# 잡음이 첨가된 연속음성에서의 자동 음절분할 알고리즘

김영섭, 김창근, 차영동, 이광석, 허강인
동아대학교 전자공학과, 진주산업대학교 전자공학과

# Automatic Syllable Segmentation Algorithm in Noise Additional Continuous Speech

Young-Sub Kim, Young-Dong Cha, Chang-Keun Kim, Kwang-Seok Lee, Kang-In Hur
Dept. of Electronic Engineering, Dong-A University
Dept. of Electronic Engineering, JinJu National University

## 요 약

본 논문에서는 잡음이 첨가된 연속음성에서의 자동 음절분할을 위해 기존에 사용되고 있는 특징 파라미터인 단구간 에너지 이외에 잡음에 강인한 특성을 가지고 있는 새로운 특징인 스펙트럼 밀도비교척도와 의사역행렬을 이용한 선형판별함수를 제안한다. 기존에 사용되는 단구간 에너지는 잡음이 없는 환경에서는 좋은 성능을 나타내지만 잡음환경에서는 그렇지 못하다. 반면에 논문에서 제안한 척도들은 반대의 성능을 가지므로 주변잡음의 크기에 따라 각각의 파라미터를 적절한 가중치로 조합하는 음절구간 결정함수와 유한상태 머신을 추가로 사용면 무 잡음 환경 뿐만 아니라, 잡음이 첨가된 연속음성에서도 일정수준 이상의 음절구간을 분리해 낼 수 있다.

## I. 서 론

음성인식에서 인식의 단위는 크게 음소, 음절, 단어등이 존재한다. 음소단위의 인식은 인식 대상 음소수가 한정되어 있어 기억용량이 한정적이며 계산량도 많지 않은 장점이 있으나 화자의 불규칙적인 발성과 잡음에 아주 민감하며 과도기 구분, 상호조음 현상, 자음의 불규칙적인 주기성 및 처리문제등 난점이 존재한다. 단어단위 인식은 화자의 발음속도, 잡음의 영향, 이음 현상, 그리고 상호조음 현상에 크게 영향을 받지 않고 구현하기가 쉽다 또한 한정된 어휘 하에서는 인식률이 높은 장점이 있으나 표준패턴 수 증가에 따른 기억용량 증가와 계산량의 증가 그리고 실시간 처리가 어려운 문제점 등을 안고 있다. 음절 단위의 인식은 단어와 음소의 중간 형태로 단어에서 표준패턴 수 증가에 따른 기억용량 증가 문제, 음소에서 분류상의 어려움을 회피할 수 있으며 음절내의 정확한 모음인식 결과에 따라 해당 모음을 포함하는 음절들만 인식 대상으로 함으로써 계산시간 단축은 물론 인식률을 증가시킬 수 있다.

본 논문에서는 실시간 뿐 만 아니라 잡음 환경에서도 신뢰성 있는 음절구간 검출을 위하여 기존에 사용되고 있는 단구간 에너지 이외에도 스펙트럼 밀도비교척도 와 의사역행렬을 이용한 선형판별함수를 추가하여 두 종류의 파라미터에 주변잡음의 크기에 따른 각각의 가중치를 곱하여 얻어진 정보를 혼합한 판정함수와 유한상태 머신을 이용해 음절경계를 검출하는 알고리즘을 제안하였다. 제안한 알고리즘의 성능평가를 위해 각기 다른 신호 대 잡음비(SNR)로 구성된 음성으로 실험하였으며 실험 결과 주변 잡음이 존재하는 환경에서도 강인한 특성을 가짐을 확인할 수 있었다.

## II. 본 론

### 1. 음절구간검출 파라미터

1) 단구간 프레임 에너지
(Short Time Frame Energy)

음절분할을 수행하기위해 사용한 첫 번째 특징으로는 각 프레임의 단구간 파워를 모두 합한 에너지를 사용하였으며 식(1)에 나타내었다.

$$E_n = \sum_{m=0}^{N-1} X^2(m) \quad (1)$$

여기서 N은 프레임 길이, $X(m)$는 현 프레임의 음성신호를 의미한다. 에너지는 일반적으로 모음구간은 파워가 크고 비교적 길고 안정된 평탄부를 형성한다. 그리고 에너지 평탄부는 모음구간에 비하여 적지만 무음구간, 비음의 정상부, 또는 마찰음구간에서도 관찰된다. 그러나 에너지만을 가지고 음절구간을 결정할 경우 잡음이 없는 환경에서는 음절구간 검출에 좋은 성능을 보이지만 잡음환경에서는 잡음신호와 음성의 시작부분을 주로 구성하는 무성음 구간과의 구별이 어렵게 된다, 또한 무성음으로 끝나는 단어의 끝점검출에서도 같은 어려움이 존재한다.

2) 스펙트럼 밀도비교 척도
(Spectral Density Comparison Measure)

두 번째로 사용한 특징으로는 MFCC에서 잡음의 추정치를 뺀 스펙트럼 밀도비교 척도를 사용하였으며 아래 식(2)에 나타내었다.

$$SD = \sum_{m=0}^{M-1} (X_m^2 - N_m^2) \quad (2)$$

여기서 $X_m$은 에너지를 포함한 음성신호의 MFCC를 의미하고, $N_m$은 잡음의 추정치, $M$은 MFCC의 차수를 의미한다. 잡음이 첨가된 환경에서는 무성음과 잡음의 구별이 어렵기 때문에 음성신호에서 미리 잡음으로 추정되는 부분을 수집하여 그것의 평균치를 빼줌으로써 음성신호에 존재하는 잡음성분을 차감하는 효과를 나타낼 수 있다.

3) 선형판별함수
(Linear Discrimination Function)

마지막으로 사용한 특징벡터로는 식(3)으로 표현되는 선형판별함수를 사용하였다.

$$Y = \sum_{m=1}^{N-1} W_m X_m + W_0 \quad (3)$$

여기서 $X_m$은 입력음성 각 프레임의 MFCC, $N$는 MFCC의 차수, $W_m$는 선형함수의 가중치, $W_0$는 성향(bias)을 의미한다. $W_m$을 구하기 위해서는 아래 식(4)로 표현되는 의사 역행렬을 이용한다.

$$W = (X^T X)^{-1} X^T Y \quad (4)$$

여기서 X는 MFCC이며 Y는 출력신호, W는 가중치를 의미한다. 가중치를 구하기 위해서는 기존의 음성구간과 잡음 또는 무음구간으로 구성되어진 음성신호를 수집하여 입력벡터 X가 음성구간이면 출력신호 Y는 1, 잡음이나 무음구간이면 출력신호 Y는 0이 되도록 학습하여야 한다. 이와 같이 학습하는 이유는 식(4)를 이용하여 구한 가중치벡터를 식(3)에 적용하였을 때 얻어지는 출력 값이 음성구간에서는 1을 출력하고, 잡음이나 무음구간에서는 0을 출력하도록 하기 위해서이다.

## 2. 음성 유/무를 위한 임계치 결정

잡음이 존재하는 음성신호에서 음절의 경계를 결정하기 위해서는 먼저 제안한 3가지의 특징벡터의 이동평균값을 구해야한다. 이것은 신호에 포함되어있는 잡음성분은 최소로 하고 음성구간에 관한 정보만을 취하기 위함이며 이는 아래 식(5)로 표현된다.

$$\overline{p_j}(n) = \alpha p_j(n) + (1-\alpha)\overline{p_j}(n-1), \; j = 1,2,3 \quad (5)$$

여기서 $p_1$, $p_2$, $p_3$는 각각 에너지, 밀도비교 척도, 선형판별함수의 출력값을 나타내며, α값은 실험에 의해 구해진 값이다. 이 α 값에 따라 특징벡터들의 변동량을 결정할 수 있다. 또한 음성구간과 잡음 또는 무음구간을 결정하는 임계값은 아래 식(6)으로 정의한다.

$$T_j(n) = a_j \overline{p_j}(n) + b_j, \quad j = 1,2,3 \quad (6)$$

여기서 $T_j(n)$은 음성신호와 무음, 잡음을 구분하기 위한 임계치를 나타내고, $\overline{p_j}(n)$는 잡음이 첨가된 음성신호를 나타낸다. 여기서 적절한 $T_j(n)$을 구하기 위해서는 앞부분에 정의되어 있는 식(4)를 이용하여 적절한 $a_j$와 $b_j$를 학습을 통하여 구하여야한다. 여기까지 과정을 아래 그림 1과 같이 나타내었다.

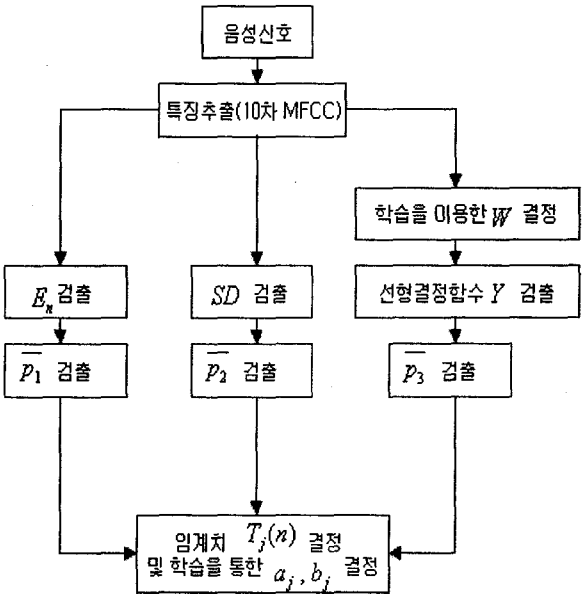


그림 1. 음절경계 결정을 위한 특징벡터 검출과정

### 3. 음절경계 결정함수

일반적으로 우리말에서의 가능한 음절은 단모음과 이중모음으로 이루어진 모음과 파열음, 파찰음, 마찰음, 비음, 설측음, 탄설음으로 이루어진 자음의 결합으로 구성되며, 반드시 하나의 음절은 하나의 음절핵인 모음을 포함하게 된다. 그러므로 시간영역의 음향 특징인 단구간 에너지와 본 논문에서 제안한 스펙트럼 밀도비교 척도, 선형결정함수를 이용한 특징 벡터 그리고 임계치를 이용하여 그림 2로 나타낸 과정을 수행한다.

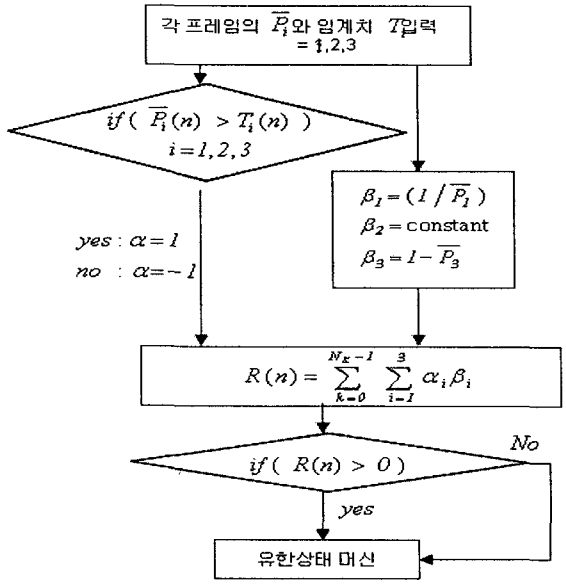


그림 2. 음성/비 음성 결정 과정

각각의 특징벡터에 서로 다른 가중치를 부여한 새로운 특징 벡터를 검출하고 이를 음성과 비음성으로 구분한 뒤 음절이 가지고 있는 평균 지속시간을 이용하여 음절경계를 결정할 수 있게 된다. 이를 구현하기 위하여 아래 그림 3으로 표현되는 유한상태 머신을 이용하여 구현할 수 있다.

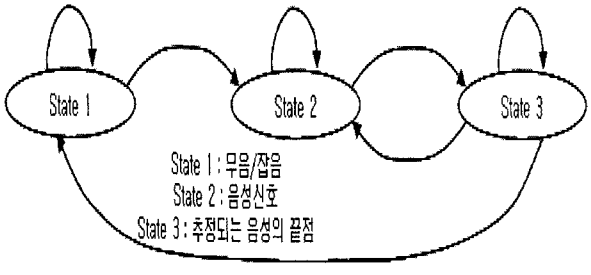


그림 3. 유한상태 머신

최초 신호의 상태를 state1로 간주하고 입력되는 $R(n)>0$면, 음성신호가 검출된 것으로 간주 하고 상태를 state2로 바꾼다. 그리고 $R(n)>0$인 프레임의 수를 카운트하여 실험으로 구한 평균 음절의 지속시간을 적용하여 음절이라고 판단될 때 원 음성신호에서 계산되어진 구간 까지를 절단하여 실시간으로 저장 공간에 저장하게 된다. 또한 $R(n)<0$ 는, 음성이 아닌 것으로 간주하여 상태를 state3로 바꾼다. 만일 음절 구간 내에 잘못 인식된 부분이 있을 수 있기 때문에 이 부분에서도 실험으로 구한 음절과 음절 사이의 임계치 및 무음으로 판단되는 프레임의 지속 시간을 적용하여 상태를 state1로 바꾸는 과정을 입력신호의 끝까지 수행을 하면서 음절로 판단되는 부분을 절단하여 저장하게 된다.

## Ⅲ. 실험 및 결과

실험에 사용한 데이터는 연구실에서 컴퓨터로 녹음한 연속 숫자 발성음과 10음절에서 15음절 사이의 낭독체 문장을 사용하였으며, 잡음환경 구현을 위해 백색노이즈로 5dB, 15dB, 25dB의 서로 다른 신호 대 잡음비(SNR)로 학습 데이터와 테스트 데이터를 만들었으며, 식 4에 나타내어진 가중치를 구하기 위한 학습 데이터는 총6명이 0부터 9까지의 숫자음 발성에 대하여 음절별로 저장하였으며, 낭독체 문장에 대해서는 각 음절별로 데이터를 절단하여 학습데이터를 만들었다. 테스트에는 학습자 이외의 10명이 각자 10문장으로 음절 분할 작업을 실행하였다. 아래 표 1은 음성데이터의 분석 조건을 나타내었으며, 그림 4는 본 논문에서 사용한 3가지 특징 벡터들의 스펙트럼을 보여준다. 그림 5는 각기 다른 신호 대 잡음비(SNR)에서의 음절분할을 수행한 결과를 나타내었다, 그림 5는 SNR이 (5dB)에서 연속 숫자음 발성을 통해 음절분할 한 결과스펙트럼을 보여주고 있으며. 마지막으로 실험용·데이터에 대한 실험결과의 평가는 삽입, 삭제오류를 기준으로 하였다, 삽입오류는 한 음절이 두 개 혹은 더 이상의 음절로 나뉘지는 경우이며 삭제오류는 한 음절로 하나 혹은 더 이상의 음절이 결합하거나 검출되지 않는 경우이다. 표 2는 10개의 문장단위의 연속음성에 대한 평가 결과이다.

표 1.음성데이터의 분석 조건

| | |
|---|---|
| Speech Format | PCM Raw data |
| A/D Conversion | 16kHz , 16bit |
| Frame Size | 384 samples (24ms) |
| Overlap Size | 128 samples (8ms) |
| Pre-emphasis | 0.97 |
| FFT Size | 512 (zero padding) |
| Mel Filter Bank Number | 20 |
| MFCC Order | 10(except power) |

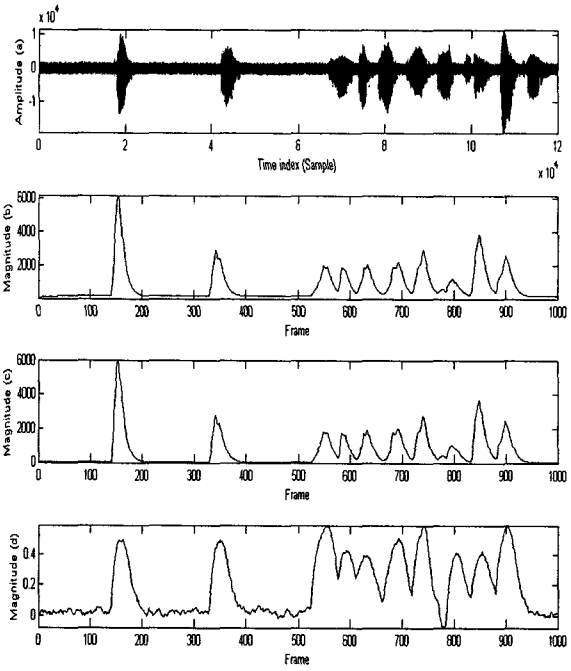


그림 4. (a)연속 숫자음 발성에 대한 음성신호, (b) 단구간 에너지, (c) 스펙트럼 밀도비교 척도, (d) 선형결정함수를 이용한 구간척도

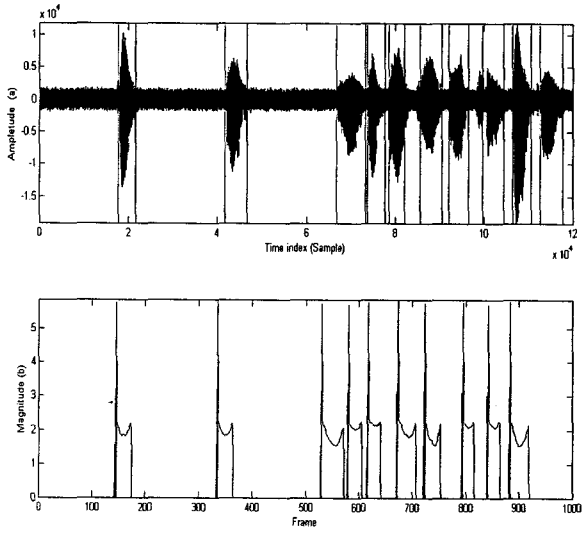


그림 5. (a)연속 숫자음 발성에 대한 음성신호, (b) 음절결정함수를 이용한 구간결정 척도

표 2. 10개 문장단위의 음절분할에 대한 실험결과

| | 5dB | 15dB | 25dB |
|---|---|---|---|
| 음절수 | 198 | 198 | 198 |
| 삽입(Insert) | 8(4%) | 11(5.6%) | 11(5.6%) |
| 삭제(Delete) | 48(24.2%) | 46(23.2%) | 44(22.2%) |
| 분할완성 | 142(71.7%) | 141(71.2%) | 143(72.2%) |

## Ⅳ. 결 론

본 논문에서는 잡음이 첨가된 연속음성에서의 적은 계산량으로 실시간 음석인식기를 구연함에 있어서의 전처리 단계인 음절분할 알고리즘을 소개하였으며 이는 서로 다른 분석기법을 조합하여 최종적으로 잡음에 강인한 새로운 척도를 구하는 것이었다. 그러나 실험결과로 확인할 수 있듯이 서로 다른 신호 대 잡음비(SNR)에서 는 강인한 성능을 보여주지만 전체적인 성능 면에서는 아직 보안해야 할 부분이 많음을 알 수 있었다.

앞으로 더 연구되어야 할 부분은 실험결과에서 삽입오류보다 삭제 오류가 4배 이상 많은 점에서 알 수 있듯이 음절분할의 오류가 발생할 수 있는 상호조음 현상을 어떤 식으로 해결하는가 하는 것이며 이는 반드시 연구되어야 하는 음성학적인 지식 표현의 분야이다. 따라서 음운의 변화를 해결할 수 있는 직접적인 방법 즉, 음운변화가 발생하는 지점을 찾을 수 있도록 특정 규칙에 바탕을 둔 기법을 도입하는 접근 방법을 찾아 계속 발전시켜 나아갈 것이다.

## 참 고 문 헌

[1] 김창근, 박정원, 권호민, 허강인, "음성인식기 구현을 위한 잡음에 강인한 음성구간 검출기법," "한국 신호처리 시스템학회 학술논문집," 1229-9480, 제4권2호, pp.18-24 , 2003

[2] L.R.Rabiner, R.W.Schafer, "Digital processing of speech signals," Prentice Hall, 1978

[3] 한학용, "우리말 음성의 최적 자동분할과 인식에 관한 연구," "공학박사학위논문," 동아대학교 대학원, 2004

[4] 이길행, 이윤준, "한국어 음성특징을 이용한 3단계 음절분할 알고리즘," "정보과학회 가을 학술발표논문집," 14권2호, 1987

[5] Richard O.Duda, Peter E.Hart, David G.Stork, "Pattern Classification(Second Edition)," A Wiley Interscience Publication, 2001

[6] 정국, 구희산, 이찬도, 김종미, 한선희, "음성 인식/합성을 위한 국어의 음성-음운론적 특성연구," "한국음향학회지," 제13권6호, pp76-84, 1994

[7] A. Ganapathiraju, J. Hamaker, J.picone, M. Ordowski, and G..R. Doddington, "Syllable-based Large Vocabulary Continuous Speech Recognition," IEEE Transaction on Speech and Audio Processing, Vol.9 No. 4, pp.358-366, May 2001