여 훈련된 데이터에 의해 이루어지는 간접적인 처리방법과 시간영역의 음성 파라메터인 ZCR (Zero Crossing Rate), LCR (Level Crossing Rate), PVR (Peak Vally Rate), 피치 그리고 음성의 지속시간과 같은 정보와 주파수 영역의 스펙트럼 동적 변화 정보와 같은 음향학적인

책임저자: 한학용 (hyhan@electro.donga.ac.kr)
604-714 부산광역시 사하구 하단동
동아대학교 전자공학과 패턴연구실
(전화: 051-200-6773; 팩스: 051-200-7712)

특징 규칙에 의한 방법 등에 의하는 규칙에 의한 방법이 널리 사용되어져 왔다. 간접적인 처리방법은 근본적으로 음소인식절차와 동일한 것으로 특징벡터로 표시된 음운패턴을 입력음성과 비교하여 얻은 정보를 이용하여 분할을 행하기 때문에 표준패턴의 작성시 발생하는 문제점 및 학습용 데이터의 양에 의존한다. 반면에 규칙에 의한 방법은 음소의 음향학적인 특징으로 이루어진 분할 파라메터들을 설정하고 이들의 임계값에 의해 사전 훈련없이 자동으로 분할할 수 있는데 반하여 임계값의 설정이 정량적이지 않는 단점이 있다[3].

음성인식시스템의 언어모델은 음소, 음절, 단어 단위로 이루어지며, 음소에 기반한 언어모델인 경우 후처리가 음절에 의한 것보다는 훨씬 복잡하다. 그러므로 신뢰할만한 음절분할이 이루어진다면 기존의 음성인식시스템의 성능향상뿐만 아니라, 전처리과정에 적용할 경우, 음절단위의 인식에 높은 인식률을 보이는 음성인식 알고리듬을 이용한 연속음성인식 시스템으로의 확장이 가능할 것이다.

본 논문에서는 우리말 연속음성에 대한 음절단위 분할법을 제안한다. 이 방법은 3개의 단계로 이루어지는데, 첫째 단계에서 음성의 시간영역 파라메터들을 이용하여 음성데이터의 음향학적 특징열인 모음 (Vowel), 자음 (Consonant), 묵음 (Silence) 단위로 라벨링한다. 두 번째 단계에서는 유한상태 오토마타를 이용하여 설계된 파서 (Parser)로 라벨열을 검색하고 이를 토큰단위로 1차분할하는 단계이다. 마지막으로 이들 토큰들은 모두 한국어의 음절구조로 되어 있지만 두번째 단계에서 조음결합 등의 영향으로 분할하지 못하는 두개 혹은 더 이상의 음절이 포함되어 있을 수 있다. 따라서 이들 토큰속에 포함되어 있는 음절의 경계를 결정하기 위해서 음절지속시간과 에너지정보를 이용하여 음절의 경계를 결정한다.

본 논문의 2장에서 분할 파라메터들을 소개하고 3장에서 음절분할 방법을 제안한다. 4장에서는 음절 분할기의 성능을 검정하기 위한 시뮬레이션 과정과 실험 결과를 제시하고 5장에서 결론을 맺는다.

II. 분할 파라메터들

음성의 분할에 사용하는 시간영역의 분할 파라메터에는 에너지, 영교차율 (ZCR), 레벨교차율 (LCR), PVR (Peak Vally Rate), 정규화된 자기상관계수, 피치 등이 있다.

2.1. 에너지

에너지는 일반적으로 모음구간은 파워가 크고 파형이 비교적 길고 안정된 평탄부를 형성한다. 에너지 평탄부는 모음구간에 비하여 적지만 무음구간, 비음의 정상부, 또는 마찰자음구간에도 관찰된다. 에너지의 상승부와 하강부는 자음에서 모음으로의 과도부분과 모음에서 자음으로 과도부분에 해당한다. 따라서 이를 효과적으로 이용하기 위해 식 (2)와 같은 일반적인 자승평균대수 에너지 대신에 신호 파형의 변화정도를 크게 하여 임계값에 여유를 주기 위해 식 (1)을 사용한다.

$$E_n = \frac{1}{N}\sum_{m=0}^{N-1} x^2(m) \tag{1}$$

$$E_n(dB) = 10\log E_n \tag{2}$$

2.2. 영교차율 (ZCR)과 레벨교차율 (LCR)

영교차률은 무성음 구간에서는 비교적 크게 나타나며 유성음이나 무음 구간에서는 적게 나타나는 성질을 가지므로 에너지와 함께 음성구간의 검출을 위해 많이 사용되어 왔다. 그러나 영교차율은 정확한 기준 레벨을 유지하지 않으면 신뢰성 있는 결과를 얻을 수 없으므로 이용에 제약이 따른다.

LCR (Level Crossing Rate)은 영교차율이 수평축의 영점을 기준으로 하기 때문에 배경잡음의 영향을 받기 쉬워 음성자체만의 특징을 제대로 나타내지 못하게 되므로 문턱값을 정하여 배경잡음의 영향을 제거시켜 음성만의 교차율을 얻기 위한 파라메타이다.

$$L_n = \frac{1}{2}\sum_{m=0}^{N-1}|\mathrm{sgn}[x(n-m)-TH] - sgn[x(n-m-1)-TH]|$$

$$\text{where,}\quad \mathrm{sgn}[s(n)] = 1, s(n)\geq 0 \tag{3}$$

2.3. PVR (Peak Valley Rate)

PVR은 음성파형에서 이전 샘플과 현재 샘플의 기울기가 (+)에서 (-)로 변하는 점을 peak로 하고 기울기가 (-)에서 (+)로 변하는 점을 valley로 하여 한 프레임에서 이들의 합으로 정의하며 식 (4)와 같다.

$$PVR_n = \sum_{m=n}^{n+N-1}(1-u[\Delta x(m)\cdot\Delta x(x+1)])\cdot w(n-m)$$

$$\Delta x(m) = x(m)-x(m-1)$$

$$u(n) = 1,\quad n\geq 0,\quad 0, \text{otherwise} \tag{4}$$

PVR은 음성의 프레임 단위의 미세한 변화를 잘 나타내 주며 음소분할에 유용하다. 자음에서의 PVR값은 크고 모음에서의 PVR값은 상대적으로 적게 나타난다. 또한 자음에서 모음으로 변하는 부분과 모음에서 자음으로 변하는 부분에서 값의 변화가 크다.

2.4. 정규화한 자기상관계수[4]

정규화된 자기상관계수는 식 (5)로 표현되며 한 프레임 내에서 인접 샘플간의 관련성을 나타내는 것으로 -1부터 1값을 갖는다. 유성음의 경우 에너지가 낮은 주파수 (3kHz 이하)에 집중되고 주기적인 안정된 특성을 나타내기 때문에 거의 1에 가까운 값을 갖는다. 반면에 무성음은 비주기적이고 과도적이며 에너지가 높은 주파수 부분에 집중되기 때문에 거의 0에 가깝다.

$$\Phi_1 = \frac{\sum_{m=0}^{N-1} x(m)x(m-1)}{\sqrt{\sum_{m=0}^{N-1} x^2(m) \sum_{m=0}^{N} x^2(m)}}, \quad 0<l<p \qquad (5)$$

2.5. 피치 (Pitch)

피치는 모음과 같은 유성의 음성신호에서 관찰되는 준주기적인 성분으로 모음구간의 결정에 결정적인 역할을 하는 파라메터이다. 근사적인 피치를 결정하는 방법들에는 자기상관함수와 상호상관함수를 이용하는 방법 등 많은 방법들이 제안되어져 있으나 본 논문에서 사용한 피치 결정법은 Medan이 제안한 상호상관함수를 이용한 알고리듬을 사용하였다[2]. 이 방법은 음성 데이터열에 대하여 일정구간을 슬라이딩시키면서 적절한 피치의 위치를 결정하는 방법으로 t=t0에서 시작하는 n개의 인접하고 가변적인 윈도우 두개가 벡터 x_n, y_n을 형성하여 범위 $N_{\min} \le n \le N_{\max}$ 내에서 상호상관계수 $\rho_n(x,y)$를 계산한다. 피치 결정은 계산된 상호상관계수가 최대가 되는 n의 값이 된다.

본 논문에서 1단계에서 적용한 분할 파라메터는 이중에서 에너지, 영교차율, PVR, 피치만을 이용하였다.

III. 음절분할법

3.1. 1단계: 음성 데이터열의 라벨링

1단계에서는 앞 절에서 소개한 분할 파라메터들을 이용하여 연속음성에 대하여 프레임별로 모음, 자음 그리고 묵음으로 라벨링하여 구문적으로 의미를 갖는 최소단위인 토큰을 형성하는 단계이다. 우리말의 음절은 단모음과 이중모음으로 이루어진 모음과 파열음, 파찰음, 마찰음, 비음, 설측음, 탄설음으로 이루어진 자음의 결합으로 구성되며 반드시 하나의 음절은 하나의 모음을 포함하게 된다. 그러므로 앞장에서의 분할 파라메터에 대한 정보를 종합하여 그림 1과 같은 알고리듬으로 라벨링하여 토큰을 생성한다. 본 연구에서는 묵음, 자음, 모음을 각각 0,1,2로 라벨링하였다.

3.2. 2단계: 라벨링 열의 스캐닝에 의한 1차분할

위에서 제안한 방법으로 라벨링 할 경우, 자음보다 모음에서 상당히 좋는 라벨링 결과를 보여준다. 그러므로 자음열에 썩여있는 0과 2의 오인식 결과로 인하여 결과적으로 모음과 자음을 완전히 분리할 수 없게 된다. 2단계에서는 유한상태 오토마타의 형태로 1단계에서 라벨링된 라벨열을 우리말의 구조에 맞추어 스캐닝하는 파서 (Parser)를 설계하여 음절분할이 이루어지도록 한다.

그림 2는 유한상태 오토마타로 형태로 표현된 파서를 보여준다. 이 파서는 15상태를 가지고 있다. 각 상태는

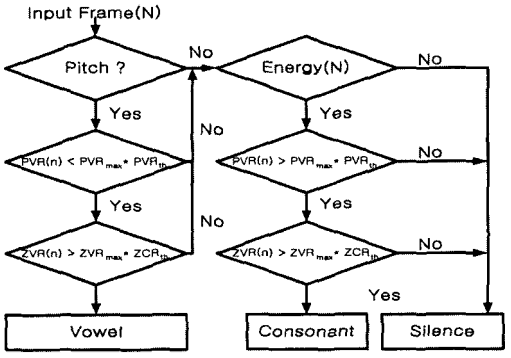


그림 1. 라벨링 알고리듬
Fig. 1. Labeling algorithm.

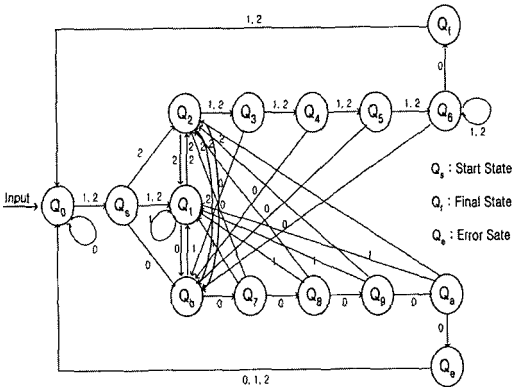


그림 2. 음절분할 파서
Fig. 2. Syllabic segmentation parser.

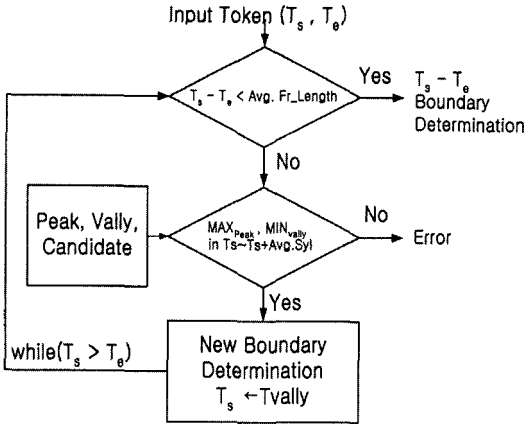


그림 3. 후처리 알고리듬
Fig. 3. Post-processing algorithm.

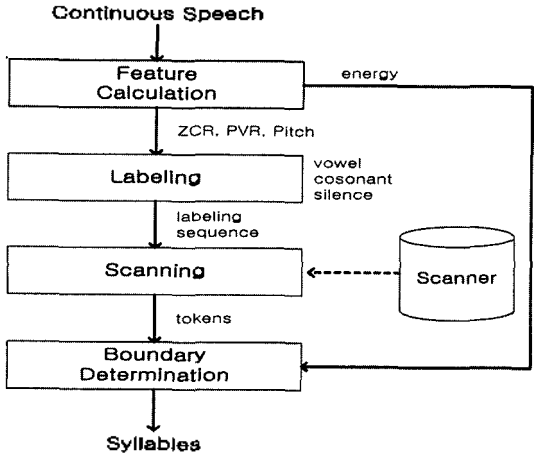


그림 5. 음절분할 전체 순서도
Fig. 5. Syllabic segmentation algorithm.

Q_i로 나타내고 초기상태는 Q_s로 음절의 시작점, 마지막 상태는 Q_f로 음절의 끝점, Q_e는 에러상태이다.

3.3. 3단계: 후처리에 의한 2차분할 경계 결정

라벨열을 스캐닝한 후, 파서로 부터 출력되는 토큰들은 모두 1차적으로 우리말 음절구조를 가진다고 볼 수 있다. 그러나, 연속음성구조에는 조음결합 등으로 두개 혹은 더 많은 음절이 포함되어 있을 수 있다. 이러한 문제는 모음의 몇 가지 특징을 가지고 있는 자음 즉, /l/,/n/,/d/와 같은 반모음을 포함하는 음절과 음절로 연결된 연속음성에서 흔히 관찰되는 특징이다. 이러한 문제는 토큰에 있는 모음부분의 스무딩 에너지 커브의 요철을 검색하고, 음절의 평균지속시간을 임계값으로 설정함으로 해결하였다. 세부 알고리듬은 그림 3와 같다.

먼저, 1차 분할의 결과 생성된 시작프레임 (T_s)와 끝 프레임 T_e를 가지는 각 토큰을 차례로 검색한다. 토큰의 길이 ($T_e - T_s$)가 평균 음절 지속시간 (Avg.Syl)보다 적으면 하나의 음절로 경계를 결정한다. 만약, 그렇지 않다면 하나 이상의 음절이 포함되어 있는 2차분할의 대상 후보가 된다. 2차분할은 토큰내에 포함되어 있는 에너지의 이동평균값에서 획득된 Peak, Valley중에서 Valley의 국부적인 최소값 (MIN_{valley}), 그리고 Peak의 국부적인 최대값 (MAX_{peak})을 의사 음절핵 (pseudo-syllabic nucleus)으로 간주하고 이를 이용한다.

즉, T_s와 T_s+Avg.Syl내에 MAX_{peak}가 존재할 경우 MIN_{valley}를 새로운 경계로 결정하고 T_s가 T_e보다 클때까지 해당 토큰내의 경계를 새롭게 갱신하는 과정을 반복하여 전체 토큰에 대하여 처리하여 최종적인 음절의 경계를

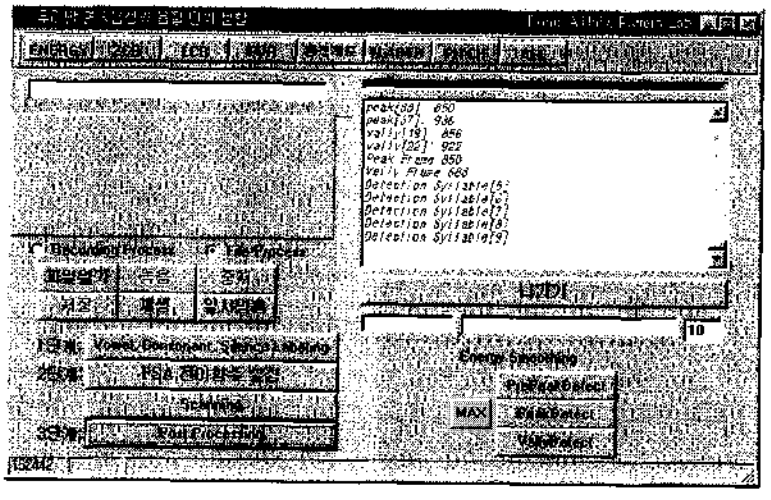
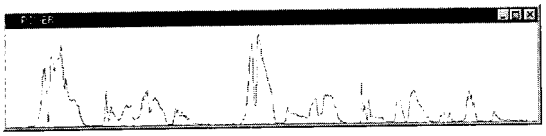
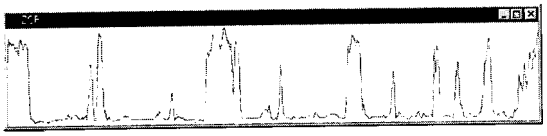


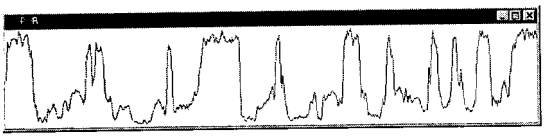
그림 4. 음절분할 시뮬레이션 환경
Fig. 4. Syllabic segmentation simulation tool.



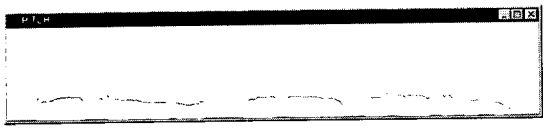
(a) 에너지 (Energy)



(b) 영교차률 (Zero Crossing Rate)

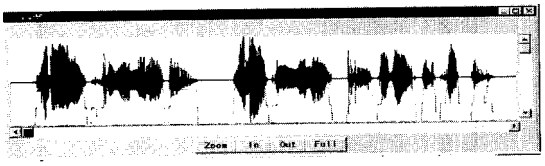


(c) PVR (Peak Valley Rate)

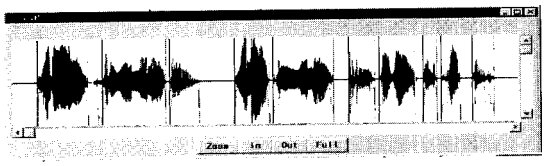


(d) 피치 (Pitch)

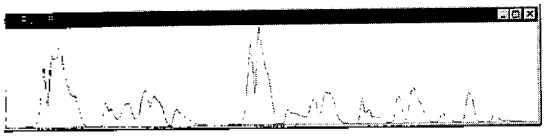
그림 6. 분할 파라메터들
Fig. 6. Segmentation parameters.



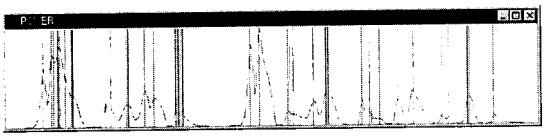
(a) 연속음성 데이터와 라벨링 결과
(Continuous Speech Data and Labeling Result)



(b) 스캐닝 결과 (Scanning Result)

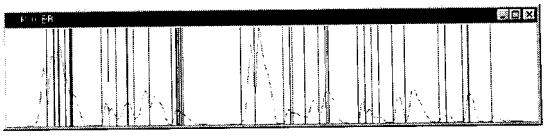


(c) 에너지에 대한 10포인터 이동평균
(10 Point Moving Average for the energy)

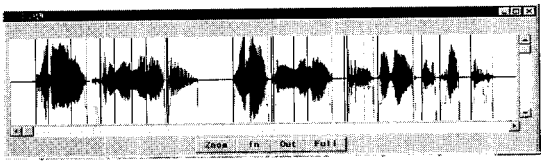


(d) 피크정보에 의한 의사 음절핵 후보 추출
(Extract the pseudo-syllabic nucleus candidates using the peak information)

그림 7. 분할절차에 따른 결과들 (a~d)
Fig. 7. Results according to the segmentation procedures..



(e) Valley에 의한 음절분할 경계 후보 추출
(Extract the boundary candidates of the syllabic segmentation using the velley)



(f) 최종 분할 결과 (Last Segmentation Result)

그림 7. 분할절차에 따른 결과들 (e~f)
Fig. 7. Results according to the segmentation procedures.

결정짓게 된다.

IV. 실험 및 고찰

4.1. 음성 데이타 및 시뮬레이션 도구

본 연구에 사용한 음성 데이터는 실험실에서 자체적으로 채집한 10개의 연속음성 문장과 50개의 연속음성 단어를 사용하였다. 이 음성 데이터는 16kHz, 16bit PCM으로 샘플링 되었으며 발성속도는 문장 데이터는 초당 6-7음절, 단어 데이터는 초당 4-5음절이 되도록 하였다. 실험은 윈도우 API함수를 이용하여 직접 음성을 입출력하도록 자체적으로 음절분할 시뮬레이션 도구(그림 5)를 제작하여 실험하였다.

분할 파라메터 계산은 음성의 PCM데이터에 대하여 Hamming 윈도우(256포인터)로 중첩간격(60포인터)을 두어 단구간 분석하여 프레임별로 분할 파라메터들을 구한다. 음절핵 정보는 10포인터 에너지 이동평균값으로 구하는데 이동평균으로 생기는 총 프레임수의 감소문제는 프레임의 전후에 이동평균 포인터에 따라 "0"값을 가지게 끔 프레임을 보상하는 방식으로 해결하였다. Peak와 Valley 프레임의 검출은 Peak-Peaking법으로 행하였다. 1단계의 라벨링 과정에서 임계값은 실험에 의하여 구하였는데 PVR과 ZCR 모두 최대 PVR, ZCR값의 87.5%의 위치를 임계값으로 설정하였다.

그림 5는 제안한 분할법의 전체 흐름도이고 그림 6,7은 실험 데이터 중 문장1인 [우리의 생활문화가 문화국민의 품격을 잃고있다]에 대한 실험 결과이다.

실험용 데이터에 대한 실험결과를 평가하기 위한 기준

표 1. 문장단위 연속음성 분할결과
Table 1. Segmentation results for the sentence units.

	음절수	삽입	삭제	분할완성
문장1	20	2(10.0%)	3(15.0%)	15(75.0%)
문장2	16	3(18.7%)	2(12.5%)	11(68.7%)
문장3	16	2(12.5%)	2(12.5%)	12(75.0%)
문장4	24	1(4.2%)	2(8.3%)	21(87.5%)
문장5	17	3(17.6%)	3(17.6%)	11(64.7%)
문장6	17	2(11.8%)	2(11.8%)	13(76.5%)
문장7	20	2(10.0%)	4(20.0%)	14(70.0%)
문장8	19	3(15.7%)	2(10.5%)	14(73.7%)
문장9	25	1(4.0%)	6(24.0%)	18(72.0%)
문장10	24	1(4.2%)	6(25.3%)	17(70.8%)
총계	198	20(10.1%)	32(16.1%)	146(73.7%)

표 2. 단어단위 연속음성 분할결과
Table 2. Segmentation results for the word units.

	음절수	삽입	삭제	분할완성
50개 단어	227	28(12.3%)	4(1.8%)	195(85.9%)

으로 삽입, 삭제오류를 기준으로 한다. 삽입오류는 한 음절이 두개 혹은 더 이상의 음절로 나뉘어질 때이며 반대로 한 음절로 하나 혹은 더 이상의 음절이 결합하거나 미검출된 음절은 삭제오류가 된다.

표 1, 2는 각각 10개의 문장단위의 연속음성과 50개의 단어단위 연속음성에 대한 분할결과이다.

V. 결 론

본 논문에서는 우리말 연속음성에 대한 음절단위 분할법을 제안하였다. 본고에서는 우리말의 자음과 모음에 존재하는 음향학적인 특징들만을 이용하여 라벨링하여 토큰(Token)을 형성하는 알고리듬과 음절구조 파서로 스캐닝(Scanning)하여 음절구조로 1차 분할하고 마지막으로 음절핵 정보를 이용하여 두개 혹은 여러 개의 음절이 존재하는 경우의 처리를 위한 알고리듬을 제안하였다.

문장과 단어단위의 연속음성에 대한 분할실험결과 각각 73.7%와 85.9%의 분할률을 보였다. 문장단위의 발성인 경우 발성 속도가 빠르고 많은 조음결합 등의 영향으로 단어단위의 비교적 안정된 발성에 비하여 낮은 분할률을 나타내었다.

우리말의 음절단위 분할은 음절단위 음성인식알고리듬이 연속음성인식 시스템으로 확장하기 위해서는 연구되어야 하는 필수적인 전처리과정이다. 향후, 더 많은 분할 파라메터들을 이용하여 더 나은 분할률을 얻을 수 있는 신뢰할만한 분할법에 대한 연구가 필요하다.

참 고 문 헌

1. Ching-Tang Hsieh, Mu-Chun Su, Eugene Lai,Chih-Hsu Hsu, "A Segmentation Method for Continuous Speech Utilizing Hybrid Neuro-Fuzzy Network," *Jornal of Information Science and Engineering*, 15, 615-628, 1999.
2. Yoav Medan, Eyal Yair, Dan Chazan, "Super Resolution Pitch Determination of Speech Signals," *IEEE Trans. On Signal Processong*, Vol. 39, No. 1, 1991.
3. 신옥근 "음절핵의 위치정보를 이용한 우리말의 음소경계 추출," 한국음향학회지, Vol.11, No.5 ,2000.
4. Lawrence Rabiner, Bing-Hwang Juang, Fundamental of speech recognition, Prentice Hall, 1993.

저자 약력

● 한 학 용 (Hag-Yong Han)

1994년 2월: 동아대학교 전자공학과 (공학사)
1994년~1997년: 경남에너지(주) 근무
1998년 2월: 동아대학교 전자공학과 (공학석사)
1999년~현재: 동아대학교 전자공학과 박사과정
※ 주관심분야: 음성신호처리, DSP응용

● 고 시 영 (Si-Young Koh)

1979년 2월: 영남대학교 전자공학과 (공학사)
1983년 2월: 영남대학교 전자공학과 (공학석사)
1992년 8월: 동아대학교 전자공학과 (공학박사)
1986년~현재: 경일대학교전자정보공학과 교수
※ 주관심분야: 음성신호처리, 생체신호처리

● 허 강 인 (Kang-In Hur)

1980년 2월: 동아대학교 전자공학과 (공학사)
1982년 2월: 동아대학교 전자공학과 (공학석사)
1990년 8월: 경희대학교 전자공학과 (공학박사)
1984년 9월~현재: 동아대학교 전기·전자·컴퓨터공학부 교수
1988년 9월~1989년 8월: 일본 筑波大学 객원연구원
1992년 9월~1993년 8월: 일본 豊橋大学 객원연구원
※ 주관심분야: DSP, 음성인식·합성, 신경회로망