스펙트럼의 변동계수를 이용한 잡음에 강인한 음성 구간 검출

김영민(ICU), 한민수(ICU)

<차 례>

<Abstract>

Noise-Robust Speech Detection Using The Coefficient of Variation of Spectrum

Youngmin Kim, Minsoo Hahn

This paper deals with a new parameter for voice detection which is used for many areas of speech engineering such as speech synthesis, speech recognition and speech coding. CV (Coefficient of Variation) of speech spectrum as well as other feature parameters is used for the detection of speech. CV is calculated only in the specific range of speech spectrum.

Average magnitude and spectral magnitude are also employed to improve the performance of detector. From the experimental results the proposed voice detector outperformed the conventional energy-based detector in the sense of error measurements.

* Keywords: Endpoint detection, Word boundary detection, Voice activity detection

1. 서 론

오늘날 음성분석 및 합성, 음성인식, 음성코딩, 음성부호화 등 음성 신호처리에 관련된 전반적 분야에서 음성 신호의 음성 구간을 정확하게 검출하는 일은 매우 중요하다. 특히 음성 인식 시스템에서 음성의 시작점과 끝점을 검출하는 것은 인식기의 전반적 성능에 미치는 영향이 매우 크므로 선결해야 하는 과제이기도 하다. 또한 디지털 이동통신시스템의 음성 신호를 부호화하는 과정에서 정확한 음성 검출을 이용해 음성 구간만을 부호화함으로써 시스템 용량 증가와 더불어 전송 효율 또한 높일 수 있다. 음성 신호에서 음성 구간의 시작점과 끝점은 발성시 여러 가지 요인에 의해 검출에 어려움이 있고 여러 잡음이 부가되었을 때는 더욱 어려워진다. 따라서 잡음이 있는 음성 신호에서 정확한 음성 구간을 검출하는 것은 쉽지 않은 일이다. 음성 신호의 잡음에는 부가 잡음과 콘볼루션 잡음 등이 있으며 일반적으로 신호대 잡음비가 충분히 큰 경우(일반적으로 30dB)에는 에너지 함수와 영교차율을 이용해서 손쉽게 음성 구간을 검출할 수 있으나[1][2] 신호대 잡음비가 10dB 이하에서는 에너지를 이용한 음성 검출은 많은 어려움이 있다고 알려져 있다.

음성 신호처리의 필수 요건으로써 음성 검출에 관한 많은 연구가 이루어졌으며 에너지의 임계값을 이용하는 방법[1, 2], 캡스트럼(Cepstrum)을 이용한 방법, LPC를 이용한 방법[4], 피치를 이용한 방법, 시간-주파수 특징 벡터를 이용한 방법[3, 5], 주기성 측정 (Periodicity measure)을 이용한 방법 등이 연구되었다.

본 논문에서는 주파수 영역에서 일반적인 음성 구간에 대한 스펙트럼의 변동계수를 특징 벡터로 이용하여 보다 잡음에 강인한 음성 검출이 가능함을 기술한다. 변동계수는 성도의 주기성으로부터 발생하는 피치의 유무와 연관성이 있으며 이 특징 벡터의 장점은 다음과 같다. 첫째, 잡음 성분이 부가되었을 경우 일반적인 피치를 구하기에 힘든 점이 많지만 변동계수로써 피치 존재의 유무만을 판단함으로써 보다 신뢰성을 높일 수 있다. 둘째, 특정 구간에 대해 일반화(normalization)되어 있는 특징 벡터로써 주파수 크기에 영향을 받지 않는다. 단, 피치가 존재하지 않는 영역의 검출에 어려움이 있는 단점이 있으나 다른 특징 벡터값들을 이용함으로써 이러한 단점을 보완하였다.

변동계수만으로 음성 구간을 검출하기보다는 시간 영역에서의 에너지와 주파수 영역에서의 에너지를 특징 벡터로 같이 이용할 경우 계산량은 늘어나지만 좀 더 나은 성능을 보여준다. 특히 자동차 소음과 같은 낮은 주파수 성분이 많은 잡음이 부가되었을 경우 음성의 주파수 영역과 중복되는 경우가 발생하기 때문에 역시 다른 특징 벡터들의 보완을 통해 해결하여 성능을 개선하였다.

2. 음성 검출 (Voice Detection)

2.1. 특징 벡터

1) Average Magnitude

시간 영역에서 에너지대신 사용한 특징 벡터이며 에너지와 비교하면 상대적으로 음성 신호가 약한 부분에서 음성 검출이 용이하다는 장점을 가지고 있다.

$$Average\ Magnitude = \frac{1}{N}\sum_{n=1}^{N} |x(n)| \tag{1}$$

(N: 한 프레임의 데이터 수)

2) Spectral Magnitude

주파수 영역에서 구한 값으로써 Average Magnitude와 구하는 방식은 같으나 본 논문에서는 실험적으로 구한 특정구간 내에서의 값의 평균으로 정의된다.

$$Spectral\ Magnitude = \frac{1}{N}\sum_{k=1}^{N} |X(k)| \tag{2}$$

(X(k) : 주파수 영역의 음성 신호)

3) 변동계수 (Coefficient of Variation, CV)

주파수 영역에서 구해지는 특징 벡터로써 실험적으로 구해진 구간 내에서 스펙트럼의 일반화된 분산 정도를 나타낸다. 이는 스펙트럼 내에 음성이 존재하는 구간에서는 피치에 의한 하모닉스 성분으로 말미암아 큰 값의 변화가 생기게 되는데 평균값으로 일반화시킨 후 이용하게 된다. 즉 피치 정보의 유무를 쉽게 판단할 수 있으며 본 논문에서는 주된 특징 벡터로 이용하였다.

$$CV = \frac{\sqrt{\frac{1}{b-a+1}\sum_{k=a}^{b}(x(k)-mean)^2}}{mean} \tag{3}$$

(a, b: 주파수 영역에서의 특정 구간의 임계값(a=125Hz, b=1000Hz),
mean: 주파수 영역에서 특정 주파수 구간의 크기의 평균값)

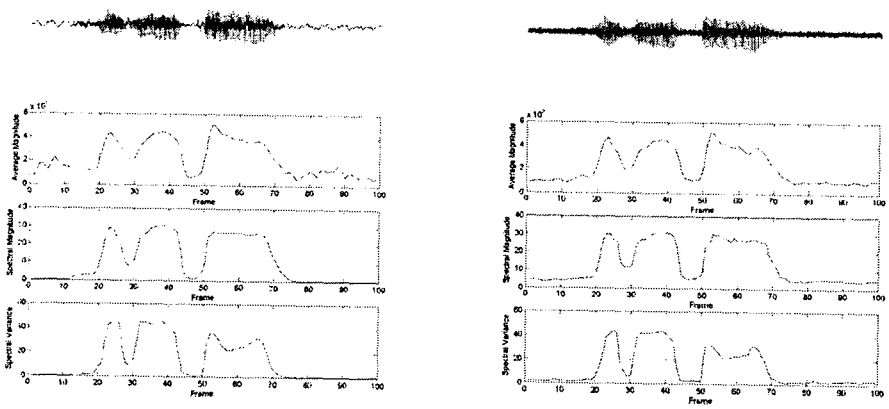
4) LCR (Level Crossing Rate)

잡음 환경 하에서 에너지를 기반으로 하는 음성 검출에 있어 무성음의 검출에 어려움이 따르므로 영교차율(Zero Crossing Rate)을 이용하는데 본 논문에서는 영교

차율보다 성능이 좋다고 알려져 있는 레벨 교차율(LCR, Level Crossing Rate)을 이용한 방법과 성능을 비교하였다.

2.2. 특징 벡터의 비교

<그림 1>은 '청와대'라는 음성을 본 논문에서 이용한 3가지 특징 벡터들의 변화를 나타낸 그림이다. 각각 10dB의 차량 잡음과 백색 잡음이 부가된 음성 신호이며 잡음 구간에서 Average magnitude 값과 Spectral magnitude 값은 음성 구간에 비해 상대적으로 변동량이 크며 전체적인 값도 크다. 이로써 CV 값이 잡음의 종류에 관계없이 음성 구간을 검출하는데 있어 좋은 성능을 보이며 적절한 특징 벡터임을 알 수 있다. 음성 신호의 시작부분이 마찰음(fricative)으로 이루어져 있으며 CV 값으로만 검출하기에 어려움이 따르므로 CV 값 외에 Spectral Magnitude 값과 Average Magnitude 값을 부가적으로 이용하였다.



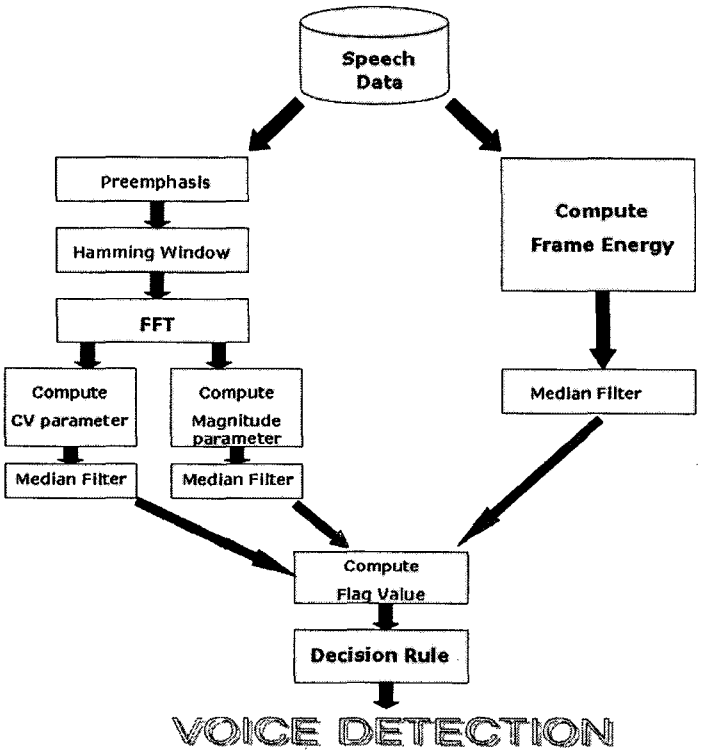
(a) 차량 노이즈가 부가된 음성 신호의 경우 (b) 백색 노이즈가 부가된 음성 신호의 경우
<그림 1> Average Magnitude, Spectral Magnitude와 CV의 잡음에 따른 비교

2.3. 제안 알고리즘

음성 신호의 입력으로부터 각각 시간 영역과 주파수 영역으로 특징 벡터들이 계산되어진다. 먼저 시간 영역으로는 프레임단위로 average magnitude 값이 계산되어 지며 median filter를 통해 smoothing되어 flag 값 계산에 포함된다. 주파수 영역에서는 preemphasis와 windowing을 통과한 후 FFT 계산을 통해 스펙트럼을 얻게 되고 이 스펙트럼으로부터 CV 값과 magnitude 값이 얻어지게 된다. 이 값들 역시

smoothing을 위해 median filter을 통과하게 되고 음성 구간을 검출하기 위한 flag 값에 기여하게 된다.

마지막으로 3가지 특징 벡터들로부터 계산된 flag 값으로부터 decision rule을 통해 음성 검출이 이루어지게 된다. 제안 알고리즘의 블록도가 <그림 2>에 보인다.



<그림 2> 제안 음성 검출 알고리즘

3. 실험 및 결과

3.1. 실험 데이터

실험에 사용된 DB는 원광대에서 제작한 PBW (Phonetically Balanced Words) 452 DB 중 남여 각각 5명씩 총 10명의 화자에 대해 1번씩 발성한 4,520개의 데이터를 이용하였다. 노이즈는 백색 잡음, 자동차 잡음, 혼합 잡음의 3가지를 이용하였고, 신호대 잡음비(SNR)를 5dB 간격으로 7가지로 설정하였다. <표 1>은 DB에

관한 정보를 나타낸 표이다.

<표 1> 사용된 DB 정보

사용 DB	원광대 PBW 452
Sampling Rate	8kHz (down sampling)
resolution	16bit
Frame size	160 samples (20msec)
Frame shift	80 samples (10msec)
화자	10명 (남자 5명 여자 5명)
잡음 종류	white, car, mixed
잡음 레벨	신호대 잡음비(SNR) 0, 5, 10, 15, 20, 25, 30dB
총 데이터수	4,520

3.2. 실험

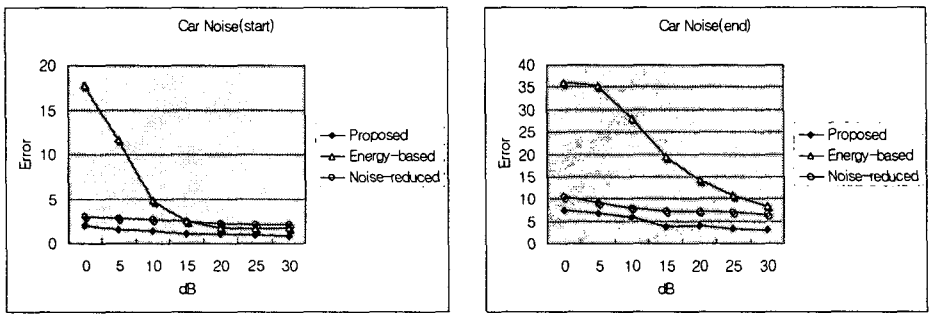
결과를 비교 평가하기 위하여 에너지와 영교차율 이용한 음성 검출과 칼만 필터를 이용하여 노이즈를 감소시킨 DB를 같은 방법으로 실험하였다. 임계값을 설정하기 위해 음성 신호의 첫 80㎳구간을 묵음으로 간주하였다. 성능 평가는 4,520개의 데이터를 웨이브 파형과 스펙트로그램으로부터 시각적으로 검출한 값을 기준으로 시작점과 끝점에 있어 벗어난 정도의 비교를 통해 이루어졌다.

주파수 영역에서 특징 벡터 계산은 일반적 음성의 분포 구간인 125Hz~1000Hz 내에서 이루어졌다. 에너지와 영교차율을 이용한 음성 검출은 잡음 부가시 상당히 에러 발생율이 높은 관계로 영교차율 대신 레벨 교차율(LCR)을 이용한 방법과 음성 데이터를 칼만 필터(Kalman Filter)를 이용하여 노이즈를 감소시킨 후 같은 방법으로 나온 결과 값을 비교하였다.

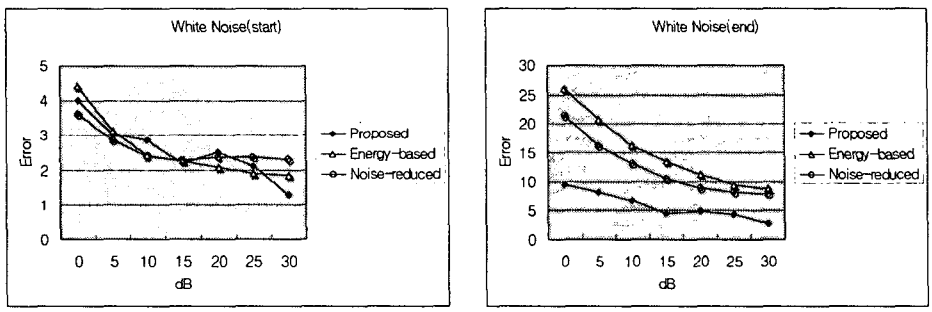
3.3. 결과

전반적으로 제안된 특징 벡터와 알고리즘을 사용한 경우가 에너지를 기반으로 하는 음성 구간 검출과 비교했을 때 보다 나은 성능을 보여주었으며 자동차 노이즈같이 에너지가 큰 노이즈의 경우 월등한 성능을 보여주었다. 칼만 필터(Kalman Filter)를 통한 잡음 제거 후 에너지로의 음성 구간 검출과 비교했을 때 시작점은 대동소이하나 끝점 검출에 많은 성능 차이를 보여주었다. 이는 끝점 근처에서 음성 신호가 서서히 감소하여 잡음 제거시 음성의 일부분까지 같이 제거됨으로 인한 결과로 추정된다. <그림 3>은 세 가지 노이즈의 부가에 따른 실험 결과 그래프

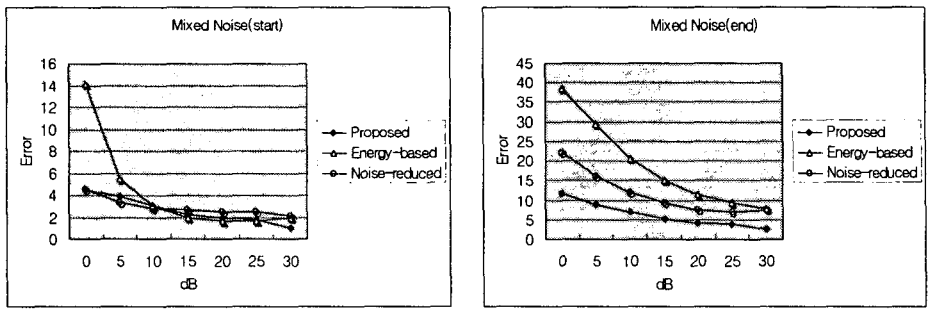
이다. <표 3>은 3가지의 노이즈에 대한 각 알고리즘별로 프레임 에러의 분포를 나타낸 표이다. 전반적으로 제안된 알고리즘이 에러가 작음을 알 수 있다.



(a)



(b)



(c)

<그림3> 검출된 음성 구간의 시작과 끝점의 프레임 에러 크기
(a): 차량 잡음, (b): 백색 잡음 (c): 혼합 잡음 (차량 잡음+백색 잡음)

<표 2> 음성의 시작 부분에 대한 노이즈별 에러 분포

		Error	Start				End			
			0dB	10dB	20dB	30dB	0dB	10dB	20dB	30dB
Proposed	car	$\|E\| \le 3$	3700	4099	4285	4367	1853	2671	4285	3413
		$3 < \|E\| \le 6$	522	344	186	120	912	592	186	642
		$6 < \|E\| \le 9$	182	42	32	20	396	244	32	176
		$9 < \|E\|$	117	36	18	14	1360	1014	18	290
	white	$\|E\| \le 3$	2544	3193	3388	4031	623	900	3388	3147
		$3 < \|E\| \le 6$	1088	763	641	330	813	1692	641	1169
		$6 < \|E\| \le 9$	541	348	275	91	976	1212	275	184
		$9 < \|E\|$	340	217	216	69	2109	716	216	21
	mixed	$\|E\| \le 3$	2171	3136	3717	4180	409	817	3717	3434
		$3 < \|E\| \le 6$	965	798	487	246	612	1336	487	948
		$6 < \|E\| \le 9$	549	343	192	58	825	1345	192	122
		$9 < \|E\|$	837	244	125	37	2675	1023	125	17
Energy-based	car	$\|E\| \le 3$	424	2797	4214	4355	3	50	4214	1898
		$3 < \|E\| \le 6$	269	755	200	133	16	232	200	1945
		$6 < \|E\| \le 9$	204	326	41	13	78	462	41	185
		$9 < \|E\|$	3624	643	66	20	4424	3777	66	493
	white	$\|E\| \le 3$	2742	3850	4283	4186	80	231	4283	2653
		$3 < \|E\| \le 6$	937	458	192	280	108	1165	192	246
		$6 < \|E\| \le 9$	365	100	13	27	275	911	13	113
		$9 < \|E\|$	477	113	33	28	4058	2214	33	1509
	mixed	$\|E\| \le 3$	1061	3340	4280	4336	124	314	4280	2752
		$3 < \|E\| \le 6$	708	677	157	152	113	607	157	314
		$6 < \|E\| \le 9$	445	212	35	11	164	932	35	106
		$9 < \|E\|$	2307	292	49	22	4120	2668	49	749
Noise-reduced	car	$\|E\| \le 3$	3652	3942	4156	4271	800	1630	4156	2151
		$3 < \|E\| \le 6$	657	356	266	216	1344	1736	266	328
		$6 < \|E\| \le 9$	99	149	60	14	1046	316	60	64
		$9 < \|E\|$	113	74	39	20	1331	839	39	1978
	white	$\|E\| \le 3$	2689	3256	3546	4215	429	604	3546	2625
		$3 < \|E\| \le 6$	756	526	641	263	338	1379	641	363
		$6 < \|E\| \le 9$	742	640	305	7	515	905	305	129
		$9 < \|E\|$	369	99	29	36	3239	1633	29	1404
	mixed	$\|E\| \le 3$	2171	3136	3717	4180	373	849	3717	2700
		$3 < \|E\| \le 6$	965	798	487	246	556	1970	487	289
		$6 < \|E\| \le 9$	548	343	192	58	855	744	192	86
		$9 < \|E\|$	837	244	125	37	2737	958	125	1446

4. 결 론

본 논문에서는 주파수 영역에서 CV라는 새로운 특징 벡터뿐만 아니라 spectral magnitude와 시간 영역에서의 average magnitude 값을 부가적으로 이용함으로써 가장 일반적 알고리즘인 에너지를 이용한 방법과 칼만 필터를 이용하여 노이즈를 제거한 후, 다시 에너지를 이용하여 끝점 검출을 한 방법과 비교하였다. 에너지를 이용한 방법은 잡음의 유무에 따라 성능의 큰 차이를 보여주었으며 잡음이 부가된 음성 신호에서 좋은 성능을 보이기 힘들며 제안된 특징 벡터와 알고리즘이 음성 검출에 있어 나은 결과를 보여주었다. 향후 알고리즘의 개선을 통해 고립 단어뿐만 아니라 문장 내에서 VAD (Voice Activity Detection)에 대한 적용을 계획 중이다.

참 고 문 헌

[1] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signal,* pp.130-134, New Jersey: Prentice-Hall, 1978.

[2] L. R. Rabiner, M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal,* Vol. 54, No. 2 pp.297-315, 1975.

[3] J. C. Junqa, B. Mark, B. Reaves, "A robust algorithm for word boundary detection in the presence of noise", *IEEE Trans. Speech and Audio Processing,* Vol. 2, No. 3, pp.406-412, July 1994.

[4] C. Tsao, R. M. Gray, "An endpoint detector for lpc speech using residual error look-ahead for vector quantization applications", *Proc. ICASSP-84.*, pp.18b.7.1-4, 1984.

[5] J. C. Junqa, B. Mark, B. Reaves, "A robust speech/non speech detection algorithm using time and frequency-based features", *Processing of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. I, pp.269-272, 1992.

접수일자: 11월 11일
게재결정: 12월 12일

▶ 김영민(Youngmin Kim)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6206
E-mail: ymkim@icu.ac.kr

▶ 한민수(Minsoo Hahn)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학원대학교
소속: 한국정보통신대학원대학교(ICU) 음성/음향 정보 연구실
전화: 042) 866-6123
E-mail mshahn@icu.ac.kr