1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

반복적 스펙트럼 차감법을 이용한 잡음 음성의 무음 구간 검출

조 훈 영 ○ 오 영 환

한국과학기술원 전산학과

The detection of Nonspeech Interval in Noisy Speech using Iterative Spectral Subtraction

Hoon-Young Cho ○ Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

{nymph,yhoh}@bulsai.kaist.ac.kr

요 약

본 논문에서는 극심한 가산 잡음에 의해 손상된 음성 신호를 스펙트럼 차감법으로 개선할 때, 잡음 스펙트럼 추정을 위한 무음 구간 추정 방법을 제안한다. 스펙트럼 차감법은 잡음을 효과적으로 제거한다고 알려져 있으나, SNR 0 dB 이하의 잡음 환경에서는 무음 구간의 검출이 힘들어 잡음 스펙트럼 추정치의 정확도가 저하된다. 일반화 스펙트럼 차감법의 과차감(oversubtraction)과 잡음 스펙트럼 추정을 반복하여 얻은 무음 구간은 SNR -10 dB ~ 0 dB의 낮은 SNR에서도 비교적 정확하며, 프레임 에너지를 이용한 무음 검출 방법에 비해 향상된 성능을 보였다.

1 서론

근래에 음성을 인간과 기계 간의 인터페이스 수단으로 활용하는 응용 시스템들이 증가하고 있다. 음성 인식을 이용한 컴퓨터의 구동, 화자 인식을 이용한 출입 통제, 이동통신 단말기에서의 음성 인식, 원격지 호텔 예약을 위한 전화 음성 인식 등 주변에서 그 예를 쉽게 찾아볼 수가 있게 되었다.

이처럼 음성을 기계 구동의 수단으로 사용함에 있어 가장 큰 문제점은 음성을 여러 형태로 왜곡시키는 잡음 신호이며, 이로 인해 응용 시스템의 성능이 크게 저하된다. 잡음은 여러 관점에서 구분할 수 있으나, 잡음이 신호를 왜곡시키는 방식에 따라 크게 가산 잡음과 곱셈성분 잡음으로 나눌 수 있다. 가산 잡음은 시간 영역에서 신호에 직접 더해지는 잡음이고, 곱셈성분 잡음은 신호에 대해 선형 필터링 효과를 나타낸다.

가산 잡음 환경에서 왜곡된 잡음 음성(noisy speech)에서 음성을 복원 또는 인식하고자 할 때 잡음의 추정이 필수적이다. 일반적으로는 신호의 무음 구간에서 잡음을 추정하므로 무음 구간의 정확한 검출은 시스템의 전체 성능에 매우 큰 영향을 미친다. 기존의 음질 개선 방법들은 무음 구간의 추정이 적절하게 수행되었다고 가정하고 그 후의 처리에 중점을 두고 있다. 그러나, 잡음의 크기가 매우 클 때에는 정확한 무음 구간의 검출이 어려워지므로, 이에 대한 연구가 필수적이다. 잡음 음성을 인식할 때에도 무음 구간의 정확한 검출은 잡음의 영향을 줄이거나, 음성 영역을 검출하여 인식률을 높이는 데 매우 중요한 역할을 한다 [8].

본 연구에서는 가산 잡음의 처리에 널리 응용되는 스펙트럼 차감법의 성질을 이용하여, SNR이 매우 낮은 잡음 음성에 대하여 무음 구간의 검출 성능을 높임과 동시에 음질 개선 성능을 더욱 향상시키고자 한다. 제안한 방법은 잡음 음성의 초기 무음 구간 추정으로부터, 무음 구간이 더 이상 변동하지 않을 때까지 일반화 스펙트럼 차감법의 과차감과 잡음의 추정을 반복한다. 수렴한 후의 무음 구간에서 추정한 잡음 스펙트럼은 좀 더 정확하며, 기존의 스펙트럼 차감법의 음질 개선 성능을 향상시킬 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 스펙트럼 차감법과 이를 일반화한 경우에 대해 알아보고, 이 방법의 문제점들을 살펴본다. 3장에서는 반복적 스펙트럼 과차감에 의한 무음 구간의 추정에 대해 설명하며, 4장에서 여러 가지 잡음 환경에 대해 수행한 실험 결과를 기술하고, 5장에서 결론을 맺는다.

2 스펙트럼 차감법

스펙트럼 차감법은 잡음 음성의 무음 구간에서 잡음 스펙트럼의 평균을 구하고, 잡음 음성의 스펙트럼에서 차감하여 음성을 복원한다. 이 방법에서는 음성과 잡음이 상관되어있지 않다는 가정하에 잡음 음성을 다음과 같이 모델링한다.

$$y[n] = s[n] + d[n] \tag{1}$$

$y[n]$은 잡음이 섞인 잡음 음성의 샘플, $s[n]$은 잡음이 섞이지 않은 음성, 그리고 $d[n]$은 잡음 신호를 의미한다. 이를 일정 구간의 프레임들로 나누어 스펙트럼 영역에서 표현하면, 프레임 m에서의 스펙트럼은

$$Y_m(\omega) = S_m(\omega) + D_m(\omega) \tag{2}$$

그림 1: 스펙트럼 차감법을 이용한 음질 개선

이며, 스펙트럼 차감법의 가장 일반화된 형태는 식 3과 같다 [3] [6].

$$\left|\hat{S}_m(\omega)\right| = \begin{cases} \left(|Y_m(\omega)|^\alpha - \beta\left|\hat{N}(\omega)\right|^\alpha\right)^{1/\alpha}, & \text{if } |Y_m(\omega)|^\alpha > \beta\left|\hat{N}(\omega)\right|^\alpha \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$\hat{N}(\omega) = \hat{D}_m(\omega) \quad (4)$$

$$arg\left(\hat{S}_m(\omega)\right) = arg\left(Y_m(\omega)\right) \quad (5)$$

α, β의 값은 양수이며, $\hat{N}(\omega)$는 무음 구간에서 추정한 잡음 스펙트럼이다. 식 5는 복원된 음성의 위상이며, 잡음 음성의 위상을 동일하게 사용한다. 원음은 $|\hat{S}_m(\omega)|$와 $arg\left(\hat{S}_m(\omega)\right)$으로부터 푸리에 역변환과 overlap-add 방식으로 얻으며 그림 1과 같다. 연구결과에 의하면, α는 명료도에 관련된 변수로서 이 값이 2인 경우가, 1 또는 0.5인 경우에 비해 높은 명료도를 보이지만, 스펙트럼 차감법 적용 후에 발생하는 음악 잡음은 더 심하다. 또, β는 차감하려는 잡음 스펙트럼의 크기를 조절하는 조절하는 변수로서 이 값이 크면 과차감(oversubtraction)이라고 하며, 음악 잡음을 효과적으로 제거할 수 있지만 음성의 명료도도 함께 저하된다. β를 과차감 인자라고도 한다.

스펙트럼 차감법은 잡음을 비교적 잘 제거하여 SNR을 향상시키지만, 잡음 제거 후의 잔여 잡음 스펙트럼 상에 비규칙적으로 나타나는 짧은 길이의 스펙트럼 피크들에 의해 음악 잡음(musical noise)이 생성되는 문제점이 있다 [1]. 이를 해결하기 위해 잡음 스펙트럼에 일정한 인자를 곱하여 잡음 음성에서 과차감함과 동시에 음악 잡음 효과를 없애기 위해 잔여 잡음의 스펙트럼을 매스킹하는 방법 [2], 스펙트럼 차감 후의 스펙트로그램 상에서 이미지 처리기법과 음성의 특별한 성질을 응용하여 음악 잡음을 없애는 방법 [6] 등이 제안되었다. 이 외에도 Ephraim과 Malah가 제안한 잡음 감쇠법을 응용하여 잔여 잡음을 감쇠하거나 [4] [5], 스펙트럼 차감 전과 차감 후의 두 단계에 심리음향학적 지식에 기반한 매스킹 문턱치를 두어 음악 잡음을 없애는 방법 등이 제안되었다 [9]

지금까지 언급한 방법들은 주로 스펙트럼 차감법 적용 후에 발생한 문제를 다룬 것들이다. 그러나, SNR이 0 dB이하인 극심한 잡음 환경 하에서는 스펙트럼 차감법을 적용하기 이전에 무음 구간의 검출이 어려워지므로, $|\hat{N}(\omega)|$의 정확한 추정이 힘들다.

기존의 에너지에 기반한 음성 및 무음 구간의 검출 방법을 간략히 서술하면 다음과 같다. 입력 신호의 처음 몇 프레임을 무음 구간이라고 가정하고, 이 프레임들의 에너지를 구한다. 이 값들의 평균 및 표준편차를 각각 μ_E, σ_E라고 할 때, 임계치 $\theta = \mu_E - \kappa \cdot \sigma_E$를 결정한다. 음성의 매 프레임을 임계치와 비교하여 에너지가 낮은 프레임을 무음 구간에 포함한다 [10]. 이 방법은 일반적으로 좋은 성능을 보여 음성 인식 등에 많이 적용되었으나, 잡음이 매우 클 때는 성능이 저하되므로 좀 더 잡음에 강인한 방법이 필요하다.

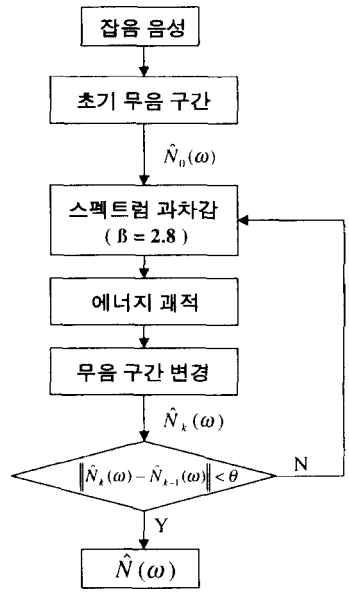

그림 2: 반복적 스펙트럼 차감을 이용한 무음 구간 및 잡음 스펙트럼의 추정

3 무음 구간 검출

잡음 음성을 과차감하여 복원한 음성의 파형은 잡음 음성에 비해 음성과 무음 구간이 훨씬 명확하게 구별된다. 따라서, 이 때의 무음 구간에서 추정한 잡음 스펙트럼은 과차감시에 사용한 잡음 스펙트럼에 비해 더욱 정확하다고 할 수 있다. 새로 추정한 잡음 스펙트럼으로 다시 과차감을 수행하여 음성을 복원하면 좀 더 원음의 파형에 가까운 신호를 얻을 수 있다.

본 연구에서는 이같은 사실을 이용하여 과차감과 차감할 잡음 스펙트럼의 추정을 반복하여 수행함으로써 무음 구간 검출의 정확도를 높이고자 한다. 제안한 방법은 그림 2에 보인 바와 같으며, 이를 초기화 단계, 반복적 무음 구간 추정 단계와 최종 단계로 나누어 설명하기로 한다.

초기화 단계

잡음 음성 신호에서 각각 분석 프레임들에 대해 샘플의 평균 제곱합을 dB 단위로 환산하여 에너지 궤적을 얻는다. 음성 신호의 전체 길이가 충분히 크다고 할 때, 최소의 에너지값을 갖는 몇 개의 프레임들은 무음 구간의 일부분이라고 판단할 수 있다. 따라서, 최소 에너지를 갖는 5개의 프레임에서 잡음 스펙트럼의 초기 추정치 $|\hat{N}_0(\omega)|$를 구한다.

반복 추정 단계

반복 회수가 $k-1$일 때의 잡음 스펙트럼을 $|\hat{N}_{k-1}(\omega)|$라고 하면, 이 추정치에 과차감 인자 $\beta = 2.8$을 곱하여 잡음 음성의 스펙트럼 $|Y_m(\omega)|$에서 차감한다. 그 후, 각 프레임에 대해 모든 ω에서 스펙트럼값을 더하여 프레임별 에너지 궤적을 구한다. 에너지 궤적의 평균치 μ_E와 표준편차 σ_E를 사용하여 임계치 $\mu_E - \kappa \cdot \sigma_E$를 정하고, 에너지가 임계치보다 작은 프레임들을 무음 구간으로 판단한다.

최종 추정 단계

새로운 무음 구간에서 추정한 잡음 스펙트럼 $|\hat{N}_k(\omega)|$와 이전 반복에서의 $|\hat{N}_{k-1}(\omega)|$의 스펙트럼 차이를 계산하여 이 값

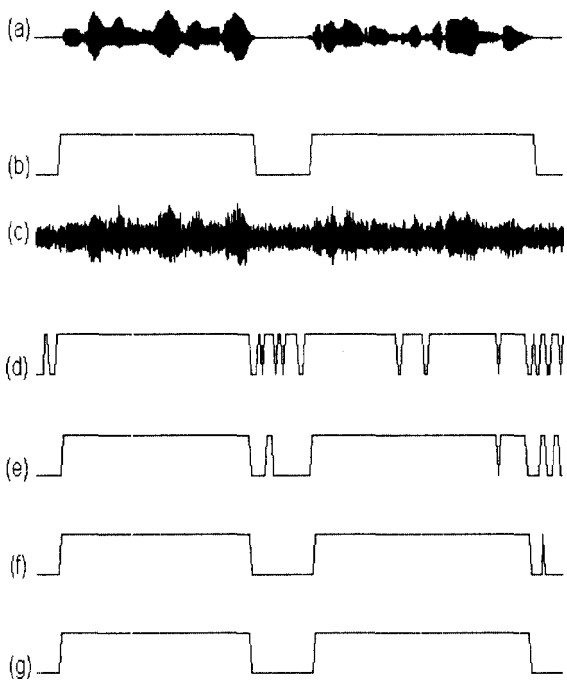

그림 3: 반복 회수의 증가에 따른 음성 및 무음 구간의 검출 (a) 음성 신호의 파형 (b) 음성 신호의 음성 및 무음 구간 (c) SNR 0 dB의 자동차 잡음이 가산된 잡음 음성 (d) ~ (g) 반복 회수가 1 ~ 4일 때의 음성 및 무음 구간의 수렴 과정

이 정해진 임계치 θ보다 작으면 반복을 끝마친다. $|\hat{N}_k(\omega)|$는 확정된 잡음 스펙트럼으로서 그림 1의 $\hat{N}(\omega)$에 해당된다.

이상에서 기술한 알고리즘에서 β값은 여러 번의 실험을 통해 구해진 값으로서 이 값이 작으면 $|\hat{N}(\omega)|$값이 수렴하지 않을 수 있다. 또한, 이 값이 너무 크면 추정한 무음 구간이 무성음 구간까지도 포함하게 된다.

4 실험 및 결과

실험에 사용된 잡음은 16 kHz의 백색 잡음과 유색 잡음, 고속도로에서 창문을 닫고 70 km/h로 달리는 자동차 내부의 소음, F16 cockpit 잡음, 100명의 사람들이 모인 곳에서 수집한 babble 잡음과 충소리 잡음이다. 음성 신호는 조용한 환경에서 16 kHz로 샘플링된 녹음한 여성 화자의 발성음을 사용하였다. 잡음 음성은 음성에 잡음 신호를 가산하였으며, SNR -10 dB, -5 dB, 0 dB, 5 dB, 10 dB 등 여러 종류의 SNR을 모의하였다.

신호의 분석은 잡음 음성 신호에서 20 ms 길이의 프레임에 해닝 창(hanning window)을 곱하여 이산 푸리에 변환하였으며, 10 ms씩 이동하였다. 잡음 스펙트럼의 추정을 위해서는 먼저 무음 구간의 추정이 필요하며, 이 때 3장에서 제안한 방법을 사용한다. 그림 3은 SNR 0 dB의 자동차 소음에 대해 제안한 방법을 적용할 경우, 반복 회수에 따라 음성 및 비음성 구간 검출이 개선되는 것을 보인다. 그림에서 알 수 있듯이 반복 회수가 3회 이상만 되어도 상당히 정확한 무음 구간 검출이 가능하다. 이 때 $|\hat{N}_k(\omega)|$는 무음 구간의 변동이 줄어듦에 따라 일정한 형태의 스펙트럼으로 수렴하는데, 이는 순수 잡음 신호에서 계산한 평균 스펙트럼과 거의 동일하였다. 과차감 인자 β의 값은 여러 번의 예비 실험을 통해 구한 2.8을 사용하였으며, 이보다 작을 경우, 무음 구간이 일정한 위치로 수렴하지 않는 경우도 있다.

표 1은 여러 종류의 잡음과 SNR에 대해서 음성 및 비음성 구간의 검출 정확도를 나타낸 것이다. 검출 정확도는 잡음이 가산되기 전의 원음에서 계산한 프레임별 음성 및 비음성 정보를 기준으로 하여, 이와 일치하는 프레임의 개수를 전체 프레임의 개수로 나누어 백분율을 구하였다. SNR이 높을수록 검출 정확도가 향상되며, SNR이 -10 dB인 극심한 잡음에 대해서도 상당히 높은 정확도를 보임을 알 수 있다. 제안한 방법은 백색 잡음, 자동차 잡음 및 F16 cockpit 잡음과 같이 거의 정상적(semi-stationary)인 잡음에 대해 효과적이며, babble 잡음 등의 시간에 따라 변경적이고, 광대역 잡음인 경우에 대해 정확도가 낮았다. 또한, 충소리 잡음과 같은 충격(impulsive) 잡음에 대해서도 비교적 좋은 성능을 보였다.

표 1: 잡음의 종류 및 SNR별 음성/비음성 구간 검출 정확도

	-10 dB	-5 dB	0 dB	5 dB	10 dB
백색 잡음	80.6	89.5	93.7	97.9	99.2
유색 잡음	-	78.9	84.8	92.4	97.9
F16 cockpit	78.1	81.9	90.7	94.5	97.5
자동차 잡음	86.9	93.2	97.9	98.3	96.2
babble 잡음	73.4	73.4	77.6	84.8	87.3
충소리 잡음	90.3	89.0	89.5	89.0	89.0

그림 4는 기존의 프레임 에너지에 기반한 음성 및 비음성 검출 방법과 제안한 방법의 성능을 여러가지 잡음에 대해 비교한 것이다. 프레임 에너지에 기반한 방법은 2장에서 설명한 바와 같으며, 이 때 임계치로는 예비 실험에서 가장 좋은 성능을 나타낸 $\kappa = 0.8$을 사용하였다.

실험 결과에 의하면, 대부분의 잡음에 대해 제안한 방법이 더 나은 검출률을 보였다. 한편, 에너지에 기반한 방법은 충소리 등의 충격 잡음에 대해 매우 낮은 정확도를 보였다. 이는 충격 잡음의 경우, 신호의 첫 부분에 잡음의 존재 여부가 불분명하므로, 적절한 임계치를 구하기 힘들기 때문이다. 잡음이 존재할 때, 무성음 부분처럼 에너지가 작은 프레임들은 무음 구간으로 추정되기 쉽다. 잡음이 존재하지 않는 때에도 신호의 다른 무음 부분에 잡음이 존재한 경우, 이를 음성으로 검출할 가능성이 높다. 제안한 방법은 신호의 여러 부분에 존재하는 무음 구간에서 잡음에 관한 추정치를 구하며, 반복 회수가 증가할 수록 개선된 잡음의 추정치를 얻게 되므로 더 나은 성능을 보였다.

5 결론

본 논문에서는 극심한 잡음 환경에서 음질을 개선할 때 필요한 무음 구간의 검출 방법을 제안하였다. 제안한 방법은 일반

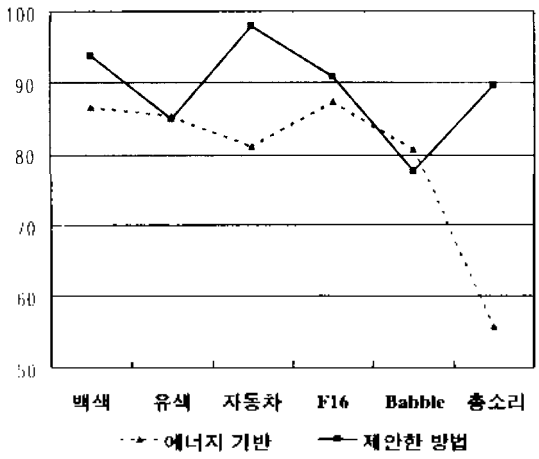

그림 4: 유성/비음성 검출 방법의 성능 비교 : SNR 0 dB 잡음 음성

화 스펙트럼 차감법의 과차감을 반복하여 수행하며, 무음 구간을 재추정한다. 다양한 잡음 환경에 대해 에너지 기반의 방법과 비교 실험해 본 결과, 제안한 방법이 유효함을 확인할 수 있었다. 제안한 방법은 특히 고속 주행 중인 자동차 소음 환경에서 음성을 인식할 때에 필수적인 끝점 검출 방법 등으로도 적용이 가능할 것으로 판단되며, 스펙트럼 차감법의 성능을 보다 향상시키기 위해서는 음악 잡음 문제를 해결하고, 실험적으로 결정되는 여러 가지 파라미터값들을 잡음의 특성에 따라 자동으로 추정하는 향후 연구가 필요하다.

참고 문헌

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. ICASSP*, pp. 208-211, Apr. 1979.

[3] B.L. Sim, Y.C. Tong, J.S. Chang and C.T Tan, "A Parametric Formulation of the Generalized Spectral Subtraction Method," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 328-337, July 1998.

[4] O. Cappé, "Elimination of the musical noise phenominon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345-349, Apr. 1994.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[6] Z. Goh, K.C. Tan and B.T.G. Tan, "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 287-292, May 1998.

[7] J.S. Lim and A.V. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

[8] C.E. Mokbel and F.A. Chollet, "Automatic Word Recognition in Cars," *IEEE trans. Speech Audio Processing*, vol. 3, pp. 346-356, Sep. 1995.

[9] T. Haulick, K. Linhard and P. Schrögmeier, "Residual Noise Suppression Using Psychoacoustic Criteria," *Proc. EUROSPEECH*, pp. 1395-1398, Sep. 1997.

[10] L.R. Rabiner and R.W. Shafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.