논문 16-41-10-06 The Journal of Korean Institute of Communications and Information Sciences '16-10 Vol.41 No.10
http://dx.doi.org/10.7840/kics.2016.41.10.1183

통계적 스펙트럼 이퀄라이저를 이용한 저 비트율 음성부호화기의 명료도 향상

이 정 훈•, 윤 덕 규*, 최 승 호°

Intelligibility Improvement of Low Bit-Rate Speech Coder Using Stochastic Spectral Equalizer

Jeong Hun Lee•, Deokgyu Yun*,
Seung Ho Choi°

요 약

디지털 음성통신에서의 저 비트율 음성부호화기는 음성발성모델의 파라미터를 사용하여 음성을 합성한다. 이 경우, 파라미터에 할당된 비트가 매우 한정적이기 때문에 합성된 음성의 스펙트럼이 크게 왜곡될 수 있으며, 이는 명료도 저하의 요인이 된다. 본 논문에서는 통계적 스펙트럼 이퀄라이저를 이용한 명료도 향상 기법을 제안한다. 본 기법은 각각의 음성부호화기별로 원음과 합성음의 스펙트럼 비율을 이용하여 통계적으로 가중치 벡터를 구하며, 이를 합성 음성에 적용한다. 객관적인 음성명료도 평가 실험을 통해, 제안한 기법이 기존의 방법보다 성능이 우수함을 확인하였다.

Key Words : Stochastic spectral equalizer, Low bit-rate speech coder, Speech intelligibility

ABSTRACT

Low bit-rate speech coder in digital speech communications synthesizes speech using vocal tract model parameters. In this case, the spectra of the synthesized speech can be much distorted since the allocated bits for the parameters are considerably limited, which results in the degradation of speech intelligibility. In this paper, we propose a speech intelligibility improvement method using stochastic spectral equalizer. This method stochastically obtains the weight vector of each speech coder using spectral ratios between original and synthesized speech, then applies this weight vector to synthesized speech. From the experiments of objective speech intelligibility tests, we found that the performance of the proposed method is better than that of the conventional method.

Ⅰ. 서 론

디지털 음성통신을 위한 대부분의 저 비트율 음성부호화기는 성도 필터(vocal tract filter)와 여기 신호(excitation signal)를 기반으로 하는 음성발성모델을 이용하며, 음성의 파라미터를 적은 수의 비트로 표현한다. 따라서 합성음의 스펙트럼이 크게 왜곡 될 수 있으며, 이는 음성 명료도를 저하시킨다.

명료도 향상을 위해 저 비트율 음성부호화기의 후처리 필터(postfilter)에 대해 많은 연구가 진행되어 왔다[1-3]. 이 방법은 음성 피치의 하모닉 성분을 강화하는 장구간(long-term) 후처리 필터, 포먼트 부분을 강화하는 단구간(short-term) 후처리 필터, 스펙트럼의 기울기 보상 등이 있다[1]. 대표적인 포먼트 강화 방법은 LPC(Linear Predictive Coding) 계수를 변경하거나, LSF(Line Spectral Frequency) 위치를 조절하는 것이다[2].

본 논문에서는 명료도 저하 현상을 해결하기 위하여, 각각의 음성부호화기별로 원음과 합성음의 스펙트럼 비를 이용하여 통계적으로 가중치 벡터를 구하며, 이를 합성 음성에 적용하는 새로운 방법을 제안한다.

Ⅱ. 통계적 스펙트럼 이퀄라이저

그림 1과 같이 각각의 음성부호화기에 대한 입력

※ 본 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.
• First Author : Department of Electronic Engineering, Seoul National University of Science and Technology, antmfdl043@naver.com, 학생회원
° Corresponding Author : Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, shchoi@seoultech.ac.kr, 정회원
* Department of Electronic Engineering, Seoul National University of Science and Technology, deokkyuyun@gmail.com
논문번호 : KICS2016-09-273, Received September 26, 2016; Revised October 18, 2016; Accepted October 18, 2016

음성 $x(n)$과 복호화기의 합성음 $y(n)$의 스펙트럼 비를 식 (1)과 같이 프레임 별로 계산한 후, 아래의 식 (2)와 같이 주파수 밴드별로 기하평균하여 수식 (3)의 가중치 벡터 $\vec{w}_{ob}$를 구한다.

$$r(i,l)=\frac{B_X(i,l)}{B_Y(i,l)} \tag{1}$$

$$w_{ob}(l)=(\prod_{i=1}^{M} r(i,l))^{\frac{1}{M}}, l=1,2,\cdots,L \tag{2}$$

$$\vec{w}_{ob}=[w_{ob}(1),w_{ob}(2),\cdots,w_{ob}(L)]^T \tag{3}$$

여기에서 i,l,M는 각각 프레임 인덱스, 1/3 옥타브 밴드 인덱스, 학습용 음성 데이터베이스의 전체 프레임 개수이다. 본 연구에서는 1/3 옥타브 스케일로 주파수 밴드를 구성하였으며, 총 밴드 개수 L은 20개, 각 밴드의 범위는 아래 표와 같다.

학습 과정에서 구한 1/3 옥타브 밴드별 가중치 벡터 $\vec{w}_{ob}$는 표 1을 기반으로 아래 식 (4)의 주파수 인덱스별 가중치 벡터 $\vec{w}_{fb}$로 변환하고, 아래 식 (5)와 같이 테스트 합성음의 스펙트럼에 적용한다.

$$\vec{w}_{fb}=[w_{fb}(1),w_{fb}(2),\cdots,w_{fb}(K)]^T \tag{4}$$

$$\vec{Y}_{eq}(i)=\vec{w}_{fb}^{\,T}\,\vec{Y}(i) \tag{5}$$

여기에서 K는 DFT(Discrete Fourier Transform) 포인트 개수이다. 이후 가중치 벡터를 적용한 스펙트럼 $\vec{Y}_{eq}(i)$를 IDFT(Inverse Discrete Fourier Transform) 한 후, 음성 프레임을 overlap add하여 최종적으로 $y_{eq}(n)$을 합성한다.

표 1. 주파수 밴드 별 주파수 범위
Table 1. Frequency range of each band

Frequency Band	Frequency Range(Hz)	Frequency Band	Frequency Range(Hz)
1	1-50	11	400-504
2	50-63	12	504-635
3	63-79	13	635-800
4	79-100	14	800-1008
5	100-126	15	1008-1270
6	126-159	16	1270-1600
7	159-200	17	1600-2016
8	200-252	18	2016-2540
9	252-317	19	2540-3200
10	317-400	20	3200-4000

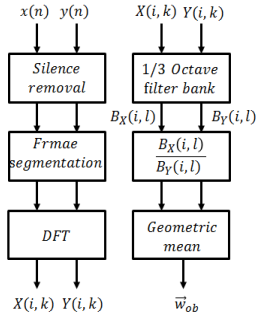

그림 1. 주파수 밴드별 가중치 계산 블록도
Fig. 1. Block diagram of calculating frequency band weights

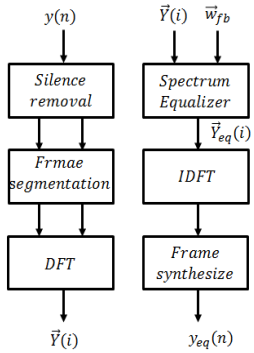

그림 2. 통계적 스펙트럼 이퀄라이저 블록도
Fig. 2. Block diagram of stochastic spectral Equalizer

Ⅲ. 실험 및 결과

본 논문에서는 음성명료도 평가 실험을 위해 객관적 음성명료도 평가 방법인 LSD(Log Spectral Distance)와 MIKNN(Mutual Information based on K-Nearest Neighbor)[3]을 사용하였다. LSD는 아래의 식 (6)과 같이 두 신호간의 스펙트럼 차이를 log scale로 계산한 것이다.

$$LSD=\frac{1}{M}\sum_{i=1}^{M}\sqrt{(\frac{1}{K}\sum_{k=1}^{N}(10\log(\frac{|Y(i,k)|^2}{|X(i,k)|^2}))^2} \tag{6}$$

MIKNN은 상호정보(Mutual Information) 기반의 음성명료도 측정방법으로서, 프레임별로 1/3 옥타브 필터뱅크를 통과 후에 K-NN을 도입하여 백분율의 점수로 환산하는 것이다.

실험에 사용한 음성 데이터베이스는 총 96개의 한국어 발성음으로 구성되며, 이중 64개를 학습에 32개

를 테스트에 사용하였다. 비교 대상은 기존의 대표적인 음성명료도 향상기법인 포먼트 강화 방법이다. 평가 대상으로 선정한 부호화기는 모두 전송률이 4.8 kbps이하이며, CELP 형 저 비트율 음성부호화기 FS-1016[4]과 LPC기반의 MELP[5,6], LPC-10e[7]이다.

표 2와 표 3은 각각 FS-1016, MELP, LPC-10e에 대한 LSD와 MIKNN 결과이다. 표 2와 표 3과 같이 제안한 방법이 원음과 처리 음성간의 스펙트럼 차이가 더 작고 MIKNN 명료도 점수는 더 높은 결과를 보임을 알 수 있다. 포먼트 강화 기법의 경우, 스펙트럼의 왜곡으로 인하여 객관적인 명료도 측정으로는 성능이 감소되었다.

표 2. 원음과 처리된 음성간의 LSD 결과
Table 2. LSD between original speech and processed speech

Method / Vocoder	synthesized speech	Formant enhancement	Stochastic spectral equalizer
FS-1016 (4.8kbps)	9.20	10.12	8.81
MELP (2.4kbps)	9.38	9.89	9.25
LPC10e (2.4kbps)	13.11	13.04	12.55

표 3. 원음과 처리된 음성간의 MIKNN 결과
Table 3. MIKNN Results between original speech and processed speech

Method / Vocoder	synthesized speech	Formant enhancement	Stochastic spectral equalizer
FS-1016 (4.8kbps)	39.13	33.41	39.38
MELP (2.4kbps)	33.85	30.09	34.08
LPC10e (2.4kbps)	26.29	24.46	27.33

Ⅳ. 결론 및 향후 연구방향

본 논문은 저 비트율 음성부호화기의 스펙트럼 왜곡에 의한 명료도 저하 문제를 극복하기 위한 통계적 스펙트럼 이퀄라이저를 제안하였다. 객관적인 음성명료도 평가 실험을 통해, 제안한 기법이 기존의 방법보다 성능이 우수함을 확인하였다.

향후 음성 데이터베이스를 확장하고, 딥러닝 기반의 스펙트럼 이퀄라이저를 연구할 계획이다. 또한, 객관적 평가와 함께 주관적 청취평가를 병행할 계획이다.

References

[1] J.-H. Chen and A. Gersho, “Adaptive postfiltering for quality enhancement of coded Speech,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 59-71, Aug. 1995.

[2] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, “Comparison of formant enhancement methods for HMM-Based speech synthesis,” *SSW*, pp. 334-339, Sep. 2010.

[3] J. Jensen and C. H. Taal, “Speech intelligibility prediction based on mutual information,” *IEEE/ACM TASLP*, vol. 22, no. 2, pp. 430-440, Feb. 2014.

[4] J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, “The federal standard 1016 4800 bps CELP voice coder,” *Digital Signal Process*, vol. 1, no. 3, pp. 145-155, 1991.

[5] Alan McCree, et al., “A 2.4 kbit/s MELP coder candidate for the new US Federal Standard,” *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on*, vol. 1, pp. 200-203, May. 1996.

[6] Y. Chun and B. Jun, “An enhanced MELP vocoder in noise environments,” *The Journal of Korean Institute of Communications and Information Sciences*, vol. 28, no. 1, pp. 81-89, 2003.

[7] T. E. Tremain, “The government standard linear predictive coding algorithm: LPC-10,” *Speech Technology*, vol. 1, no. 2, pp. 40-49, 1982.