한국음향학회지 제29권 제2호 pp. 148~153 (2010)

연속음성신호에서 IMBE 모델을 이용한 SNR 추정 연구

IMBE Model Based SNR Estimation of Continuous Speech Signals

박 형 우*, 배 명 진*
(Hyung-Woo Park*, Myung-Jin Bae*)

*숭실대학교 정보통신공학학과
(접수일자: 2009년 12월 8일; 수정일자: 2010년 1월 27일; 채택일자: 2010년 2월 7일)

음성 신호처리 환경에서 잡음이 섞인 신호를 개선할 목적으로 음성향상 기법이 많이 이용되고 있다. 잡음추정 알고리즘은 변화하는 환경에 빠르게 적응할 수 있어야 하며 음성신호의 영향을 줄이기 위해 음성신호가 존재하지 않는 구간에서만 잡음의 파워를 갱신한다. 이러한 방법은 음성구간검출이 선행되어야 한다. 그러나 잡음에 열화된 음성신호에 묵음구간이 존재하지 않을 경우, 위와 같이 음성검출을 통한 묵음구간에서의 잡음 추정 방법 및 SNR 추정 방법이 적용될 수 없다. 본 논문에서는 묵음구간이 존재하지 않는 연속음성신호에서 SNR을 추정하는 기법을 제안한다. 음성신호는 MBE (Multi-Band Excitation) 발성 모델에 따라 유·무성음으로 구분할 수 있다. 그리고 에너지가 유성음에 대부분 분포하기 때문에, 부가성 잡음 환경에서 유성음의 에너지를 음성신호의 에너지로 근사화하여 SNR을 추정할 수 있다. 제안하는 방식은 연속음성신호를 IMBE (Improved Multi-Band Exciation) 보코더를 이용해 유·무성음 대역으로 구분하고, 각각 대역의 에너지 정보를 이용하여 단구간 음성신호의 SNR을 계산한다. 전체 음성구간의 SNR은 단구간 SNR의 평균값을 통해 추정한다.
핵심용어: SNR 추정, 연속음성신호, IMBE 모델, 에너지 비교
투고분야: 음성처리 분야 (2.3)

In speech signal processing, speech signal corrupted by noise should be enhanced to improve quality. Usually noise estimation methods need flexibility for variable environment. Noise profile is renewed on silence region to avoid effects of speech properties. So we have to preprocess finding voice region before noise estimation. However, if received signal does not have silence region, we cannot apply that method.

In this paper, we proposed SNR estimation method for continuous speech signal. A Speech signal consists of Voice and Unvoiced Band in The MBE excitation model. And the energy of speech signal is mostly distributed on voiced region, so we can estimate SNR by the ratio of voiced region energy to unvoiced. We use the IMBE vocoder for classifying the Voice or Unvoice band of segmented speech signal. Continuously we calculate the segmented SNR using that information and the energy of each band. And we estimate the SNR of continuous speech signal.
Keywords: SNR Estimation, Continuous Speech, IMBE Model, Energy Compare
ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

우리는 생활에서 음성을 기본적인 의사소통 수단중 하나로 사용한다. 이러한 음성은 문자나 그림과 같이 견고한 매체를 통해 전달되거나 저장되지 않고 공기, 통신선로, 다양한 형태 변질되는 저장매체를 통해 유연하게 전달 및 저장된다. 유연한 환경은 사용하기 편리하다는 장점이 있지만, 주변 환경에 영향을 크게 받는다. 즉, 의사소통을 하는데 있어 환경 잡음의 크기가 정보 전달 양과 질을 결정하게 된다.

오늘날 컴퓨터를 이용한 정보기술이 발전함에 따라 이를 이용한 음성신호처리 분야도 눈부시게 발전하고 있다. 궁극적으로 사람과 컴퓨터가 원활하게 의사소통을 하거나, 사람과 사람이 정보통신기기를 이용해 정보를 전달하기 위해 다양한 분야에서 연구가 이루어지고 있다. 그중 음성분석, 인식시스템을 통해 사람의 의사를 컴퓨터에 전하는 경우, 실험실 수준의 조용한 환경에서는

책임저자: 배 명 진 (mjbae@ssu.ac.kr)
156-743 서울시 동작구 상도동 511 숭실대학교 정보통신공학과
(전화: 02-820-0902; 팩스: 02-814-0608)

우수한 성능을 보인다. 하지만 다양한 잡음이 존재하는 실제 상황에서는 성능이 저하된다. 그래서 음성신호를 향상하기위해 다양한 방법이 제안되었다 [3][4].

음성신호의 품질향상을 위해서는 잡음의 크기를 아는 것이 중요하다. 특히 스펙트럼 차감법에서는 잡음의 양을 알아내어 스펙트럼 상에서 빼기 때문에 그 크기를 안다는 것은 매우 중요하다. 이러한 의미에서 잡음의 양을 추정하거나 측정하기 위해 기존에는 신호의 묵음구간에서 잡음의 에너지를 측정하는 방법이 이용되었다. 하지만 이 방식은 잡음에 열화된 음성신호에 경우 묵음구간을 검출하기 어렵고, 이것은 잡음의 크기를 검출하지 못하게 된다 [5][6]. 그리고 연속음성신호에서 확률적으로 잡음을 추정하여 잡음을 제거하는 경우 추정된 잡음이 변화하면 음성신호에도 영향을 주게 된다 [7][10].

본 논문에서는 연속 음성 입력신호가 잡음에 열화된 경우, IMBE (Improved Multi-Band Excitation)모델을 이용하여 단구간으로 나눈 음성구간의 SNR을 추정하고 이렇게 구해진 SSNR (Segmented SNR)을 종합하여 전체 음성신호의 SNR을 구하는 방법을 제안하고자한다. 제안하는 방법은 기존의 방식인 음성구간 검출기를 사용하지 않아도 되고, 연속음성 구간의 유·무성음 대역의 정보를 이용하므로, 잡음의 양이 변화하는 경우에도 적용할 수 있다. 논문의 구성은 서론에 이어 2장에서 음성생성모델과 성도성분을 추출하기 위한 선형예측분석에 대해 기술한다. 3장에서는 제안한 SNR 추정기법을 설명하고, 4장에서는 실험 및 결과를 설명하고 마지막으로 결론을 맺는다.

II. 음성 생성 모델

2.1. 일반적인 음성생성모델

음성생성에 대한 선형모델은 50년대 후반 Fant에 의해 개발되었는데, 음성출력을 음원이 여파기를 통과하여 나오는 신호로 가정하고, 음원과 성도의 각 부분을 독립적인 것으로 간주하는 선형예측모델을 제시하였다. 음원에 대한 모델로 유성음의 음원은 준주기적인 펄스, 무성음의 음원은 백색잡음을 사용하였고, 성대에서 성문이 음원에 미치는 영향은 성문모델로 모델링하였으며 [8], 이는 그림 1과 같이 나타낼 수 있다.

2.2. IMBE 음성생성 모델

음성생성 모델은 Fant에 의한 선형예측 모델을 주로 이용하였다. 이 선형예측모델은 음질이 전반적으로 저하되며, 주위 환경의 잡음에 민감하다는 단점이 있다 [1][2]. 이러한 단점을 극복하기 위해 여기원에 초점을 맞춰, Multi-Band Excitation (MBE) 모델이 제안되었다. 유·무성음 분류를 갖은 스펙트럼 구간을 이용해 기존의 음성생성 모델의 여기원 부분을 다양화 하였으며, 음질 개선을 이루었다 [1][2][9].

MBE음성 생성모델과 기존의 선형예측을 이용한 음성생성모델에 차이점은 다음과 같다. 기존의 방법은 단구간의 음성 구간을 유·무성음 구간으로 나눈다. 하지만, MBE 음성 생성모델에서는 단구간의 음성 신호를 기본주파수를 이용해 주파수 대역에서 몇 개의 대역으로 단구간 신호의 스펙트럼을 나누고 각각 대역의 유·무성음 정보를 결정한다. 유·무성음 대역으로 구분하는 방법은 단구간의 음성 에너지, 각각 밴드의 에너지, 그리고 이전 구간에서의 유·무성음 대역정보를 파라미터화 하여 비교하는 것이다 [2].

그림 2는 MBE 음성생성모델의 블록도 이다. 그림 1의 선형예측모델과 다른점은 유·무성음을 선택하는 변환 스위치대신 짧은 음성구간의 대역별 유·무성음 정보가 음성신호의 여기원이 되는 점이다 [8]. 최종적으로 음성을 합성하는 부분에서는 주기적인 스펙트럼은 유성음으로 결정된 대역에서 이용하고, 백색잡음 스펙트럼은 무성음으로 결정된 구간에서 이용하여 Fant가 제안한 선형모델에 비하여 음질이 우수한 음성 생성 모델이 된다.

IMBE 모델은 Digital Voice System에서 구현한 음성

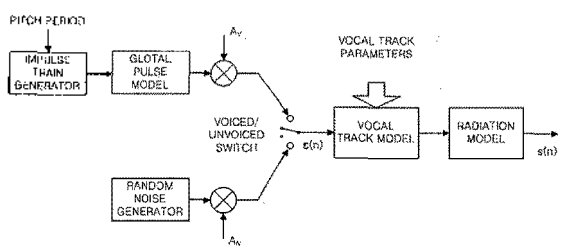

그림 1. 일반적인 음성생성모델
Fig. 1. Speech system model.

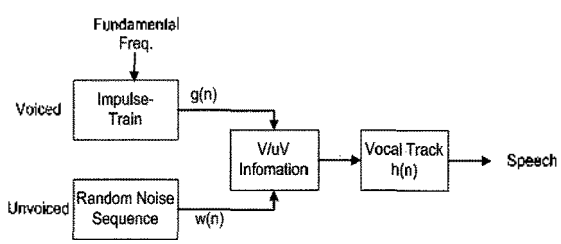

그림 5. MBE 음성생성 모델
Fig. 2. MBE Speech system Model.

부호화 기술로 MBE 음성 생성 모델에 기반을 하고 실제 유·무성 대역 구분을 통하여 합성되는 음질을 높이는데 이용된다 [2][9].

III. 제안하는 SNR 추정법

본 절에서는 잡음 환경에서 수신된 음성신호를 IMBE 모델로 분석하여 SNR을 추정하는 알고리즘을 제안한다. 연속음성 신호의 경우, 묵음구간이 존재하지 않으므로 기존의 음성구간 검출기의 출력을 이용해 SNR을 구할 수 없다. 하지만, 제안하는 방식을 이용하면 수신된 음성을 묵음이나 잡음 구간으로 나누는 것이 아닌 주파수 대역에서 신호와 잡음의 크기를 추정하므로 연속음성신호의 SNR을 얻을 수 있다. 따라서 VAD와 같은 음성구간 검출 방법이 필요하지 않으며, 잡음에 열화된 음성을 통해서 직접 SNR을 추정할 수 있다.

기존의 방식은 VAD (Voice Active Detector)를 이용해 묵음 구간과 음성 신호구간을 나누고 묵음구간의 에너지 계산을 통해 SNR을 추정하게 된다 [6]. 기존의 방식으로는 음성이 지속적으로 발생하는 구간이나 잡음 환경이 변화하는 등의 경우에는 잡음의 양을 알아내기 어렵다 [5][10]. 하지만, 제안 하는 방식은 연속음성신호를 이용하고 묵음구간만을 이용하는 것이 아니라 전체 구간의 분절된 시간에서의 에너지양을 이용하므로 이용도중 잡음환경이 변하거나 묵음 구간을 검출할 수 없는 때에도 이용이 가능하다.

제안하는 SNR 추정방법의 원리는 음성신호의 특성상 대부분의 에너지가 유성음에 존재하는 점을 이용한다. 음성신호에 백색 가산 잡음이 부가되는 경우, 주파수 스펙트럼 상에서 전 대역에 잡음이 더해진다. 하지만 잡음의 영향은 유성음과 무성음에서 다르게 영향을 주게 된다. 유성음의 경우 기본적인 에너지의 크기가 크기 때문에 상대적인 잡음의 영향이 작게 나타나고, 무성음은 크게 된다. 따라서 잡음환경에서 연속음성신호를 수신한 경우 IMBE 모델을 이용하여 신호를 유·무성음 대역으로 구분하여, 대역의 에너지를 이용하여 잡음이 부가된 척도를 가늠하는 파라미터를 만들어 신호의 SNR을 추정하게 된다. 아래 식 (1)이 잡음의 양을 추정하는 파라미터를 구하는 수식이다.

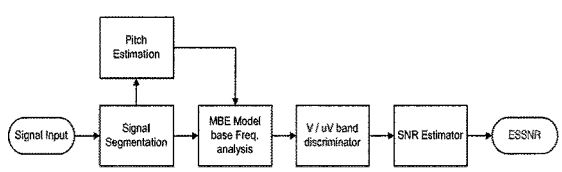

그림 3. 제안하는 방법의 순서도
Fig. 3. Flowchart of Proposed Algorithm.

$$ESNR = 10\log_{10}\left(\frac{\frac{1}{N}\sum_{i\in Voiced}^{N}\sum_{i=k_p}^{B} f_i^2}{\frac{1}{M}\sum_{j\in unvoiced}^{M}\sum_{j=k_q}^{B} f_j^2}\right) \quad (1)$$

여기서 f_i는 유·무성음으로 구분된 대역에 주파수의 크기 값이고, B는 구분된 주파수 대역의 크기이다. k_p와 k_q는 구분된 주파수의 시작점이고, N은 유성음으로 구분된 대역이고 M은 무성음으로 구분된 대역이다. 유성음과 무성음의 에너지로 근사화 하기 위해 각각의 개수로 정규화 하였다. 이렇게 얻어진 값을 이용해 추정된 SNR값인 ESNR (Estimated Signal to Noise ratio)를 구한다.

제안하는 방법을 그림 3에서 순서도로 표현 하였다. 신호가 입력되면 분석 가능한 짧은 시간단위로 분절을 하고 분절된 프레임의 피치를 추정한다. 추정된 피치를 이용해 IMBE 알고리즘으로 해석을 하여 대역을 구분한다. 그리고 SNR Estimator에서 유·무성음 대역의 정보를 이용하여 분절된 프레임별 SSNR을 계산하고, 식 2와 같이 SSNR의 평균을 취하여 전체 음성신호의 SNR을 얻는다. 식2에서 n은 전체 음성구간의 프레임 개수이고 SSNR은 각각 프레임에서의 추정된 SNR이다.

$$SNR = \frac{\sum_{k=1}^{n} SSNR_n}{n} \quad (2)$$

3.1. 피치 추정

IMBE 알고리즘을 이용하여 음성신호를 분석하는데 기본이 되는 정보는 피치주기이다. 피치 정보를 이용하여 주파수 대역을 나누고 분석구간을 결정하기 때문에 정확한 피치주기를 알아내야 한다. 피치 주기에 앞서 신호를 짧은 구간으로 분절하는 과정이 있는데 이는 FFT 윈도우 사이즈를 고려하여 512 샘플이 되도록 하였으며 해밍윈도우를 이용하였다. 512 샘플 구간에서 피치 검색과정은 시간영역 자기상관함수를 이용하여 주기를 강조하고 그 반복성이 검출되는 첨두치를 검출하였다. 그림 4에 음성파형과 검출된 피치 주기를 표현하였다.

3.2. 유·무성음 대역 판단

3.1절에서 구해진 피치 정보를 이용해 분석하는 프레임의 주파수 대역을 나누고, IMBE 모델을 분석을 통해 분석하는 프레임의 유·무성음 대역을 판단한다. 대역을 나누고 IMBE 알고리즘으로 해석을 하는 과정은 연속음성신호의 SNR을 추정하는데 중요한 파라미터가 된다.

그림 5는 잡음이 부가되지 않은 신호의 IMBE 모델을 이용해 얻어진 정보를 이용해 유·성음 스펙트럼을 구분하여 나타냈다. 유성음으로 유력하게 판단하는 시간영역의 프레임에서도 그림 5와 같이 가운데의 유성음 대역과 아래의 무성음 대역으로 나누어지게 되고, 각 대역의 에너지를 파라미터화 하여 프레임의 SNR을 추정하는데 이용하였다.

3.3. SNR 추정

IMBE 모델을 이용하여 유·무성음 대역으로 나누고 대역의 에너지를 파라미터화 하여 SNR을 추정한 하였다. 이때 추정된 SNR은 식 (1)을 이용하여 구하였다. 시간의 변화에 따른 SSNR선도는 그림 6과 같이 얻을 수 있었다. 이 실험 결과는 백색잡음을 SNR이 0 dB가 되도록 더하여 제안된 추정방법에 의해 SNR을 추정한 결과 선도 이다. 점선은 잡음을 부가하기 전의 원음성신호와 알고 있는 크기의 잡음과 원신호의 SSNR을 계산하여 나타낸 값이다. 실선은 제안된 방법으로 SSNR을 추정하여 얻은 선도이다. 실선의 SSNR 추정치가 완벽하게 SSNR (점선)을 추종하지는 못하지만, 그 형태와 경향이 유사하게 나타남을 그림 6을 통하여 확인하였다.

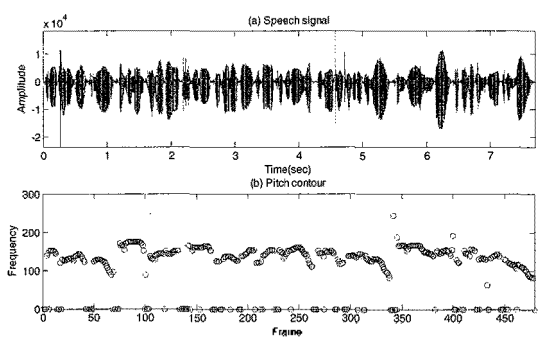

그림 4. 음성파형과 피치 추정기 결과 파형
Fig. 4. Speech signal and Pitch estimatation.

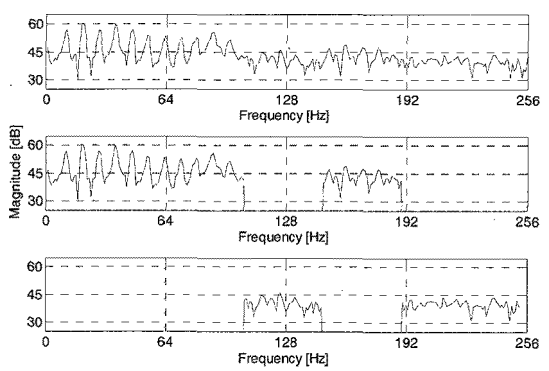

그림 5. 유·무성 대역으로 나누어진 스펙트럼
Fig. 5. Voice and Unvoice Spectrum after MBE model.

IV. 실험 및 결과

4.1. 실험 및 데이터 수집 환경

제안된 알고리즘의 성능평가를 위해 백색 잡음을 포함하여 다양한 잡음환경에서 객관적 테스트를 수행하였다. 실험에 사용된 데이터는 남성과 여성 각 5명이 반복적으로 발성한 '국민교육헌장'을 이용하였다. 연속 음성신호에 대해서 테스트하기 위해 신호구간 중 묵음구간을 검출하여 샘플에서 제거하였으며, 이때 실험 데이터는 8 kHz로 샘플링하고, 샘플당 양자화 비트수는 16 bits/sample를 사용하였다. 시간영역에서 프레임 (윈도우)의 길이는 일반적으로 20 ~ 40 ms가 사용되지만, 본 논문에서는 FFT 윈도우 사이즈를 고려하여 64 ms로 하였다. 이때 프레임에 따른 샘플의 개수는 512개이다. 윈도우 오버랩은 음성신호의 SNR 추정시 일반적으로 50 %이하에서 좋은 성능을 나타내며 [7][8], 본 논문에서는 25 %로 오버랩하여 프레임 단위로 SNR 추정을 수행하였다. 프레임을 나누는데 이용한 윈도우는 해밍 윈도우 (Hamming Window) 512 샘플을 이용하였다.

4.2. 실험 결과

그림 7과 그림 8에서 진행된 실험결과를 확인 할 수 있다. 그림 7에서는 백색잡음 환경과 선박 엔진 잡음환경을 실험 하였고, 그림 8에서는 자동차 내부 잡음 환경과 전투기 조종석 잡음 환경에서 진행된 실험을 확인 할 수 있다. 그림 7에서 보면, 두 가지 환경 모두 0dB를 기준으로 기울기가 변화하게 되는 것을 확인 할 수 있다. 백색잡음 환경은 0dB보다 큰 입력에서 완만하게 변화하며, 선박

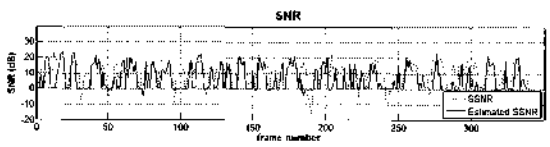

그림 6. 백색 잡음이 추가된 신호의 SNR 추정 선도
Fig. 6. Estimate SNR with White noise condition.

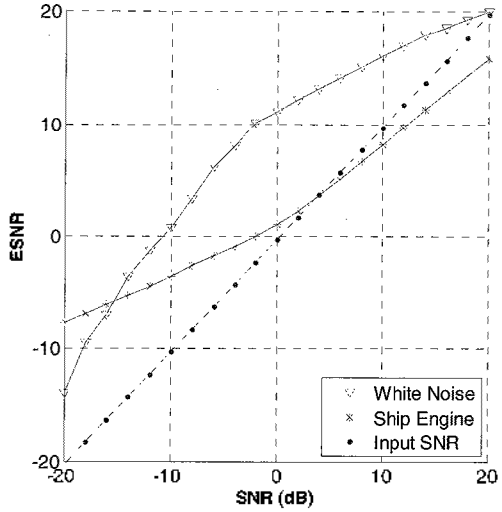

그림 7. SNR 추정 결과 선도
Fig. 7. Experiment Result.

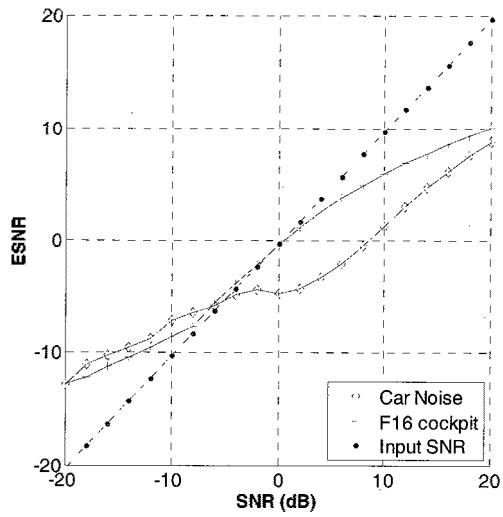

그림 8. SNR 추정 결과 선도
Fig. 8. Experiment Result.

엔진잡음 환경에서는 0 dB보다 낮은 입력에서 점점 큰 오차가 발생되었다. 입력된 SNR과 비교하면 오차가 존재하지만, 파라미터를 조절하면 선형적인 결과를 얻을 수 있겠다.

그림 8에 자동차 내부 잡음 환경에서는 0 dB 부근에서 입력이 변화하지만 출력에는 큰 변화가 없는 결과가 얻어졌다. 하지만 0 dB보다 큰 입력에서는 선형적으로 증가하는 경향을 나타냈다. 전투기 조정석 잡음 환경은 0 dB부근에서는 선형적인 응답이 있었지만, ±10 dB를 초과하는 입력에서는 점점 오차가 커지는 결과가 얻어졌다.

그림 7과 8에서 다양한 잡음 환경을 만들어 제안한 SNR 추정기의 성능을 검증하였다. 입력에 대하여 완벽하게 추정치가 나오지 않았지만, 자동차 내부 환경을 제외하고는 모두 입력이 증가하면 출력도 증가하는 결과를 얻었다.

V. 결론

본 논문에서는 연속음성신호의 잡음량 추정방법을 제안하고 그 성능을 실험하였다. 기존의 잡음추정 알고리즘은 음성신호의 영향을 줄이기 위해 음성신호가 존재하지 않는 구간에서만 잡음양을 계산한다. 이러한 방법은 잡음만 존재하는 구간을 찾아내기 위한 음성구간검출이 필요하다. 그러나 잡음에 열화된 음성신호에 묵음구간이 존재하지 않을 경우, 음성검출을 통한 묵음구간에서의 잡음 추정 방법 및 SNR 추정 방법이 적용될 수 없다. 그리고 신호의 중간에 잡음의 크기가 변화하는 경우에 잡음의 양을 측정하기 어렵다. 또한 잡음에 열화된 음성신호의 경우 무성음과 잡음의 형태가 유사하여 묵음 구간 검출이 어렵다. 하지만 다양한 환경에서 음성신호를 향상하기 위해서는 잡음량을 추정하는 것이 필요하고 중요하다.

본 논문에서는 연속 음성 입력신호에서 음성신호가 잡음에 열화된 경우, 수신된 음성신호 구간에서 SNR을 추정하는 방법을 제안하였다. 수신된 음성신호를 발생원리에 따라 유성음과 무성음 대역으로 구분하고 각 대역의 에너지를 파라미터화 하여 SNR을 추정하도록 하였다. 제안된 SNR 추정기를 이용하면 VAD와 같은 음성구간 검출기를 통하여 묵음구간에만 잡음의 에너지를 계산하는 것이 아니라 연속음성신호에서도 잡음 레벨의 추정이 가능한 장점을 가진다.

제안한 방법의 성능을 확인하기 위해 백색 잡음을 포함하여 다양한 잡음 환경에서 실험하였으며, 잡음의 양의 변화에 따라 대체적으로 선형적인 응답결과가 얻어졌다. 하지만 잡음의 종류에 따라 SNR 추정 정도가 다르게 나타나는 경우가 발생하였다. 이는 잡음으로 간주하는 무성음 대역과 잡음의 시간영역과 주파수영역에서의 형태가 유사하여 잡음과 무성음의 유사성 때문에 발생되는 문제로 판단된다. 대체적으로 선형적인 응답결과가 얻어졌지만 더욱 정확한 SNR 추정을 위해서는 잡음을 분석하여 그 차이점을 구분할 필요가 있다. 그리고 연산량 감소를 위한 최적화 방법이 추후 연구되어야 하겠다.

참 고 문 헌

1. D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Transactions on Acoustics, Speech and Signal processing*, vol. 36, no. 8, 1988.
2. *IMBE VOCODER DESCRIPTION*, Digital Voice System, 1993.
3. M. Kleinschmidt, J. Tchorz, and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Communication*, vol. 34, no. 1–2, pp. 75–91, 2001.
4. A. J. Accardi and R. V. Cox, "A Modular Approach to Speech Enhancement with an Application to Speech Coding," *IEEE ICASSP*, vol. 1, no. 1, pp. 1245, 1999.
5. J. Sohn, N. S. Kim and W. Sung, "A statistical model – based voice activity detector," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan, 1999.
6. 이희원, 장경아, 배명진, "G.723.1 보코더에서 잡음환경에 강인한 음성활동구간 검출기에 관한 연구," 한국음향학회, *한국음향학회지* 21권, 2호, pp. 173–181, 2002.
7. 송영환, 박형우, 배명진, "연속음성신호의 SNR 추정기법에 관한 연구," *한국음향학회지* 제28권 제4호, pp. 1–9, 2009.
8. 배명진, 이상효, *디지탈 음성분석*, 동영출판사, 1998.
9. 김을제, 김형태, 한창문, 배명진, "MBE 부호화용 스펙트럼 V–UV 구간 검출에 관한 연구," 한국음향학회, *학술논문발표회 논문집* 제11권, pp. 43–48, 1992.
10. I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

저자 약력

•박 형 우 (Hyung-Woo Park)

1980년 3월 12일생
2004년 2월: 숭실대학교 정보통신전자공학부 (공학사)
2009년~현재: 숭실대학교 정보통신공학과 (공학석사과정)

•배 명 진 (Myung-Jin Bae)

현재: 숭실대학교 정보통신전자공학부 교수
한국음향학회지 제 26권 제4호 참조