음성의 유성음 특성을 이용한 음성/비음성 판별 방법

이성주*, 정호영*, 이윤근*, 김형순**
*한국전자통신연구원
**부산대학교
e-mail : lee1862@etri.re.kr

A Robust Speech/Non-Speech Decision Using Voiced Characteristics of Speech

Sung Joo Lee*, Ho-Young Jung*, Yun-Keun Lee*, Hyung Soon Kim**
*Electronics and Telecommunications Research Institute (ETRI)
**Pusan National University (PNU)

요 약

자동음성인식 시스템을 이용하는 사용자 입장에서 보면 음성인식시스템을 사용하기 위하여 음성을 입력할 때마다 버튼을 눌러야 하는 Push-To-Talk (PTT) 방식은 여간 번거로운 일이 아닐 수 없다. 그리고 사용자가 원거리에서 음성을 입력하는 경우처럼 PTT 방식 자체가 용이하지 못 한 음성인식 응용분야에서는 Non-Push-To-Talk (NON-PTT) 방식의 필요성이 대두되게 된다. NON-PTT 방식의 음성 전처리를 위해서는 입력신호로부터 음성신호만을 구분해내는 음성판별기술이 필수적이다. 하지만 일상적인 잡음환경에서 음성신호만을 구분해내는 일은 매우 어려운 일이 아닐 수 없다. 본 논문에서는 일상적인 가정잡음환경에 강인한 음성판별방식을 제안한다. 여기서는 음성판별을 위해서 음성의 유성음 특성을 이용하였다. 즉, 일정구간 이상의 음성신호에는 일정구간이상의 유성음 구간이 존재하며 만약 잡음환경에서도 유성음 구간을 잘 검출할 수 있다면 이러한 음성의 특성을 이용하여 검출된 신호가 음성인지 아닌지를 판별할 수 있다. 이를 위하여 여기서는 가정잡음환경에서도 유성음을 잘 검출할 수 있도록 11 가지 유성음 특징들과 이를 이용한 음성판별방법을 제안하였다. 제안된 방법의 성능 평가를 위하여 음성의 끝점검출방법과 통합하여 음성/비음성 판별 테스트를 수행하였으며 테스트 수행결과 열악한 잡음환경에서 80%이상의 비음성을 거절하는 성능을 보였다.

1. 서론

우리가 생활하는 가정에는 많은 잡음원들이 존재하며 이러한 잡음원들로부터 발생되는 잡음들은 일반적은 비정적(non-stationary)인 특성을 가지고 있고 때로는 텔레비전에서 들려오는 중얼중얼거리는 소리라든지 라디오에서 들려오는 음악과 같이 음성과 유사한 특성을 가지기도 한다. 이러한 잡음환경에서 에너지 파라미터를 이용하는 음성의 끝점검출 방법[1]을 사용할 경우, 비음성 잡음을 사용자 음성으로 검출하는 현상을 흔히 발견할 수 있다. 이러한 약점을 보완하기 위하여 피치[10] 혹은 엔트로피[2] 기반의 끝점검출 방법이 제안되었으나 가정환경과 같은 복잡한 잡음환경에서 음성만을 검출하는 것은 여전히 어려운 문제로 남아 있다. 통계적 모델에 기반한 음성/비음성 판별 방법[11]은 각종 파라미터 기반의 음성검출 방법[1][2][10]에 비해 그 성능이 우수하나 음성과 유사한 특징을 갖는 잡음환경하에서 그 성능이 저하되는 단점을 가지고 있어 음성과 유사한 특성을 갖는 잡음이 빈번히 발생하는 가정환경에서는 그 성능을 발휘하지 못 하는 문제점이 있다.

본 논문에서는 음성의 유성음 특성을 이용한 음성/비음성 판별 방법을 제안한다. 즉, 음성의 유성음 특징을 세밀하게 표현하면서 잡음에 강인한 11 가지 유성음 특징을 이용하고 이를 기반으로 유성음을 판별해 낸다. 그런 다음 유성음 프레임 수의 비를 이용하여 검출된 신호가 음성인지 아닌지를 판별해 내게 된다.

서론에 이은 2 장에서는 제안된 전체 음성전처리 시스템에 대하여 간략히 설명하고 3 장에서는 제안된 음성/비음성 판별 방법에 대해 설명한다. 제안된 방법의 성능 평가 및 테스트 결과는 4 장에서 그리고 결론은 5 장에서 각각 설명한다.

2. 음성 전처리 시스템

제안된 음성전처리 시스템은 크게 두 단계로 구성되어 있다. 첫 번째 단계에서는 단일채널음질향상기법, TF 파라미터[1]와 EE feature[2]를 이용한 음성의 끝점검출법 그리고 유성음 특징을 이용한 음성/비음성 판별법이 유기적으로 결합되어 있으며 두 번째 단계에서는 GSAP 파라미터[3]를 이용하여 보다 세밀한 음성의 시작점과 끝점을 검출하게 된다. 즉, 첫 번째 단계에서 음질을 향상시킨 입력신호를 이용하여 음성 구간만을 판별한 후, 두 번째 음성전처리 단계에서 음성의 시작점과 끝점을 보다 세밀하게 검출하는 방식을 취하고 있다.

3. 음성/비음성 판별법

음성과 비음성을 판별하기 위해서는 먼저 입력신호에서 음성의 유성음 구간만을 정확히 검출해야 하는데 여기서는 음성의 유성음 구간 검출을 위해 다음과 같은 11 가지 유성음 특징을 추출하였다.

1. Modified TF 파라미터
2. High-to-Low Frequency Band Energy Ratio (HLFBER)
3. Tonality[4]
4. Cumulative Mean Normalized Difference Valley (CMNDV) based on YIN algorithm[5]
5. Zero Crossing Rate (ZCR)[6]
6. Level Crossing Rate (LCR)[6]
7. Peak-to-Valley Ratio (PVR) of autocorrelation function
8. Adaptive Band Partitioning Spectral Entropy (ABPSE)
9. Normalized Autocorrelation Peak (NAP)
10. Spectral Entropy[2]
11. Average Magnitude Difference Valley (AMDV)[8]

위와 같은 11 가지 유성음 특징들을 추출한 후, 현재 프레임이 유성음인지 아닌지를 판별하게 되는데 여기서는 유성음 임계값 혹은 경계값과 추출된 유성음 특징 파라미터들을 비교하는 방식으로 음성신호의 유성음 구간을 검출하였다. 음성의 유성음 구간의 경우 TF 파라미터와 NAP 는 잡음구간에 비하여 큰 값을 가지는 반면 CMNDV 파라미터는 작은 값을 가진다. 그리고 그외 특징 파라미터들은 일정구간의 값을 가지는 특성을 가지고 있다. 따라서 특징 파라미터들과 유성음 임계값 혹은 경계값을 비교함으로써 현재 프레임이 유성음인지 아닌지를 쉽게 판단할 수 있다. 첫 번째단계 끝점검출을 통해 검출된 음성구간과 유성음 구간의 비를 이용하여 검출된 신호가 음성인지 아닌지를 판별할 수 있다. 즉, 유성음 구간의 비가 임계값 보다 작으면 비음성으로 그 반대의 경우 음성으로 판단 할 수 있다. 그리고 일정구간 이상의 무성음 프레임이 연속적으로 발생하는 경우 음성의 끝점이 검출된 것으로 판단하므로 비정적 잡음에 의해 음성의 끝점검출 방법이 오동작하는 현상을 방지한다.

4. 성능평가 및 결과

제안된 음성/비음성 판별 방법의 성능을 측정하기 위하여 NON-PTT 모드에서 비음성 거절 성능을 측정하였다. 성능평가를 위해 49 명의 화자로부터 시뮬레이션 가정환경에서 음성을 녹취하였다. 다양한 가정환경잡음을 시뮬레이션 하기 위하여 20 여종의 가정환경 잡음을 녹취하고 화자의 음성발성 시 5 종 이상의 잡음을 마이크 주변에서 스피커를 통해 재생하면서 음성을 녹취하였다. 녹취된 신호는 16kHz, 16bits 로 디지털화되었고 발성된 발화는 가정자동화 영역의 50 개 명령어로 구성되어 있다. 각각의 발성 사이에 10 초간의 간격을 두어 발성 시뿐만 아니라 발성 사이 사이에도 다양한 가정잡음들의 간섭 현상을 데이터베이스에 반영하였다. 사용자가 원거리에서 발성하는 경우 나타나는 형상을 음성데이터에 반영하기 위하여 사용자로부터 각각 네 가지 거리(30cm, 50cm, 1m, 1.5m)에 마이크를 배치하여 사용자의 음성을 녹취하였다. 시뮬레이션 환경의 Signal-to-Noise Ratio (SNR)은 -5~10dB 사이의 값을 가지는 것으로 측정되었다. 아래 표 1 은 시뮬레이션 환경에서 비음성 거절 성능을 테스트한 결과이다.

<표 1> 시뮬레이션 환경에서 비음성 거절 성능

Mic. Distance	Correct Rejection	False Rejection
30 cm	85.2 %	0 %
50 cm	84.6 %	0.2 %
1 m	80.4 %	5.6 %
1.5 m	78.7 %	9.8 %
Total	82.2 %	3.9 %

위의 Correct Rejection 은 음성/비음성 판별 기능이 없는 2 단계 음성전처리 시스템에서 검출된 비음성 신호들의 개수를 기준으로 음성/비음성 판별 기능이 있는 경우 검출된 비음성 신호들이 얼마나 줄어 들었는가를 보여준다. False Rejection 은 음성/비음성 판별 기능 때문에 사용자 음성이 검출되지 못한 경우를 나타낸다.

5. 결론

본 논문에서는 다양한 가정잡음환경에서 음성의 유성음 구간만을 세밀하게 검출할 수 있는 11 가지 유성음 특징을 제안하고 이를 이용한 음성/비음성 판별 방법에 대해 설명하였다. 제안된 방법의 성능을 평가한 결과, 다양한 시뮬레이션 가정환경잡음 속에서 음성검출 성능의 열화가 거의 없이 80%이상의 비음성을 거절 할 수 있었고 향후, 정확한 성능평가 위하여 실제환경에 대한 테스트가 필요할 것으로 생각된다. 그리고 향후 제안된 방식의 유성음 특징벡터들이 통계적 모델기반의 신호 판별 방식 혹은 인공지능을 이용한 신호 판별 방식과 결합될 경우 성능향상이 기대된다.

참고문헌

[1] Jean-Claude Junqua, Brain Mak and Ben Reaves, “A Robust Algorithm for Word Boundary Detection in the Presence of Noise”, IEEE Trans. Speech and Audio Proc., VOL. 2, NO. 3, pp. 406~412, JULY 1994.

[2] Liang-Sheng Huang and Chung-Ho Yang, “A Novel Approach Robust Speech Endpoint Detection in Car Environments”, IEEE ICASSP2000, VOL. 3, pp 1751~1754, 2000.

[3] Nam Soo Kim and Joon-Hyuk Chang, “Spectral Enhancement Based on Global Soft Decision”, IEEE Signal Proc. Letters, VOL. 7, NO. 5, MAY 2000.

[4] James D. Johnston, “Transform Coding of Audio Signals Using Perceptual Noise Criteria”, IEEE Journal On Selected Areas In Communications, VOL. 6, NO. 2, FEBRUARY 1988.

[5] Alain de Cheveigne and Hideki Kawahara, “YIN, A Fundamental Frequency Estimator for Speech and Music”, Journal of the Acoustical Society of America, 111(4), 2002.

[6] Lawrence R. Rabiner, “On the Use of Autocorrelation Analysis for Pitch Detection”, IEEE Trans. On Acoustics, Speech, And Signal Proc., VOL. ASSP-25, NO. 1, FEBRUARY 1977.

[7] Bing-Fei Wu and Kun-Ching Wang, “Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments”, IEEE Trans. On Speech and Audio Processing, VOL. 13, NO. 5. SEPTEMBER 2005.

[8] Myron J. Ross, Harry L. Shaffer, Andrew Cohen, Richard Freudberg, and Harold J. Manley, “Average Magnitude Difference Function Pitch Extractor”, IEEE Trans. On Acoustics, Speech And Signal Proc., VOL. ASSP-22, NO. 5, OCTOBER 1974.

[9] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.2 (2000-04), April 2000.

[10] K. Iwano and K. Hirose, “Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and its Use for Continuous Speech Recognition”, IEEE ICASSP’99, VOL. 1, pp. 133-136, MAY 1999.