논문 2009-46SC-1-4

리아프노프 지수를 이용한 음성신호 종점 탐색 방법

(Endpoint Detection of Speech Signal Using Lyapunov Exponent)

장 한*, 김 정 연*, 정 길 도**

(Xian Zang, Jeong Yeon Kim, and Kil To Chong)

요 약

음성 인식 연구에서 잡음이 존재하는 음성 발음의 처음과 끝을 찾아내는 것은 매우 중요하다. 음성 종점 탐지를 위한 기존의 방식으로는 2개의 간단한 시간 영역 측정법인 단시간 에너지와 단시간 영점교차 비율 방법이 있다. 위의 방법들은 낮은 신호 대 잡음비의 환경에서는 정확한 결과를 보장 할 수 없기 때문에 본 논문에서는 시간 영역 파형의 리아프노프 지수를 이용하여 음성의 시작과 종점을 구별하는 새로운 접근법을 제시하였다. 제안한 방법은 Mel-Scale 특징 방법에서 요구되는 종점 탐지 과정을 위한 주파수 영역 매개변수를 얻는 과정이 필요 없기 때문에 보다 간단하다. 제안한 방법의 성능 검증을 위해 아라비아 숫자의 음성단어 분석에 적용해 보았으며, 결과를 통하여 제안한 방법이 인식률을 현저히 증가시킴을 확인하였다.

Abstract

In the research of speech recognition, locating the beginning and end of a speech utterance in a background of noise is of great importance. The conventional methods for speech endpoint detection are based on two simple time-domain measurements - short-time energy, and short-time zero-crossing rate, which couldn't guarantee the precise results if in the low signal-to-noise ratio environments. This paper proposes a novel approach that finds the Lyapunov exponent of time-domain waveform. This proposed method has no use for obtaining the frequency-domain parameters for endpoint detection process, e.g. Mel-Scale Features, which have been introduced in other paper. Accordingly, this algorithm is low complexity and suitable for Digital Isolated Word Recognition System.

Keywords: Digital Isolated Word Recognition; Time-domain; Time-dependent Lyapunov exponent

I. 서 론

배후 잡음으로부터 음성을 구별하기 위한 음성 발음의 종점 탐지는 음성 처리 과정에서 매우 중요한 기술이다. 정확한 종점 위치 정보는 음성 인식 정확도를 향상시킨다. 특히, 고립 단어에서 자동인식 시스템을 사용할 경우 발생하는 주요 오류는 시험데이터와 참고 템플릿의 시작과 끝 경계의 부정확한 정보에서 상당수 기인된다. 따라서 각 낱말에 대응하는 음성 신호의 영역을 찾아내는 것은 매우 중요하다. 또한 음성 신호의 시작과 끝점을 찾아내는 과정은 음성 신호에서 입력의 부분을 구분할 수 있게 하므로 많은 계산량을 제거하기 위하여 사용될 수도 있다.

기존의 종점 검출 방법은 잡음이 적은 순수한 음성 신호의 경우 실행할 수 있는 간단한 에너지 탐지 방법을 주로 사용했다. 예를 들면 높은 신호 대 잡음 환경의 경우에 적용하는 단시간 에너지 및 영점교차 비율 방법은 시작과 끝 점을 찾아내는 알고리즘을 활용할 수 있지만, 잡음 환경에서는 효율이 현저히 감소된다.

지금까지 인식 정확도를 개선하기 위해 종점 탐지의 많은 연구가 수행된바 있다. 그러나 대부분의 알고리즘은 여러 부분에서 많은 시간을 소모하고 실시간 음성

* 정회원, 전북대학교 제어계측공학과
(Control and Instrumentation Department, Chonbuk National University)
** 정회원-교신저자, 전북대학교 전자정보공학부
(Electronics and Information Departemtn, Chonbuk National University)
접수일자: 2008년11월22일, 수정완료일: 2009년1월12일

체계의 적응성을 감소시키는 다량의 계산을 요구한다.

본 연구에서 제안한 방법은 시간 영역 음성 신호의 리아프노프 지수 계산을 통하여 시간 영역에서 특징을 추출하는 방법을 제안하였다. 이 알고리즘에 바탕을 두고, 패턴 인식 과정에서 매치 과정을 찾기 위해 특성 매개변수로 멜-주파수 켑스트럴 계수(MFCC, Mel-Frequency CeDStral Coefficients)[1~3]를 사용하였다. 또한 동적 시간 왜곡(DTW, Dynamic Time Warping) 알고리즘을 적용하였다[4~7]. 실험을 통하여 디지털 고립 단어 인식 체계에 있어서 제안한 종점 검출 방법의 성능을 확인하였다.

II. 본 론

1. 기존의 종점 검출 방법

음성 발음의 종점 검출의 문제 해결을 위한 기존의 알고리즘은 일반적으로 두 종류로 분류되며, 간단한 시간 영역 측정, 즉, 단시간 에너지와 단시간 영점교차 비율이 있다. 기존의 연구 내용[8]을 살펴보면 위에서 언급한 두 방법을 활용하여 새로운 알고리즘을 제안하였다. 예를 들어 만약 높은 신호 대 잡음비 환경이라면 시작점과 종점의 위치를 찾기 위한 유용한 알고리즘으로써 단시간 에너지와 영점교차 비율 방법의 결합을 들 수 있다. 그러나 잡음비율이 높은 환경에서는 그 결합에 의한 효과가 감소된다.

무성 또는 잡음 영역 진폭이 일반적으로 유성 영역의 진폭보다는 매우 낮다는 것을 알 수 있다. 음성 신호의 단시간 에너지는 이 진폭 변이를 반영하는 편리한 방법을 제공한다. 일반적으로 단시간 에너지는 다음과 같이 정의할 수 있다.

$$Em = \sum_{n=m}^{m+N-1} s_w^2(n) \tag{1}$$

여기에서, $S_w(n)$은 윈도우를 설정한 후의 음성 신호이다. 식 (1)은 다시 식 (2)와 같이 표현할 수 있다.

$$Em = \sum_{n=m}^{m+N-1} \hat{s^2}(n) \cdot h(n-m) \tag{2}$$

여기에 다음과 같은 공식이 적용되고,

$$h(n) = w^2(n) \tag{3}$$

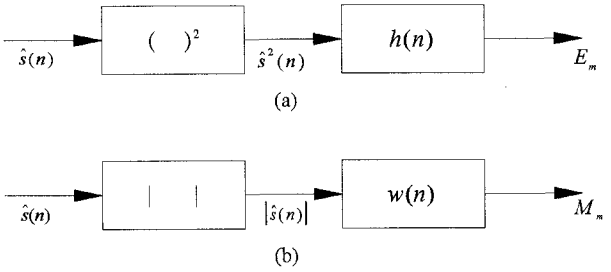


그림 1. (a)단시간 에너지의 블록 다이어그램; (b)단시간 평균 크기

Fig. 1. Block diagram representation of (a) the short-time energy; and (b) the short-time average magnitude.

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) : 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad\qquad\quad : else \end{cases} \tag{4}$$

여기에서 $w(n)$는 해밍(Hamming) 윈도우이다. 은 한 개의 창에 존재하는 수로써 256과 동등하다. 식 (2)는 그림 1(a)와 같이 표현될 수 있다. 즉, 신호 $\hat{s^2}(n)$은 식 (3)에 주어진 임펄스 응답을 갖는 $h(n)$선형 필터에 의해 여과되는 것을 볼 수 있다.

음성 신호의 단시간 평균 크기는 식 (5)와같이 표현할 수 있다.

$$Mm = \sum_{n=m}^{m+N-1} |s_w(n)| = \sum_{n=m}^{m+N-1} |\hat{s}(n)| \cdot w(n-m) \tag{5}$$

여기에서 신호의 절대 값의 총 합을 제곱 합 대신에 계산할 수 있다.

그림 1(b)는 어떻게 식 (5)가 $|\hat{s}(n)|$상에서 선형 필터처럼 실행되는 지를 보여준다. 또 다른 매개변수는 단시간 영점교차 비율이다. 대략적 정의는 식 (6)과 같다.

$$Z_m = \frac{1}{2}\sum_{n-1}^{N-1} |sgn[s_w(n)] - sgn[s_w(n-1)]| \tag{6}$$

여기에서,

$$sgn[x] = \begin{cases} 1 & , x \ge 0 \\ -1 & , x < 0 \end{cases} \tag{7}$$

이다.

그림 2는 식(6)에 관련된 동작을 블록선도로 표현한 것이다.

두 종류의 시간 영역 표현 조합에 기초를 두어, 음성 신호의 시작점과 종점을 구별하기 위한 알고리즘은 고

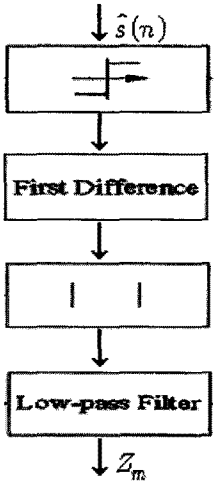


그림 2. 단시간 영점교차 비율의 블록다이어그램
Fig. 2. Block diagram representation of short-time zero-crossing rate.

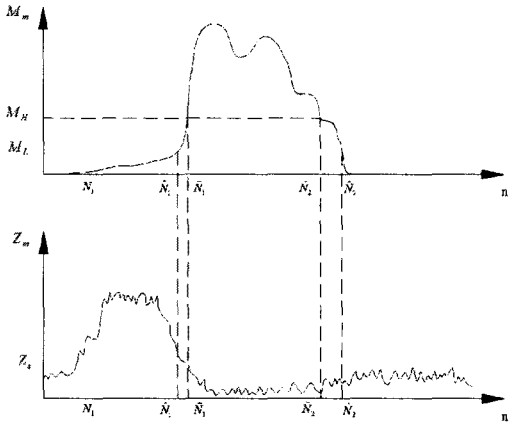


그림 3. 고립 단어 음성의 시작과 끝을 탐지하기 위한 단시간 평균 크기 및 단시간 영점교차 비율 측정
Fig. 3. Short-time average magnitude and short-time zero-crossing rate measurements for detecting the beginning and end of an isolated word speech.

립 단어 음성-인식 체계의 문맥에서 Rabiner와 Sambur [9~10]에 의해 연구되었다. 그림 3에 알고리즘의 방법을 나타내었다. 단시간 평균 크기 방법은 매우 보편적인 문턱값(그림 3의 M_H)을 항상 초과하는 구간을 찾는 방법으로, 시작점과 종점이 이 구간의 바깥에 놓여있음을 확인할 수 있다. 즉 M_H에 따라 $\widetilde{N}_1$과 $\widetilde{N}_2$사이가 음성 구간임을 알 수 있다. 그 후에 $\widetilde{N}_1$으로부터 뒤로 검색단계, M_M점을 맨 처음 만나는 낮은 문턱 값(그림3에서 N_1) M_L보다 작은 점 M_M을 시작점으로 임시 설정한다. 유사한 방법으로 $\widehat{N}_2$를 임시 종점으로 설정한다. 이중 문턱 값 과정은 평균 크기 기능에 있는 복각이 틀린 종점을 출력하지 않는다는 것을 보장한다. 이 단계에서는 시작점과 종점이 $\widehat{N}_1$에서 $\widehat{N}_2$ 간격 안에 없다고 가정하는 것이 비교적 안전하다. 다음 단계는 단시간 영점교차 비율을 문턱 값(그림 3에서 Z_s) 과 비교하면서 $\widehat{N}_1$에서부터 뒤에서 앞으로 검사한다. 이것은 $\widehat{N}_1$을 수행하는데 있어서 25ms로 제한된다. 만약 영점교차 비율이 3개 이상 시간 문턱을 초과하는 경우라면 시작점 $\widehat{N}_1$이 첫 번째 그림 3에서 N_1점인 시작점으로 뒤로 이동하고 그 점은 영점교차 문턱 값이 초과된다. 그렇지 않으면 $\widehat{N}_1$을 시작점으로 정의한다. 유사한 절차를 종점에서도 수행한다.

2. 제안한 종점 탐지 방법

잘 알려진 대로 수학에서 동적 시스템의 리아프노프 지수는 미소하게 가까운 궤적과의 거리 비율을 특징으로 선택하는 것으로써 많은 계산량을 요구한다. 처음 떨어진 위상 공간에 있는 궤도를 δZ_o로 표시하면 다음과 같다.

$$|\delta Z(t)| \approx e^{\lambda t}|\delta Z_o| \tag{8}$$

여기에서 시스템의 복잡한 정도를 표현하는 리아프노프 지수[11~12]는 λ이다. 이 값이 적다면 보다 덜 복잡한 시스템이 된다.

본 논문에서는 리아프노프 지수의 이론적 원리를 적용하여 음성인식 시스템을 설계하였다. 실험을 통하여 데이터를 수집하였으며, 음성 신호를 1초당 $8KHz$로 샘플링 하여 시간영역상의 8000개의 분산된 데이터를 수집하였다. 음성 분석을 통해 상위 리아프노프 지수를 이용하여 전체 간격에서 잡음 부분을 추출할 수 있도록 하였고 하위 리아프노프 지수로는 음성 부분을 추출하였다. 이러한 방법으로 배경 잡음을 제거할 수 있는 상위 리아프노프 지수 필터의 문턱 값을 결정할 수 있다. 따라서 본 알고리즘은 종점을 검색하는 편리한 방법을 제공한다.

알고리즘의 세부 과정은 다음과 같다:

(가) 종점 탐지를 하기 전에, 기존의 방법처럼 몇 가지 전 처리 작업을 수행해야 한다.

(1) 필터링 작업 - 저주파필터 $f_s/2(4kHz, f_s$ 샘플링 주파수)와 고주파 필터 $50Hz$

(2) 표본 추출

샘플링 주파수 $f_s = 8kHz$와 8비트 양자화 사용.

(3) A/D 변환

아날로그 음성 신호를 디지털화한 후에, 일련의 음성 샘플 $s(n)$을 얻는다.

(4) 프리엠퍼시스

신호 스펙트럼을 평활화하기 위하여 식 (9)를 사용하여 음성에너지의 옥타브당 $-6dB$를 제거한다.

$$\hat{s}(n) = s(n) - 0.97s(n-1) \tag{9}$$

(5) 분할

프레임 길이는 $32ms$(256 샘플)로 설정하였고 인접한 2개의 프레임은 $16ms$(128 샘플))를 중첩한다. 따라서 61프레임을 획득하였다.

(6) 윈도우설정

각 프레임에 256 포인트 허밍 윈도우를 적용하였다. 식 (4)에 허밍 윈도우의 수학식을 표현하였다.

(나) 첫 번째 프레임을 선택하고 프레임 내에서 최대와 최소의 진폭을 구한다.

(다) 최대와 최소 사이를 256 구역으로 분할한다. 분할한 형태는 그림 4와 같다.

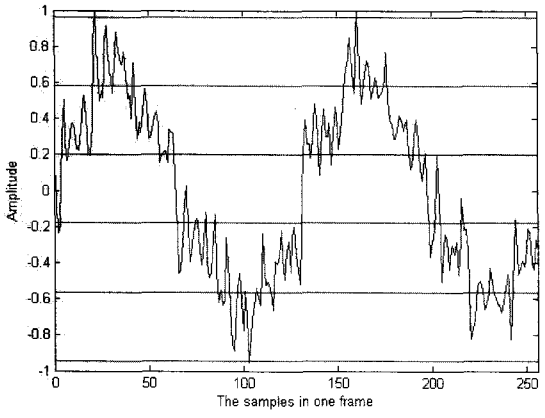


그림 4. 각 프레임에서 분활한 개략도 (실제는 256개로 분활한다.)

Fig. 4. The sketch map of segmentation in each frame (In fact, the number of red line is 256).

(라) 256개로 나누어진 각 영역에서 한 영역에 존재하는 샘플의 수를 구한 다음 처음 시작점을 최소의 진폭으로 설정하고 나머지 255개 영역의 샘플의 수를 반복하여 구한다. 각 단계에서는 다음의 작업을 진행한다.

(1) 샘플수 n을 조사하여 만약에 $n > 2$이면 그 영역의 첫 번째 샘플로부터 2개의 인접한 샘플사이의 진폭 분산 d0', d1', d2', … dn'을 계산한다.

(2) 샘플들의 각 쌍에 단계 a를 적용하는 동안 또 다른 2개의 샘플들을 찾는 과정을 거친다. 만약 영역사이에 2개의 샘플을 찾을 수 없다면 새로운 샘플 쌍의 진폭 분산 d0', d1', d2', … dn'을 계산한다.

(3) 식 10을 사용하여 리아프노프 지수를 계산한다.

$$lyapunov = \frac{\sum_{i=0}^{n} \log_2(d'(i)/d(i))}{n+1} \tag{10}$$

(4) a)단계를 반복한다.

(5) 256단계를 끝낸 후에, 리아프노프 지수의 평균값을 구한다.

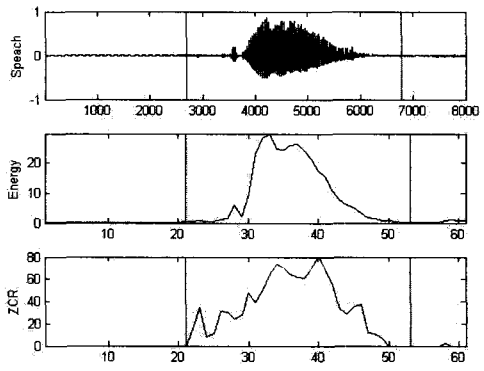
(마) 단계 2로 넘어가서 다음 프레임에 위의 과정을 반복하여 총 61프레임을 반복한다.

(바) 잡음 부분을 제거하기 위하여 61 리아프노프 지수 사이의 문턱 값을 설정한다. 따라서 음성과 배경잡음을 구별할 수 있다.

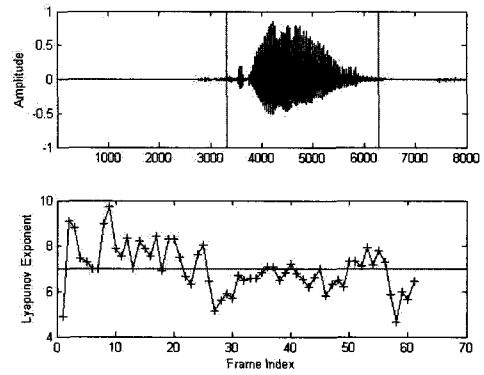
III. 실 험

영어 숫자 0~9를 기록하기 위해 소프트웨어 "GoldWave"를 사용하였다. 각 음성 신호는 $4KHz$로 저주파 필터링 하였다. 샘플링 주파수는 $8KHz$이며 8bit로 양자화하였다. 각 프레임의 길이는 32ms(256 샘플)로 설정하였고, 2개의 인접한 프레임사이에 16ms(128 샘플) 중첩되어 있다. 그리고 데이터 포인트를 선택하기 위해 각 프레임에 250 포인트 허밍 윈도우를 적용하였다. 제안한 종점 탐지 방법을 사용하여 무성영역과 배경 잡음 영역을 제거하였다. 그리고 각 프레임에 특성 매개변수로 24차 MFCC 적용하였다. 마지막 단계로 DTW 알고리즘을 사용하여 패턴 인식 과정을 수행하였다.

숫자 0~9의 테스트 샘플을 사용하여 시뮬레이션을 진행하였다. 그림 5는 기존의 종점 탐지 방법과 제안한 방법의 성능 비교 결과를 나타내었다. 결과에서, 우리는 리아프노프 시간종속 지수에 근거한 종점 탐지의 성과가 기존의 방법보다 향상되고 정확성이 증가됨을 볼 수 있다. MFCC와 DTW가 결합된 인식 비율은 93% 이상을 보였다.



(a)



(b)

그림 5. (a)기존의 알고리즘을 사용한 종점 탐지 결과 ; (b)제시된 알고리즘을 사용한 종점 탐지 결과

Fig. 5. The endpoint detection result (a) using conventional algorithm; (b) using the proposed algorithm(b)

IV. 결 론

숫자 음성 분석 시스템은 음성의 시작과 끝 점에서 매우 높은 정확도를 요구한다. 종점 탐지는 음성 인식에 있어서 중요한 방법이며 또한 난해한 연구방법이다. 본 논문에서는 종점 탐지의 성능을 향상하기 위해 시간 종속 리아프노프 지수에 근거한 종점 탐지 방법을 제안하였으며, 시뮬레이션 결과를 통해 배경 잡음을 가진 음성에서도 향상된 식별 정확도를 얻을 수 있었다. 향후 연구로는, 숫자 음성뿐만 아니라 다양한 음성에서 탐색 능력을 향상시킨 방법에 대해 연구할 것이며 정확도를 향상시킬 수 있는 방안에 대해 연구하고자 한다.

참 고 문 헌

[1] Joseph W. Picone, "*Signal Modeling Techniques in Speech Recognition*", Proceeding of the IEEE, vol.81, No.9, pages 1215-1247, 1993.

[2] Steven B. Davis and Paul Mermelstein, "*Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, No.4, August 1980.

[3] M. De Wachter, M. Matton, K. Demuynck, P.Wambacq, R. Cools, D. Van Comernolle, "*Template-based Continuous Speech Recognition*", IEEE Trans. ASLP 15 (2007) 1377-1390.

[4] F. Itakura, "*Minimum prediction residual principle applied to speech recognition*", IEEE Trans. ASSP 23 (1975) 67-72.

[5] H. Sakoe, S. Chiba, "*Dynamic programming algorithm optimization for spoken work recognition*", IEEE Trans. ASSP 26(1978) 43-49.

[6] H. Silverman, D. Morgan, "*The application of dynamic programming to connected speech segmentation*", IEEE ASSP Mag. 7, no.3(1990) 7-25.

[7] Zebulum, R.S.; Vellasco, M.; Perelmuter, G.; Pacheco, M.A.; "*A COMPARISON OF DIFFERENT SPECTRAL ANALYSIS MODELS FOR SPEECH RECOGNITION USING NEURAL NETWORKS*", IEEE 39th Midwest symposium on Circuits and Systems, 1996., Volume 3, 18-21 Aug. 1996 Page(s):1428 - 1431 vol.3.

[8] L. R. Rabiner and M. R. Sambur, "*An Algorithm for Determining the Endpoints of Isolated Utterances*", Bell Syst. Tech. J., vol. 54, No. 2, pp. 297-315, February 1975.

[9] M. R. Sambur and L. R. Rabiner, "*A Speaker Independent Digit-Recognition System*", Bell Syst. Tech. J., vol. 54, No. 1, pp. 81-102, January 1975.

[10] Kokkinos, I.; Maragos, P., "*Nonlinear speech analysis using models for chaotic systems*", Speech and Audio Processing, IEEE, volume 13, Issue 6, Nov. 2005 Page(s): 1098-1109.

[11] Adriano. Petry, D. A. C. Barone, "*Preliminary experiments in speaker verification using time-dependent largest Lyapunov expontent*", Computer Speech and Language, 17 (2003), 403-413.

저 자 소 개

장 한(학생회원)
2007년 중국 명문대학교 회해 공과대학 공정장비제어 학과 학사졸업.
2008년 현재 전북대학교 전자정보 공학부 석사과정

<주관심분야 : Mechanical Engineering, auto theory, automotive design, electronic control technology.>

정 길 도(정회원)
1984년 Oregon State University 기계공학 학사졸업.
1986년 Georgia Institute of Technology 기계공학 석사졸업.
1992년 Texas A&M University 기계공학 박사 졸업.
2008년 현재 전북대학교 전자정보 교수
<주관심분야 : Time-Delay, Robotics, 인공지능, Web 기술>

김 정 연(학생회원)
2005년 전북대학교 산업정보 시스템공학과 학사졸업.
2008년 현재 전북대학교 전자정보 공학부 석사과정
<주관심분야 : RADAR, Navigation, Robotics >