입술정보 및 SFM을 이용한 음성의 음질향상알고리듬*

Speech Enhancement Using Lip Information and SFM

백 성 준** · 김 진 영**
Seong-Joon Baek · Jinyoung Kim

ABSTRACT

In this research, we seek the beginning of the speech and detect the stationary speech region using lip information. Performing running average of the estimated speech signal in the stationary region, we reduce the effect of musical noise which is inherent to the conventional MMSE (Minimum Mean Square Error) speech enhancement algorithm. In addition to it, SFM (Spectral Flatness Measure) is incorporated to reduce the speech signal estimation error due to speaking habit and some lacking lip information. The proposed algorithm with Wiener filtering shows the superior performance to the conventional methods according to MOS (Mean Opinion Score) test.

Keywords: Speech Enahncement, Spectral Flatness Measure

1. 서 론

인간의 기본적인 의사소통 수단은 음성이지만, 실생활에서 인간은 표정, 몸짓, 글자 등 다양한 수단을 사용한다. 특히 말소리의 정확한 이해를 위하여 인간은 무의식적으로 영상정보를 이용한다[1]. 인간이 이렇듯 멀티모달 의사소통을 한다는 것은 멀티모달 정보를 이용하게 되면 음성신호를 처리할 때 도움을 얻을 수 있다는 걸 뜻한다. 특히 잡음에 심하게 오염된 음성의 음질을 개선하고자 할 때, 음성만으로는 원래 음성의 정보를 추출하기가 어려워 큰 폭의 음질 개선이 이루어지지 않는다. 이때 입술 움직임 정보를 읽고 이로부터 원음성의 정보를 추출하는 데 도움을 얻을 수 있다면 음성 정보만을 이용하여 처리하는 경우에 비해 성능 향상을 기대할 수 있을 것이다.

음성을 이용한 음질향상 알고리듬에는 주파수 차감법[2], Wiener 필터를 이용한 방법[3], 선형예측모델을 이용한 방법[4], Kalman 필터를 이용한 방법[5], 그리고 HMM을 이용한 방법 등이 있다[6]. 이 중 본 연구에서는 계산량과 출력음질에 있어서 적절한 균형을 유지하고 있는 Wiener 필터를 이용한 방법을 사용하였으며 이 알고리듬의 성능을 개선하기 위해 입술 정보를 사용하였다.

* 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음.(KRF-2002-003-D00333)
** 전남대학교 공과대학 전자컴퓨터정보통신공학부

입술 정보를 음질 개선 알고리듬에 사용할 때에는 다음 두 가지 점을 주의해야 한다. 첫 번째는 입술정보로부터 음성정보를 얻어내는 데에는 한계가 있다는 점이다. 애초부터 입술정보는 세부적인 음성정보를 전부 보여주지 못하지만, 설령 다 보여준다고 하더라도 모든 음성에 대한 입술 정보를 가지고 있어야 하며, 그것도 한 사람에 대해서만이 아니라 모든 사람에 대해서 입술정보를 가지고 있어야 하므로 현실적으로 입술정보로부터 세부적인 음성정보를 얻어내는 것은 매우 어렵다고 할 수 있다. 두 번째는 입술정보의 편차가 사람마다 꽤 크다는 점이다. 그것은 발성습관이나 화자의 감정상태에 따라 어떤 사람은 입술을 크게 벌리며 발음하고 또 다른 사람들은 입술을 아주 조금씩만 움직이면서 발음하는 것에 기인한다. 때문에 본 연구에서는 입술의 절대적인 크기나 모양에 따라 음성정보를 유추하기보다는 상대적인 움직임으로부터 정보를 읽어 이것을 기존의 음질향상 알고리듬에 결합하여 사용하였다.

2. 제안된 방법

제안된 방법은 크게 입술 정보를 어떻게 이용하는가와 음성신호를 어떻게 필터링할 것인가 하는 두 개의 부분으로 나눌 수 있다. 먼저 입술 정보를 어떻게 처리했는지 살펴본 다음 음성 신호의 필터링은 어떻게 했는지 보도록 한다.

2.1 입술정보의 이용

Wiener필터를 이용하고자 할 때 잡음신호의 정확한 추정은 매우 중요하다. 이를 위해서는 정확한 음성검출이 요구되는데, 잡음이 심한 경우 음성정보만을 이용하여 잡음구간을 검출하는 데에는 한계가 있다. 따라서 본 연구에서는 음성구간 검출을 위해 입술정보를 사용하였으며, 이를 위해 특별히 센서를 부착한 카메라를 이용하였다. 입술 정보는 입술 모델에 근거하여 높이, 폭, 안, 바깥입술경계, 윤곽의 파라미터를 이용하는 방법[7]을 사용하였다. 그림 1에는 많이 사용되는 파라미터를 나타내었다.

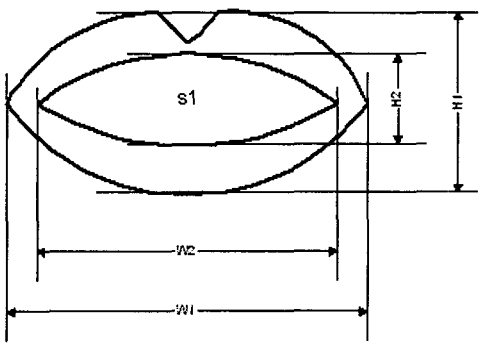

그림 1. 입술 파라미터

이 파라미터들을 이용할 경우 시작점과 끝점의 검출은 매우 쉬운데, 그림에서 안쪽 입술의

높이를 나타내는 파라미터 H2가 0에 가까워지면 입술이 닫힌 것으로 판단할 수 있으므로 H2가 임계치보다 작은 경우에 시작점 혹은 끝점으로 판단하였다. 하지만 실제 구현에 있어서는 따로 고려해야 할 점이 있는데, 그것은 자음으로 시작되는 어휘들은 발성이 막 시작되는 시점에서 입술의 움직임이 그리 크지 않다는 점이다. 이것을 보정하기 위해 H2를 이용하여 검출된 지점으로부터 약 66 msec 정도의 여유 구간을 두어 시작점을 검출하였다. 끝점의 경우에도 같은 이유로 여유를 두어 처리하였다.

Wiener필터를 이용할 경우에 음성신호성분에 대한 예측 성능 개선을 위해 한 프레임 내부의 데이터뿐만 아니라 그 전 프레임의 데이터도 같이 이용하는 방법을 사용할 수 있다. 이를 위해서는 매 분석 프레임마다 음성신호가 얼마만큼 변했는지 측정할 수 있어야 한다. 입술 파라미터 중 H2와 W2는 특히 모음이 발성될 때, 즉 입술의 움직임이 순간적으로 완만해질 때 크게 변하지 않는다. 따라서 이 파라미터를 이용하여 H2와 W2의 유클리드 거리와 그 크기의 비가 임계치보다 적을 경우에 동일한 음성이 발화되는 것으로, 즉 분석 프레임 사이에 음성신호가 많이 변화하지 않은 것으로 판단한다. 이 두 조건을 식으로 나타내면 다음과 같다.

$$H2(i)W2(i)-H2(i-1)W2(i-1)\leq\varepsilon_{hw}$$
$$H2(i)/W2(i)-H2(i-1)/W2(i-1)\leq\varepsilon_{h/w}$$

여기서 첨자 i는 i 번째 프레임을 뜻한다. 이 식을 사용하여 비슷한 음성-특히 모음-이 발성된 프레임으로 판별되는 경우에 음성신호를 추정할 때, 현재 프레임과 그 전 프레임 간의 Running Average를 취함으로써 잡음의 영향을 훨씬 줄일 수 있게 되어 음질향상 알고리듬의 성능을 크게 올릴 수 있다. 세부적인 알고리듬은 다음 절에서 다루기로 한다.

2.2 Wiener 필터링

Wiener 필터링 방법은 음성신호의 Short Time Spectral Amplitude를 추정하여 음성 신호의 음질을 향상시키는 방법이다. 잡음 $n(k)$가 섞인 음성신호 $x(k)$를 식으로 나타내보면 $y(k)=x(k)+n(k)$와 같다. 음성신호와 잡음이 서로 상관관계가 없다고 하면(uncorrelated) 이 신호의 파워스펙트럼은 $S_y(w)=S_x(w)+S_n(w)$로 표시되며, $y(k)$로부터 $x(k)$를 얻기 위한 Wiener 필터는

$$H(w)=\frac{S_x(w)}{S_x(w)+S_n(w)}=\frac{S_y(w)-S_n(w)}{S_y(w)}$$

와 같이 주어진다. 이 중 $S_y(w)$는 관측된 신호로부터 바로 계산이 가능한 값이지만 $S_n(w)$는 그렇지 못하다. 따라서 잡음신호에 대한 스펙트럼은 음성 신호가 없는 구간에 있는 스펙트럼의 평균값을 이용하여 추정한다.

$$S_n(w)=E\,N(w)^2\approx\frac{1}{N}\sum_{i=1}^{N}N_i(w)^2$$

여기서는 음성신호가 없는 구간이 N 개의 프레임이고, $N_i(w)$가 각 프레임의 스펙트럼을 나타내고 있다고 가정하였다. 전달함수 $H(w)$를 $SNR(w)=S_x(w)/S_n(w)$를 이용하여 나타내면 다음 식과 같이 좀더 간편하게 쓸 수 있다.

$$H(w)=\frac{SNR(w)}{1+SNR(w)}$$

이 식에 의하면 특정 주파수 대역에서 SNR이 높으면 $H(w)$가 1에 가까워져서 필터는 관측신호를 전부 통과시키는 반면 SNR이 낮으면 $H(w)$가 0에 가까워져서 관측신호를 통과시키지 않게 된다. 따라서 Wiener 필터는 잡음성분이 큰 주파수 성분은 크게 감쇄시키고, 음성신호성분가 큰 주파수 성분은 그냥 통과시키는 바람직한 특성을 갖는다는 것을 알 수 있다. Wiener 필터 혹은 이의 간단한 구현인 주파수 차감법(Spectral Subtraction)을 사용하게 되면 벌어지는 현상 중 하나가 musical noise의 발생이다[3]. 이러한 musical noise의 완전한 제거는 매우 힘들지만 그 효과를 줄이는 것은 가능한데, 본 연구에서는 분석구간 내부 평활화(intra-frame smoothing) 그리고 분석구간 사이 평활화(inter-frame smoothing)를 이용하여 musical noise의 영향을 줄이고자 하였다.

먼저 분석구간 내부 평활화는 파워스펙트럼을 추정할 때 전체 주파수 대역을 몇 개의 대역으로 나누어 대역별 파워스펙트럼을 구하는 방법을 이용하였다. 이 방법을 사용할 경우에는 파워스펙트럼의 변이가 줄어들어 음성신호가 과소추정되어 musical noise가 생기는 현상을 줄일 수 있다. 이렇게 대역별 분석을 도입할 경우에는 대역을 어떻게 나누느냐가 또 다른 문제가 된다. 균일하게 대역을 나누는 방법과 청각특성을 고려하여 Critical Band나 Mel Band를 사용하는 방법이 가능한데 예비실험에 의하면 균일하게 대역을 나누는 것이 보다 좋은 성능을 보였다. 따라서 본 연구에서는 관측된 신호를 균일한 대역폭을 갖는 12 개의 주파수 대역으로 나누어 분석하였다.

두 번째 문제는 분석구간 사이, 즉 분석 프레임과 분석 프레임 사이의 평활화인데, 앞 절에서 언급한, 입술의 움직임이 상대적으로 적은 구간의 정보를 사용하였다. 하지만 이 입술정보만으로 발화여부와 동일음성발화여부를 전적으로 판단하기는 곤란하다. 그 이유는 발성습관에 따라 많은 사람들이 발화가 끝난 묵음 구간에서도 입을 다물고 있지 않거나 입술을 거의 움직이지 않은 채로 발성하는 일반적이지 않은 상황이 발생하기 때문이다. 이러한 문제에 대응하기 위해서 본 연구에서는 음성신호로부터 측정된 SFM (Spectral Flatness Measure)를 동시에 사용하여 발화여부와 유무성음 판별에 이용하였다. 분석구간의 길이를 N이라고 하고 $y(k)$에 대한 스펙트럼을 $Y(w)$라고 하면 SFM은 다음과 같이 정의된다[8].

$$SFM=10\log_{10}\frac{\left(\prod_{w=1}^{N}|Y(w)|\right)^{1/N}}{\frac{1}{N}\sum_{w=1}^{N}|Y(w)|}$$

다음 그림에는 일반적인 음성신호에 대해 측정된 SFM을 보였다. 이 그림에는 SFM이 가

는 실선으로, 임계값이 가는 직선으로 표시되어 있다. 그림으로부터 SFM값이 유성음과 무성음 사이에 큰 차가 있고, 잡음의 경우에는 SFM값이 특히 0에 가까워진다는 사실을 볼 수 있는데, 이로부터 SFM의 크기가 유무성음 판별은 물론 발화여부를 판단하는 데 도움이 된다는 것은 자명하다(모음 사이에 있는 자음은 유성음화 되는 경우가 있으므로 이를 고려하고 보아야 함)[8].

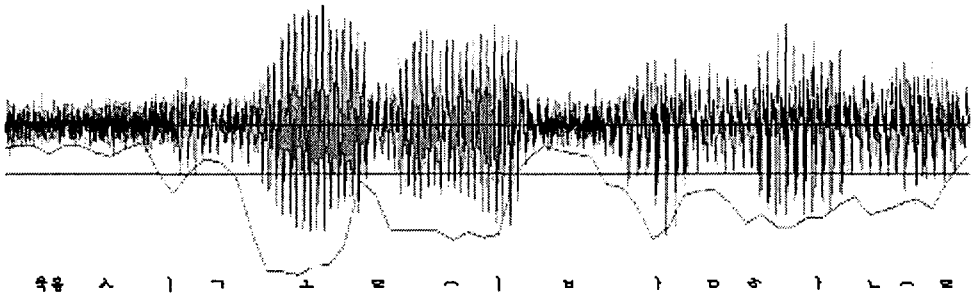

그림 2. 5dB SNR을 갖는 음성신호("시골의 밤하늘")에 대한 SFM 곡선

현재 프레임 i에서 음성파워스펙트럼 $S_x^i(w)$는 관측된 신호 파워스펙트럼 $S_y^i(w)$에서 잡음파워스펙트럼 $S_n(w)$을 빼면 된다. 이때 $S_y^i(w)=|Y_i(w)|^2$이 되는데, 프레임 사이의 평활화를 이용한 Wiener 필터링은 다음과 같은 C 코드로 표현할 수 있다.

```
rCurSx[w]=(rSy[w]=rMagY[w]*rMagY[w])-rSn[w];
if(rCurSx[w]<0) rCurSx[w]=0;
rSNR=(alpha*rPrevSx[w]+(1-alpha)*rCurSx[w])/rSn[w];
rMagX[w]=rMagY[w]*(rSNR/(rSNR+1.0));
rPrevSx[w]=rMagX[w]*rMagX[w];
```

여기서 최종출력, 즉 잡음이 줄어든 음성신호는 추정된 음성신호 스펙트럼의 크기 rMagX[w]와 입력신호의 각도 rAngleY[w]를 이용하여 역푸리에 변환하면 얻을 수 있다. 평활화 인자 alpha는 기존 알고리듬[3]에서는 상수값으로 설정이 되어 있었으나 본 연구에서는 다음 식과 같이 SFM에 따라 변하는 가변치로 설정하였다.

$$\text{alpha}=\begin{cases}0.9, & \text{if } SFM \geq threshold\\ 0.1, & otherwise\end{cases}$$

이 식의 의미는 SFM이 임계값보다 크면 유성음으로 판단하고 프레임 간 평활화를 보다 강하게 하며, 그렇지 않은 경우에는 매우 약하게 평활화를 하겠다는 것이다. 이것은 유성음의 경우가 평활화의 효과를 무성음에 비해 크게 볼 수 있을 것이라는 점에 착안한 것이다. 이때 입술정보로부터 얻은 정상상태에 관한 정보도 동시에 사용하는 데, 두 가지 조건이 동시에 만

족하는 경우에 alpha값을 0.9로 설정하였다. 이것은 예외적인 상황, 즉 소리를 내지 않으면서 입술을 벌리고 있거나, 닫고 있는 경우, 혹은 유성음 경향을 가진 잡음이 들어오는 경우 등에 대처하기 위한 것으로 좀더 Robust한 알고리듬을 만들기 위한 것이다. 예비 실험에 의하면 alpha 값을 상수로 고정시켜 사용하는 경우에 비해 전체적으로 고주파 음이 좀더 살려져 듣기에 보다 나은 소리를 들려준다는 평가를 얻을 수 있었다. 본 논문에서 제안한 알고리듬을 프레임 내부 평활화, 즉 밴드별 분석을 제외하고, 그림 3에 블록도로 나타내었다.

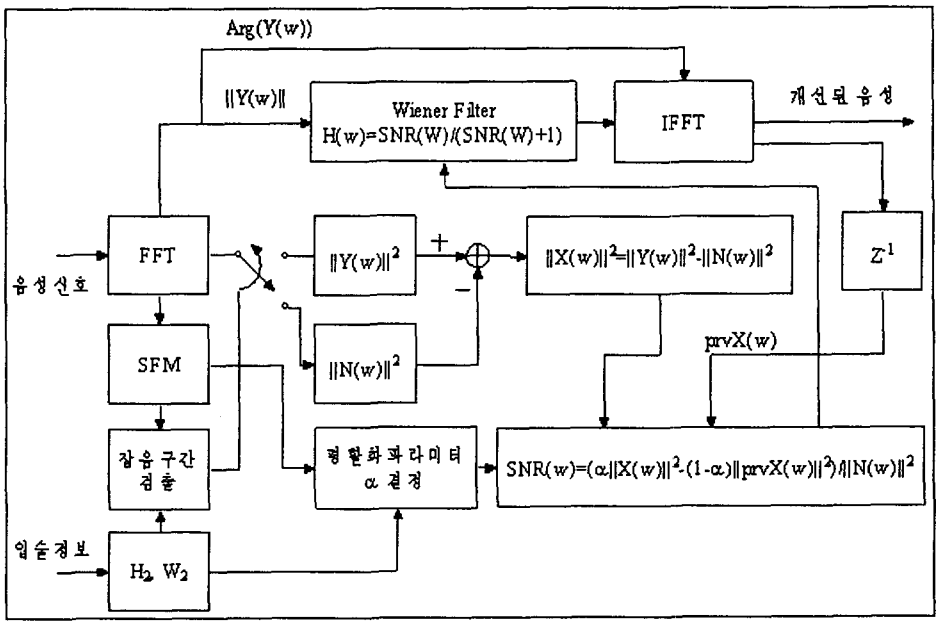

그림 3. 제안된 알고리듬에 대한 블록도

3. 실험 및 결과

제안된 음질 향상 알고리듬의 성능을 평가하기 위해 열악한 환경이라고 할 수 있는 5 dB와 0 dB 신호대 잡음비(SNR)를 가지는 데이터를 남녀 각 1 인에 대해 준비하였다. 시료는 동화책을 낭독한 음성에 백색잡음을 첨가하여 만든 것인데 단문 5 개과 장문 2 개를 각각 준비하였으며, 표본화는 16 비트, 8 kHz로 하였다. 입술정보는 Smarttec사의 Famous Tracker를 이용하여 센서를 화자의 입술 주위에 부착한 후 직접 좌표값을 입력받았다.

음질향상 알고리듬의 성능을 측정하는데, 신호대 잡음비의 향상정도와 같은 객관적인 지표는 큰 도움이 되지 않는다는[9] 점을 미루어, 본 연구에서는 가장 일반적으로 사용되는 주관적 평가 방법이라고 할 수 있는 MOS 테스트 결과를 표 1에 나타내었다. 실험 결과는 화자를 전혀 모르는 청자 20 명에게 헤드폰을 착용하게 하고, 기존 방법과 제안된 방법으로 각각 처리한 단문 5 개와 장문 2 개를 각각 2 회씩 임의로 들려주고, 그로부터 얻은 점수에 대한 평균이다. 실험 결과를 보면 여성과 남성화자 모두에게 제안된 방법이 우수하다는 걸 볼 수 있다.

표 1. MOS 평가 결과

발성화자	SNR	MOS(기존방법)	MOS(제안된 방법)
Male	5dB	3.74	3.98
	0dB	3.51	3.82
Female	5dB	3.62	3.94
	0dB	3.32	3.78

4. 결 론

본 논문에서는 입술정보로부터 음성의 시작점을 찾아내고, 입술 움직임이 상대적으로 적은 구간을 검출하여 이 구간에서는 음성신호가 정상적(stationary)이라고 보고, 분석 프레임 사이의 Running Average를 취함으로써 Musical Noise를 줄이는 방법을 제안하였다. 그와 더불어 입술정보만 이용하여 끝점 및 정상구간을 검출하는 경우 화자의 발성습관에 따라 유발되는 오차를 줄이기 위해 음성신호로부터 SFM을 구하여 이를 입술정보와 결합하여 사용하였다. 제안된 음질향상 알고리듬은 Wiener 필터링을 이용하였는데, MOS 평가법에 따른 실험결과에 의하면 기존의 방법에 비해 좀더 우수한 성능을 보임을 확인할 수 있었다. 하지만 제안된 알고리듬은 SFM을 계산해야 하므로 기존의 알고리듬에 비해 약간의 계산량 증가가 뒤따른다.

참 고 문 헌

[1] Chen, T., H. Peter Graf & K. Wang. 1994. "Speech-assisted video processing: Interpolation and low-bitrate coding." *28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA.

[2] Boll, S. F. 1979. "Suppression of acoustic noise in speech using spectral subtraction." *IEEE Trans. ASSP.*, ASSP-29, 113-120.

[3] Ephraim, Y. & D. Malah. 1984. "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator." *IEEE Trans. on ASSP*, ASSP-32, 1109-1121.

[4] Lim, J. S. & A. V. Oppenheim. 1978. "All-pole modeling of degraded speech." *IEEE Trans. ASSP.*, ASSP-26, 197-210.

[5] Rong, Wen. & P. C. Chen. 1998. "Subband Kalman filtering for speech enhancement." *IEEE Trans., Circuits and Systems*, 45.

[6] Kang, S. K., S. J. Baek, K. Y. Lee. & K.-M. Sung. 2000. "Mixture IMM for speech enhancement under nonstationary noise." *IEEE Trans.. Speech and Audio Processing*, 8, 637-641.

[7] Benoit, C., T. Lallouache, T. Mohamadi. & C. Abry. 1992. "A set of visual French visemes for visual speech synthesis." in *Talking Machines: Theories, Models and Designs*, 485-504.

[8] Yantorno, R. E. 2000. "A study of the spectral autocorrelation peak valley ratio (SAPVR) as a method for identification of usable speech and detection of co-channel speech." *Final Report for Summer Research Faculty Program.*

[9] Grin, L., J. L. Schwartz. & G. Feng. 2001. "Audio-visual enhancement of speech in noise." *JASA*, 109(6), 3007-3020.

접수일자: 2003. 4. 29.
게재결정: 2003. 6. 5.

▲ 백성준
광주광역시 북구 용봉동 300번지 (우: 500-757)
전남대학교 전자컴퓨터정보통신공학부
Tel: +82-62-530-1795 Fax: +82-62-530-1759
E-mail: tozero@chonnam.ac.kr

▲ 김진영
광주광역시 북구 용봉동 300번지 (우: 500-757)
전남대학교 전자컴퓨터정보통신공학부
Tel: +82-62-530-1757 Fax: +82-62-530-1759
E-mail: kimjin@dsp.chonnam.ac.kr