변형된 이득함수를 이용한 잡음 환경에서의 음성인식

진호성*, 이상호**, 홍재근***
*경북대학교 전자전기컴퓨터학부
**(주)삼성전자
***경북대학교 IT대학 전자공학부
e-mail : icarous1@ee.knu.ac.kr

Speech Recognition in Noisy Environments Using Modified Gain Function

Ho-Sung Jin*, Sang-Ho Lee**, Jae-Keun Hong***
*School of Electrical Engineering and Computer Science,
Kyungpook National University
** Samsung Electronics Co., LTD.
*** College of IT Engineering, Kyungpook National University

요 약

본 논문에서는 2단계 잡음제거 방법의 이득함수를 이용한 고조파 복원 잡음제거 방법의 이득함수를 조정하여 기존의 방법보다 음성개선을 향상시켰고, 제안한 방법으로 개선된 음성을 음성인식 기술에 적용하였다. 본 논문에서는 기존 방법으로 음성개선 결과 묵음구간에서 음성구간으로 변화는 구간에서 이전 프레임의 추정된 음성신호로 스펙트럼의 이득함수가 구해져서 음성이 발생하는 구간에서 왜곡이 발생한다. 따라서 본 논문에서는 이러한 현상을 개선시키기 위해 2단계 잡음제거 방법의 이득함수를 추정된 *a priori* SNR과 비교하여 이득함수를 조정하고, 2단계 잡음제거 방법의 이득함수를 고조파 복원 방법의 이득함수와 비교하여 이득함수를 조정하여 음성을 개선하는 방법을 제안하였다. 그리고 음성인식을 위한 특징벡터 추출을 위해 제안한 방법으로 개선된 음성의 대수 에너지를 정규화 하는 대수 에너지 정규화 방법(Log Energy Normalization)을 음성인식 방법에 적용하였다.

1. 서론

음성개선 방법 중에 스펙트럼 향상 방법으로는 스펙트럼 차감법(Spectral Subtraction)[1], Wiener 필터링[2], 최소 평균 자승 오차(MMSE-STSA, Minimum Mean Square Error-Short Time Spectral Amplitude)[3]등이 있다.

2단계 잡음제거(TSNR, Two-step Noise Reduction) 방법은 이득함수와 음성부재를 사용하여 잡음이 섞인 음성으로부터 깨끗한 음성을 추정하는 방법이다. 따라서 전형적인 최소 평균 자승 오차 방법으로 *a posteriori* SNR과 *a priori* SNR의 두 파라메터를 추정하고, *a priori* SNR로 스펙트럼의 이득함수를 구한다. 이렇게 구해진 스펙트럼 이득함수를 *a priori* SNR을 재추정 하는데 사용하고, 재추정된 *a priori* SNR로 다시 스펙트럼 이득함수를 구하고 이것을 잡음이 섞인 음성신호에 곱하여 음성을 개선시킨다.

고조파 복원 잡음제거(HRNR, Harmonic Regeneration Noise Reduction)방법[4]에서 고조파 복원은 2단계 잡음제거 방법으로 개선된 음성신호와 개선된 음성신호의 고조파 복원신호와 콘볼루션하여 고조파가 복원된 신호를 구한다. 2단계 잡음제거 방법으로 구한 스펙트럼 이득함수를 이용하여 고조파 복원된 신호의 스펙트럼의 *a priori* SNR을 추정하고, 추정된 *a priori* SNR로 스펙트럼의 이득함수를 구하여 잡음이 섞인 음성신호에 곱하여 잡음이 섞인 음성을 개선시킨다.

2단계 잡음제거 방법으로 개선된 음성신호를 얻을 때 묵음구간에서 낮은 스펙트럼 이득함수로 인해 음성구간에서도 왜곡이 생겼다. 그리고 2단계 잡음제거 방법의 이득함수가 고조파 복원 잡음제거 방법의 이득함수 추정에 영향을 주어 개선된 음성신호의 크기가 줄어들었다. 이러한 문제점들은 보완하기위해 본 논문에서는 묵음구간에서 음성구간으로 변화는 구간에서 낮은 스펙트럼이득함수를 보완하기위해 2

단계 잡음제거 방법으로 얻은 스펙트럼의 이득함수와 추정된 *a posteriori* SNR을 비교하여 스펙트럼의 이득함수를 구하였고, 고조파 복원 잡음제거 방법의 스펙트럼 이득함수와 2단계 잡음제거 방법의 스펙트럼의 이득함수를 비교하여 고조파 복원 잡음제거 방법의 새로운 스펙트럼 이득함수를 구하였다. 이 결과 묵음구간에서 음성구간으로 변하는 구간에서 음성신호의 왜곡을 보완되었다. 본 논문에서 제안한 방법으로 개선된 음성신호는 묵음구간에서 기존의 방법보다 잡음제거가 덜 되었지만 음성구간에서 왜곡은 기존방법 보다 좋아졌다. 하지만 본 논문에서 제안한 방법도 음성신호의 크기가 줄어드는 결과를 보였다. 본 논문에서는 음성인식의 위해 개선된 음성이 원음성보다 신호의 크기가 작아지는 결과로 인해 대수 에너지의 문제가 발생하여 대수 에너지 정규화 방법을 통하여 음성인식에서 문제 되는 대수 에너지의 문제점을 보안하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 2단계 잡음제거 방법을 이용한 고조파 복원 잡음제거 방법에 대해 소개한다. 3장에서는 변형된 이득함수를 이용한 2단계 잡음제거 방법의 고조파 복원 방법을 제안하고, 4장에서 실험 및 결과를 보인 후 5장에서 결론을 맺는다.

2. 2단계 잡음제거 방법을 이용한 고조파 복원 잡음제거 방법

잡음이 섞인 음성신호 $x(t)$는 깨끗한 음성신호 $s(t)$와 잡음신호 $n(t)$의 합으로 표현된다. 전형적인 *a posteriori* SNR과 *a priori* SNR의 추정은 식 (1)로 나타낸다.

$$SNR_{post}(p,k)=\frac{|X(p,k)|^2}{E[|N(p,k)|^2]}$$
$$SNR_{prio}(p,k)=\frac{E[|S(p,k)|^2]}{E[|N(p,k)|^2]} \quad (2)$$

여기서 $E[.]$는 기대치이고, $S(p,k), N(p,k)$와 $X(p,k)$는 각 신호의 p번째 프레임에서의 k번째 스펙트럼 성분을 나타낸다.

2단계 잡음제거 방법의 *a posteriori* SNR과 *a priori* SNR을 추정은 식 (2)와 같이 표현된다,

$$\hat{SNR}_{post}(p,k)=\frac{|X(p,k)|^2}{\hat{\gamma}_n(p,k)}$$
$$\hat{SNR}_{prio}(p,k)=\beta\frac{|\hat{S}(p-1,k)|^2}{\hat{\gamma}_n(p,k)}+(1-\beta)P\left[\hat{SNR}_{post}(p,k)-1\right] \quad (2)$$

여기서 $P[.]$는 양의 값을 가지기 위한 연산자이고, $\hat{\gamma}_n(p,k)$는 잡음전력스펙트럼 밀도 $E[|N(p,k)|^2]$를 나타낸다. $\hat{S}(p-1,k)$는 이전 프레임에서 추정된 음성신호의 스펙트럼이다. *a priori* SNR을 추정하기 위해 사용된 파레메터 β는 실험을 통하여 β=0.98일 때 가장 좋은 결과를 보였다. 추정된 *a posteriori* SNR과 *a priori* SNR의 두 파라메터를 통하여 얻어진 스펙트럼 이득함수는 식 (3)과 같이 표현된다.

$$G(p,k)=\frac{\hat{SNR}_{prio}(p,k)}{1+\hat{SNR}_{prio}(p,k)} \quad (3)$$

식 (4)는 전형적인 스펙트럼 잡음제거 방법으로 개선된 음성을 추정하는 것을 나타낸다.

$$\hat{S}(p,k)=G(p,k)X(p,k) \quad (4)$$

2단계 잡음제거 방법은 전형적인 스펙트럼 잡음제거 방법에서 사용하는 이득함수로부터 이용하여 *a priori* SNR을 재추정하여 스펙트럼 이득함수를 다시 구하고 이것을 잡음이 섞인 음성신호에 곱함으로써 음성을 개선시킨다. 식 (5)은 2단계 잡음제거 방법의 *a priori* SNR 재추정을 표현한다.

$$\hat{SNR}_{prio_{2step}}(p,k)=\frac{|G(p,k)X(p,k)|^2}{\hat{\gamma}_n(p,k)} \quad (5)$$

식 (5)로부터 재추정된 *a priori* SNR을 이용하여 얻어진 스펙트럼의 이득함수는 식 (6)으로 표현된다.

$$G_{2step}(p,k)=\frac{\hat{SNR}_{prio_{2step}}(p,k)}{1+\hat{SNR}_{prio_{2step}}(p,k)} \quad (6)$$

식 (7)은 2단계 잡음제거 방법으로 개선된 음성을 추정하는 것을 나타낸다.

$$\hat{S}(p,k)=G_{2step}(p,k)X(p,k) \quad (7)$$

고조파 복원은 2단계 잡음제거 방법으로 개선된 음성과 이 음성신호에 대응하는 단위 임펄스신호와의 곱하여 개선된 신호의 고조파 성분을 복원한다. 식 (8)는 고조파 성분 복원을 표현한다.

$$s_{harmo}(t) = NL(\hat{s}(t)) = \hat{s}(t)p(\hat{s}(t))$$
$$p(u) = \begin{cases} 1, & \text{if } u > 0 \\ 0, & \text{if } u < 0 \end{cases} \quad (8)$$

여기서 NL은 비선형 함수를 나타낸다. 식 (9)는 고조파 복원 잡음제거 방법의 *a priori* SNR 추정을 나타낸다.

$$\hat{SNR}_{prio_{HRNR}}(p,k) = \frac{\rho(p,k)|\hat{S}(p,k)|^2 + (1-\rho(p,k))|S_{harmo}(p,k)|^2}{\hat{\gamma_n}(p,k)} \quad (9)$$

여기서 파라메터 $\rho(p,k)$는 2단계 잡음제거 방법을 통하여 개선된 음성에서 고조파의 성분을 복원한 음성과 2단계 잡음제거 방법을 통하여 얻어진 음성과의 관계를 제어하는 파라메터이다. 식 (10)은 고조파 복원 방법의 스펙트럼 이득함수를 표현한다.

$$G_{HRNR}(p,k) = \frac{\hat{SNR}_{prio_{HRNR}}(p,k)}{1 + \hat{SNR}_{prio_{HRNR}}(p,k)} \quad (10)$$

고조파 복원 잡음제거 방법으로 음성을 개선하는 방법은 식 (11)로 표현된다.

$$\hat{S}(p,k) = G_{HRNR}(p,k)X(p,k) \quad (11)$$

3. 변형된 이득함수를 사용한 고조파 복원 방법

기존의 2단계 잡음제거 방법을 사용하는 고조파를 복원하여 음성을 개선하는 방법은 스펙트럼의 이득함수를 *a priori* SNR추정에 계속적으로 사용하여 최종 스펙트럼의 이득함수를 얻게 되어 있다. 여기서 *a priori* SNR을 계속적으로 사용하면서 묵음 구간에서 스펙트럼의 이득함수를 아주 작게 주어서 묵음구간에서 음성구간으로 변화하는 구간에서 음성신호의 왜곡이 발생하는 결과를 보였다. 따라서 묵음구간에서 스펙트럼의 이득함수를 아주 작게 주는 경우를 개선하기 위해 스펙트럼의 이득함수를 이전 단계의 스펙트럼의 이득함수와 비교하여 스펙트럼의 이득함수를 추정하는 방법을 제안하였다.

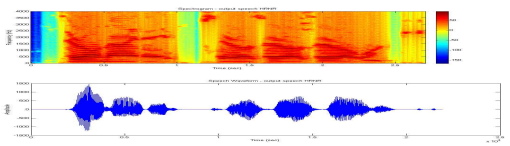
식 (12)는 2단계 잡음제거 방법에서 변형된 스펙트럼 이득함수를 구하는 방법이다.

$$G_{2step}(p,k) = \begin{cases} G_{2step}(p,k) & \text{if } G_{2step}(p,k) < \hat{SNR}_{post} \\ \hat{SNR}_{prio} & \text{otherwise} \end{cases} \quad (12)$$

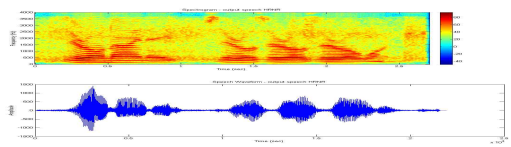
식 (13)은 고조파 복원 잡음제거 방법에서 2단계 잡음제거 방법의 이득함수를 이용하여 고조파 복원 잡음제거 방법의 이득함수를 재추정한다.

$$G_{HRNR}(p,k) = \begin{cases} G_{HRNR}(p,k) & \text{if } G_{HRNR}(p,k) < G_{2step}(p,k) \\ G_{2step}(p,k) & \text{otherwise} \end{cases} \quad (13)$$

그림 1은 기존의 방법과 제안한 방법으로 복원한 음성신호의 스펙트로그램과 파형을 보여준다.



(a) 기존의 HRNR 방법



(b) 제안한 방법

[그림 1] 스펙트로그램과 음성파형

제안한 방법은 음성개선이 향상되었다. 하지만 음성인식 실험 결과는 잡음이 많이 섞인 음성의 경우 인식률이 좋아지는 반면 잡음이 없는 깨끗한 음성에서는 인식률이 조금 떨어지는 결과를 보였다. 따라서 깨끗한 음성에서 인식률 저하를 막기 위해 제안한 방법으로 개선된 음성신호의 대수 에너지 정규화 방법[5]을 제안하였다. 대수 에너지 정규화 방법은 음성신호의 n개의 프레임에서 에너지의 최댓값과 최솟값을 찾고, 대상(target)의 최솟값을 구하여, 대상의 최솟값보다 작으면 모든 프레임에 대해서 대수 에너지를 정규화 하는 방법이다. 따라서 개선된 음성신호의 대수 에너지의 최솟값과 최댓값에 따라 신호의 대수 에너지가 동적인 범위를 가지는 대수 에너지 정규화 방법을 사용하였다. 이러한 대수 에너

지 정규화 방법을 통하여 깨끗한 음성에서도 인식률이 향상 되는 결과를 보였다.

본 논문에서는 음성의 왜곡을 최소화하면서 음성개선을 위해 변형된 스펙트럼의 이득함수를 갖도록 스펙트럼 이득함수를 조정하고, 음성개선 후 에너지의 문제로 깨끗한 음성에서 인식률이 저하 되는 것을 보완하기 위해 대수 에너지 정규화 방법을 사용하였다.

4. 실험 및 결과

본 논문에서는 음성인식 성능의 평가를 위해 European Telecommunications Standards Institute (ETSI) STQ-AURORA DST Working Group[6]에서 제시한 AURORA2 데이터베이스를 사용하고, 기본 인식 시스템은 AURORA2 에서 제공하는 AURORA2-HTK를 사용하였다. 테스트 음성으로 AURORA2 데이터베이스에서 제공되는 set A를 사용하였다. 표 1은 인식방법에서 많이 사용되는 MFCC방법으로 특징을 추출하여 인식한 결과이고, 표 2는 2단계 잡음제거 방법의 이득함수를 사용한 고조파 복원 잡음제거 방법으로 개선된 음성을 MFCC방법으로 특징을 추출하여 인식한 결과를 보여준다. 그리고 표 3은 제안한 방법으로 인식한 결과를 보여준다.

[표 1] 잡음제거를 하지 않는 인식률 (%)

	Subway	Babble	Car	Exhibition	Avg.
Clean	99.93	99.00	98.96	99.20	99.27
20dB	97.05	90.15	97.41	96.39	95.25
15dB	93.49	73.76	90.04	92.04	87.33
10dB	78.72	49.43	67.01	75.66	67.71
5dB	52.16	26.81	34.09	44.83	39.47
0dB	26.01	9.28	14.46	18.05	16.95
-5dB	11.18	1.57	9.36	9.60	7.93
Avg.	65.51	50.00	58.76	62.25	59.13

[표 2] HRNR 방법의 인식률 (%)

	Subway	Babble	Car	Exhibition	Avg.
Clean	93.74	92.53	93.65	93.92	93.46
20dB	85.35	81.05	92.01	80.16	84.64
15dB	79.00	74.94	88.73	73.03	78.93
10dB	68.93	61.19	82.43	60.63	68.30
5dB	52.75	44.07	69.76	42.27	52.21
0dB	32.67	21.43	50.52	22.59	31.80
-5dB	17.38	9.95	26.78	9.32	15.86
Avg.	61.40	55.02	71.98	54.56	60.74

[표 3] 제안한 방법의 인식률 (%)

	Subway	Babble	Car	Exhibition	Avg.
Clean	98.04	97.70	97.46	97.44	97.66
20dB	96.16	94.72	97.08	94.29	95.56
15dB	94.01	91.02	95.65	92.29	93.24
10dB	87.84	79.87	90.31	84.45	85.62
5dB	73.23	57.56	73.04	66.55	67.60
0dB	50.02	31.38	38.53	36.01	38.99
-5dB	22.08	13.36	15.69	14.96	16.52
Avg.	74.48	66.52	72.69	69.43	70.74

표 1과 표 3을 비교하면 깨끗한 음성일 경우는 인식률이 조금 낮아지나, 신호대잡음비가 작아질수록 제안한 방법이 우수한 성능을 나타내는 것을 확인할 수 있다. 그리고 표 2와 표 3을 비교하면 기존의 방법 보다 전체적인 인식률에서 성능이 향상되었다.

5. 결론

본 논문에서는 잡음환경에서 음성의 개선을 목적으로 기존의 고조파 복원을 이용한 잡음제거 방법에서 사용 하였던 2단계 잡음제거 방법의 이득함수를 추정된 *a priori* SNR과 비교하여 이득함수를 조정하고, 2단계 잡음제거 방법의 이득함수를 고조파 복원 방법의 이득함수와 비교하여 이득함수를 조정하여 음성을 개선하는 방법을 제안하였다. 그리고 음성인식을 위한 특징벡터 추출을 위해 제안한 방법으로 개선된 음성의 대수 에너지를 정규화 하여 음성개선을 통해 왜곡된 대수 에너지 값을 정규화 하여 인식률에 있어서 성능이 향상되는 결과를 보였다.

참고문헌

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoustic Speech Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1980.

[2] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a MMSE short-time spectral amplitude estimator,” *IEEE Trans. on Acoustic Speech Signal Processing,* vol. 32, no. 6, Dec. 1984.

[4] C. Plapous, C. Marro, and P. Scalart,

"Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 12, no. 6, pp. 2098-2108, Nov. 2006.

[5] W. Zhu and D. O'Shaughnessy, "Log-energy dynamic range normalization for robust speech recognition," *IEEE International Conference on Acoustics Speech Signal Processing 2005,* pp. 245-248, 2005.

[6] ETSI standard document, "Speech processing, transmission and quality aspects(STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm," *ETSI ES 201 108 v1.1.1*