EVRC 패킷에서 LSP 거리를 이용한 음성 끝점 검출

An End Point Detection Technique Using the LSP Distance in EVRC Packets

민 병 준*, 강 명 수*
(Byung Jun Min*, Myoung Soo Kang*)

요 약

음성 인식 기능의 성능 향상을 위해서는 처리 속도가 빠르면서도 잡음 환경에서 정확하게 동작하는 음성 끝점 검출이 선행되어야 한다. 본 논문에서는 저잡음 환경에서의 음성 끝점 검출을 위한 간단하면서도 빠른 알고리즘을 제안한다. 제안된 알고리즘은 LSP 거리를 측정 기준으로 문턱값 논리를 사용하며 입력 음성으로는 EVRC로 보코딩된 패킷을 이용한다. 제안된 알고리즘을 이용한 실험 결과는 디코딩된 음성 파형으로부터 결정한 결과와 비교되었다. 실험 결과에서 제안된 알고리즘은 만족할만한 정확성을 나타내었다.

ABSTRACT

This paper presents a simple and fast method for end point detection under low-level noisy environment. The proposed algorithm uses a threshold logic with LSP distances and takes vocoded packets as input to the recognition system. The results from the proposed method are compared with those manually checked in decoded speeches. From the result it exhibits acceptable accuracy.

I. 서 론

음성 인식 시스템에서 부정확한 음성 끝점검출은 음소를 잘라내거나 잡음을 음성 구간에 포함시킴으로써 인식기의 인식률을 떨어뜨리는 결과를 초래한다. 따라서, 정확한 음성 끝점 검출을 위한 대표적인 몇가지 알고리즘들이 제안되어 왔다[1-5]. 여기서, 이러한 대부분의 알고리즘들은 에너지 contour, 영 교차율 혹은 레벨 교차율, LPC등의 파라미터들에 기초를 두고 있다. 하지만, 이들 알고리즘들은 잡음 환경에서 음성 인식 시스템의 전반적인 정확성을 현저하게 떨어뜨리는 부정확성을 지니고 있다[1, 2]. 반면에 정확성이 높은 알고리즘들은 음향학적인 파라미터들에 기초하고 있어 그 성능이 우수한 반면, 시스템의 응답 시간이 느린 단점을 지니고 있다[5].

본 고에서는 저잡음 환경 하에서 발음된 음성 구간을 검출하기 위하여 간단하고 빠르면서도 정확한 알고리즘을 제안한다. 제안된 알고리즘은 프레임 단위로 입력되는 EVRC 보코딩된 패킷에서 LSP 파라미터를 디코딩하여 이로부터 구한 거리 정보를 측정 기준으로 문턱값 논리를 사용한다. 이는 음성 끝점 검출에 필요한 최소의 파라미터만 추출함으로써 처리 시간을 줄일 수 있게 하고, 사용하는 모든 파라미터들에 고정 소수점 연산을 수행함으로써 보다 빠른 동작을 가능하게 하였다. 또한, 입력 음성으로 EVRC로 보코딩한 패킷을 사용하여 잡음의 영향을 최소화한 상태에서 알고리즘을 적용하므로 실제 환경에서도 인식기의 우수한 성능을 보장할 수 있다.

제안된 방법은 운전 중인 자동차 안, 사무실 등의 잡음 환경에서 획득한 EVRC 패킷을 이용하여 테스트 되었고, 알고리즘으로 결정된 음성 구간과 수작업으로 결정된 음성 구간을 비교하여 정확성을 검토하였다.

본 논문에서 서술할 내용은 다음과 같다. II 절에서는 음성 끝점 검출을 위한 기본 가정에 대해 기술하고, III 절에서는 제안된 알고리즘을 설명한다. 마지막으로 IV 절에서는 향후 개선해야 할 문제점 및 결론을 제시한다.

II. 음성 끝점 검출을 위한 기본 가정

일반적으로 음성 끝점 검출을 위해서는 입력된 음성 파형을 그대로 이용한다. 하지만, 본 논문에서 제안된 알고리즘은 EVRC 단말기에서의 음성 인식기 구현에 목적이 있으므로, 이를 위해서는 단말기에서 음성이 처리되는 과정을 살펴볼 필요가 있다.

그림 1에서와 같은 단말기의 음성 처리 과정에서 대부분의 음성인식 알고리즘들이 PCM 데이터를 사용하는 반면 제안된 알고리즘은 EVRC로 보코딩된 패킷을 사용한다. 이로써 얻어지는 장점은 여러가지가 있겠지만 서론에서도 언급하였듯이 EVRC 보코더의 잡음 제거 기술을 그대로 사용할 수 있고, 음성 인식 과정에서 패턴 매칭에

* SK텔레콤 중앙연구원
접수일자 : 1999년 3월 5일

사용되는 LSP 파라미터를 패킷에서 디코딩해서 사용하므로 음성에서 따로 특징 파라미터를 추출하는 추가적인 과정이 필요없게 된다.

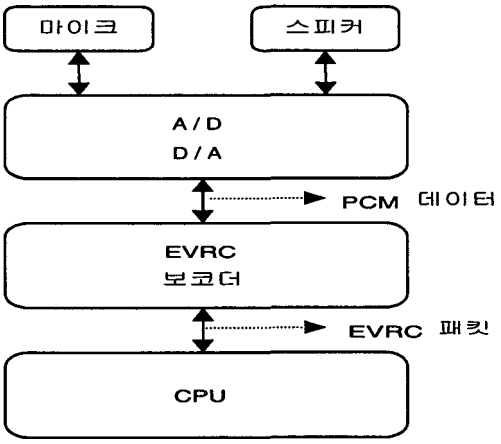


그림 1. 단말기에서의 음성 처리 과정
Fig. 1. Voice processing procedure in a EVRC phone.

지금까지 언급한 내용이 그림 2에 나타나있다. 동작 속도의 관점에서 보면 CPU(Arm 7 사용)에서 LSP 파라미터를 디코딩하는 시간은 각 단계에서의 처리 시간을 비교한 표 1에서 알 수 있듯이 전체 디코딩 시간에 결정적인 영향을 미치지 않는다. 또한, LSP 파라미터들에 고정 소수점 연산을 사용함으로써 처리 속도를 최소화하게 된다.

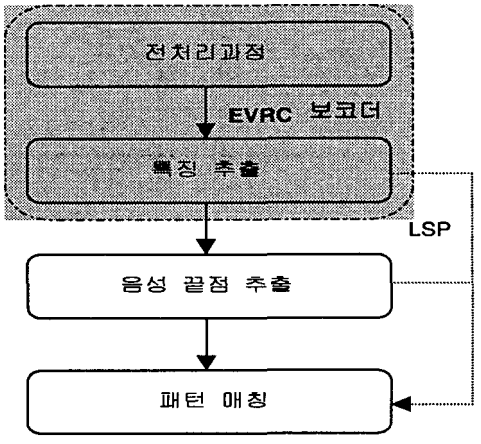


그림 2. 제안된 음성 인식 과정
Fig. 2. Proposed speech recognition procedure.

마지막으로, 제안된 알고리즘은 가변 데이터 율로 보코딩하지 않고, rate 1으로만 보코딩함으로써 디코딩 과정에서 복잡도를 최소화하였다.

표 1. 디코딩 단계에서의 평균 처리 시간
Table 1. Average processing time in decoding procedure.

디코딩 단계	평균 처리 시간 (고정 소수점)	전체 시간 대 비율
LSP 디코딩	0.008 s	3.7 %
Adaptive Codebook Gain & Delay 디코딩 Fixed Codebook Gain & Index 디코딩 Adaptive Codebook Memory 갱신	0.007 s	3.3 %
LSP 를 LSP 로 변환	0.1 s	46.5 %
LSP 필터링	0.1 s	46.5 %

III. LSP 거리를 이용한 음성 끝점 검출

본 장에서는 음성 구간에서의 LSP 변화와 그를 이용한 음성 끝점 검출에 대하여 자세히 설명하도록 한다.

3.1 LSP 거리 계산

제안된 방법은 비음성 구간과 음성 구간에서 LSP 벡터들의 분포가 다른 점을 이용한다. 조용한 환경이나 일정한 잡음이 존재하는 환경에서 비음성 구간의 LSP 벡터는 그림 3과 그림 4에서처럼 그 값들이 변하지만 비교적 작은 범위 내에 있음을 알 수 있다. 반면, 동일한 환경에서 음성이 존재하는 경우는 LSP 벡터가 포먼트를 형성하기 위한 방향으로 이동하므로 정상적인 범위에서 벗어나 분포하고 있음을 알 수 있다.

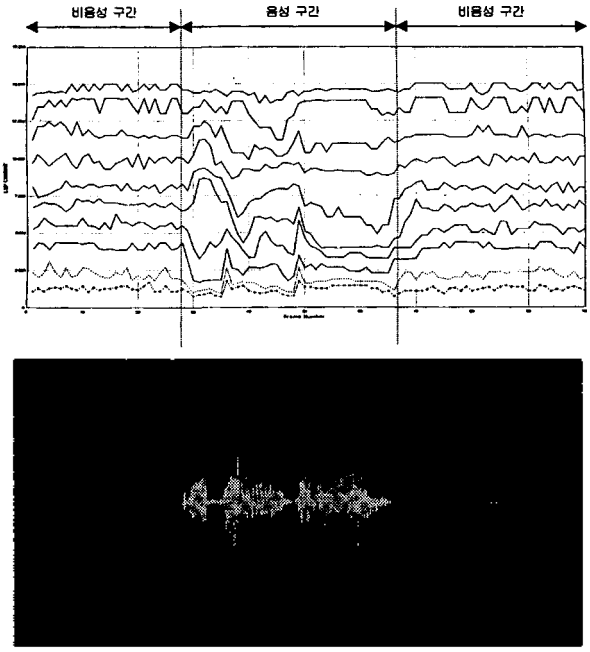


그림 3. 조용한 환경에서 음성 및 비음성 구간의 LSP 벡터 변화와 실제 음성 파형
Fig. 3. LSP vector contour in clean environment and speech waveform.

따라서, 잡음의 레벨 및 특성이 일정한 환경에서 LSP 벡터 정보만으로도 충분히 비음성, 음성 구간을 구별할 것으로 예상할 수 있다. 하지만, 그림 5와 같이 전화벨 소리나 babble 잡음(주위의 왁자지껄하는 소리)처럼 잡음 특성이 음성과 비슷하거나 그 레벨이 일정하지 않는 경우에는 정상적인 동작을 보장하지 못할 수도 있을 것이다.

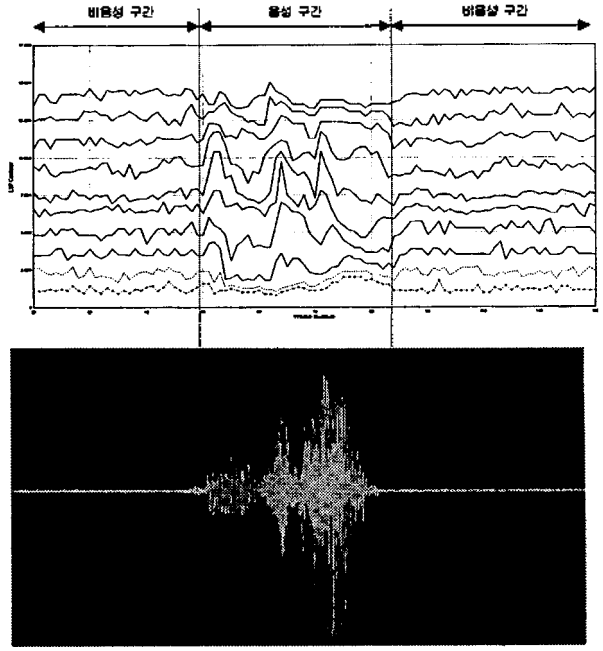


그림 4. 일정한 잡음 환경에서 음성 및 비음성 구간의 LSP 벡터 변화와 실제 음성 파형

Fig. 4. LSP vector contour in constant noise environment and speech waveform.

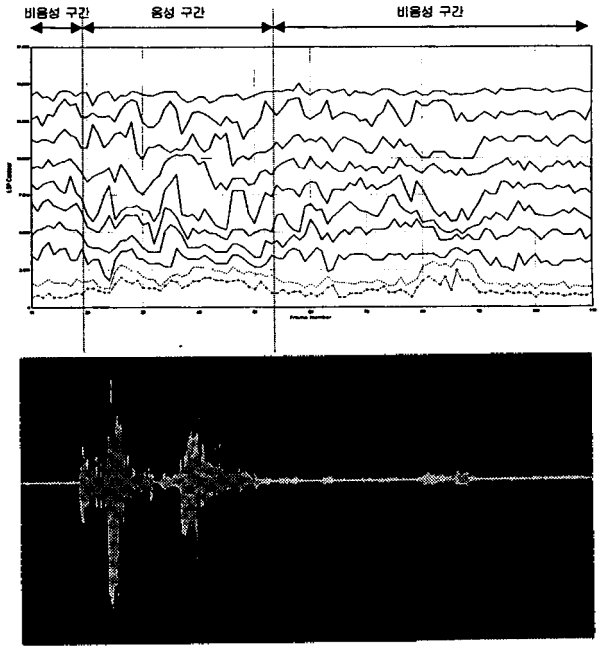


그림 5. babble 잡음 환경에서 음성 및 비음성 구간의 LSP 벡터 변화와 실제 음성 파형

Fig. 5. LSP vector contour in babble noise environment and speech waveform.

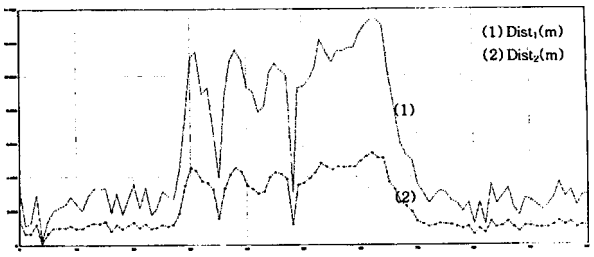


그림 6. 조용한 환경에서 음성 존재 시 LSP 거리

Fig. 6. LSP distance in clean environment.

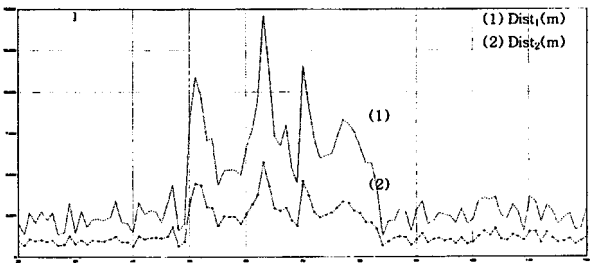


그림 7. 일정한 잡음 환경에서 음성 존재 시 LSP 거리

Fig. 7. LSP distance in constant noise environment.

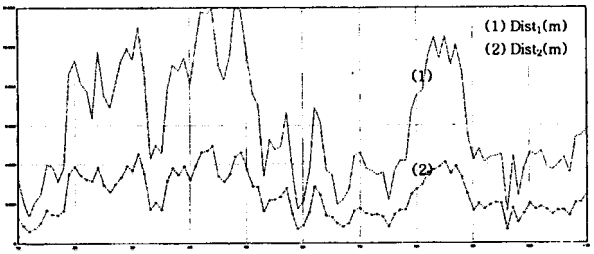


그림 8. babble 잡음 환경에서 음성 존재 시 LSP 거리

Fig. 8. LSP distance in babble noise environment.

실제로 음성 끝점 검출에는 위의 정보를 바탕으로 한 LSP 거리를 이용한다. 즉, 패킷을 입력 받기 시작해서 5~7프레임(100~140ms) 정도를 음성이 없고 주변 잡음만 존재한다는 가정하에 훈련 구간(TRAINING)으로 정하여 이 구간에서 LSP 벡터들을 구한 후 각각의 LSP에 대한 중간값을 취한다. 여기서 중간값을 취하는 이유는 훈련 구간에서 순간적인 잡음에 따른 LSP 변화의 영향을 막기 위함이다. 따라서, 훈련 구간에서는 중간값들로 이루어진 기준 LSP 벡터를 얻게 된다.

기준 LSP 벡터를 구한 후에는 입력된 첫번째 프레임부터 매 프레임마다 구해진 LSP 벡터와 기준 LSP 벡터 사이의 Euclidian 거리를 계산한다. 이를 식으로 나타내면 식 (1)과 같다.

$$Dist_1(m) = \sqrt{\sum_{i=1}^{10}(LSP_m(i) - LSP_{median}(i))^2}$$

$LSP_m(i)$: m번째프레임의 i 번째 LSP

$LSP_{median}(i)$: 기준프레임의 i 번째 LSP (1)

하지만, 곱하는 과정과 제곱근을 취하는 과정은 속도를 저하시키는 원인이 되므로 식 (2)와 같이 간략화된 거리를 계산한다.

$$Dist_2(m) = abs(LSP_m(i) - LSP_{median}(i))$$

$LSP_m(i)$: m번째프레임의 i 번째 LSP

$LSP_{median}(i)$: 기준프레임의 i 번째 LSP (2)

그림 6, 7, 8에서는 그림 3, 4, 5에 각각 해당하는 LSP 거리를 보여주고 있다. 각각의 그림에서 (1)은 식 (1)에 해당하고, (2)는 식 (2)에 해당하는 결과이다. 그림들에서 알 수 있듯이 LSP 거리가 비음성 구간에서는 아주 작고, 음성 구간에서는 반대로 어느 일정한 값을 넘으므로 이로써 음성 구간의 유무를 판단할 수 있게 된다.

3.2 문턱값 (Threshold) 논리

본 절에서는 II 장에서 언급한 LSP 거리에 대한 정보를 바탕으로 문턱값 논리를 이용하여 음성 구간의 시작점 및 끝점을 검출하는 방법에 대하여 설명하기로 한다.

가. 시작점 검출

우선, 훈련 구간에서 결정된 기준 LSP 벡터와 훈련 구간 동안의 LSP 벡터들과의 거리들을 구한다. 사실, 이 거리들은 작은 값들이지만 음성 구간임을 판단하기 위한 문턱값을 설정하기 위하여 구하는 것이다. 즉, 이 거리들 중 최대값을 문턱값으로 설정하여 훈련 구간 이후의 프레임에 대한 LSP 거리가 이 문턱값보다 어느 이상 크게 되면 음성 구간으로 판단한다. 문턱값도 최대값 이외의 다른 조건으로 결정한 값을 사용할 수 있다.

위와 같이 문턱값이 설정되고 난 후 LSP 거리가 이 값을 넘어서는 순간이 바로 음성 구간의 시작점이 될 수 있다. 하지만, 어느 정도의 잡음 역시 문턱값을 넘을 수 있으므로 이 경우를 배제하기 위하여 LSP 거리가 문턱값을 넘고, 그 상태가 미리 정한 프레임 수 이상 유지되면 시작점으로 간주하도록 한다.

나. 끝점 검출

끝점도 시작점과 유사한 방법으로 구하게 된다. 끝점은 시작점과는 반대로 LSP 거리가 문턱값 아래로 떨어지는 순간이 될 수 있다. 하지만 이 경우 또한 처음 음절이 끝나고 다음 음절이 시작되기 전에 끝점이 검출된 것으로 오판할 수 있으므로 시작점 검출과 유사하게 문턱값 아래로 떨어지고 난 후 그 상태가 미리 정한 프레임 수 (NO_FRAME_END)이상 유지되면 끝점으로 간주하도록 한다. 즉, 처음 끝점이 검출되고 NO_FRAME_END 프레임이 지나지 않아서 또 다른 시작점이 검출되면 이를 연결된 한 단어로 간주하여 검출된 끝점을 취소하고 다시 끝점 검출 조건을 검색하게 된다. 여기서, 또 한가지 고려할 것은 NO_FRAME_END 프레임 동안 순간적인 잡음에 의해 짧은 구간동안 문턱값을 넘는 경우가 끝점 검출에 영향을 미치지 못하도록 그 구간의 길이는 문턱값 아래 값처럼 간주하여야 한다는 것이다. 마지막으로 끝점과 시작점의 차이가 미리 정한 최소 프레임 이하이거나 최대 프레임 이상인 경우에는 정상적인 음성이 아닌 것으로 간주하고 다시 음성을 입력 받도록 한다.

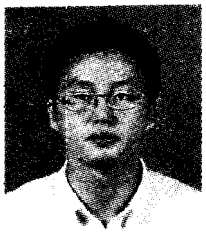
IV. 결 론

본 논문에서는 단말기에서 구현 가능한 EVRC 패킷을 입력으로 하는 음성 인식기에서 음성 끝점을 검출하는 방법을 제시하였다. 제안된 방법으로 남녀 각각 2명씩의 화자가 100단어를 2회씩 시속 100km정도로 운전 중인 차 안에서 라디오를 틀어놓고 (평균 20dB SNR) 실험하였다. 여기서, 끝점은 프레임 단위로 계산되므로 이를 비교할 때, 전체 디코딩된 음성 파형에서 판단한 끝점 샘플이 해당 끝점 프레임 내에 있으면 일치하는 것으로 간주한 결과 95% 정도가 음성 끝점 검출에 성공하였다. 따라서, 잡음의 정도가 일정한 환경에서는 처리 속도가 빠르면서도 정확하게 음성 끝점을 검출해 내어 만족할 만한 성능을 보임을 알 수 있었다. 하지만, 제안된 방법은 잡음의 레벨이 일정치 않거나 babble 잡음 같은 환경 하에서는 정상적인 동작을 보장할 수 없는 한계가 있었다. 이를 해결하기 위해서는 에너지 정보를 이용하는 방법을 고려할 수 있겠는데, 이 역시 디코딩 시간을 줄이기 위해 LPC를 구하지 않고, 적응 코드북의 에너지와 그 이득만으로 주변 잡음의 영향을 감소시키면서 음성 구간의 에너지 정보를 구하는 방법을 고려할 수 있다.

참 고 문 헌

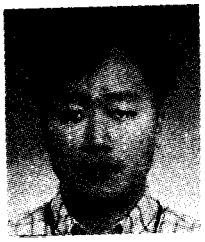
1. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., Vol. 54, pp. 297-315, FEB. 1975.
2. L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", IEEE ASSP, Vol. 29(4), pp. 777-785, Aug. 1981.
3. D. Childers, M. Hahn, J. Larar, "Silent and Voice/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech", IEEE ASSP, Vol. 37(11), pp. 1771-1774, Nov. 1989.
4. M. H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals", Speech Communication 8 (1989), pp. 45-60.
5. L. R. Rabiner, C. E. Schmidt, B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech", Bell System Technical Journal, pp. 455-487. Mar. 1977.

▲민 병 준 (Byung Jun Min) 1962년 2월 20일생



1981년 ~ 1985 : 서울대학교 공과대학 전자공학과(공학사)
1985년 ~ 1987년 : 한국과학기술원 전기 및 전자공학과(공학석사)
1987년 ~ 1993년 : 한국과학기술원 전기 및 전자공학과(공학박사)
1987년 ~ 1997년 : 디지콤 정보통신연구소 책임연구원
1997년 ~ 현재 : SK텔레콤 중앙연구원 수석연구원
※주관심분야 : 음성 데이터 압축, 음성 인식, 음성 합성

▲강 명 수(Myoung Soo Kang) 1971년 7월 24일생



1990년 ~ 1994년 : 한양대학교 전기공학과(공학사)
1996년 ~ 1998년 : 포항공과대학교 대학원 전자전기공학과(공학석사)
1998년 ~ 현재 : SK Telecom 연구원
※주관심분야 : 음성 데이터 압축, 음성 인식