논문 20-45-05-05 The Journal of Korean Institute of Communications and Information Sciences '20-05 Vol.45 No.05
https://doi.org/10.7840/kics.2020.45.5.798

자기부호화기를 이용한 비트 열 기반 변형 MELP 보코더

이 한 나•, 윤 덕 규*, 최 승 호°

A Bit-Stream-Based Modified MELP Vocoder Using Autoencoder

Hannah Lee•, Deokgyu Yun*,
Seung Ho Choi°

요 약

본 논문에서는 자기부호화기를 이용하여 비트 열을 기반으로 MELP 보코더를 변형하는 방법을 제안한다. 본 기법은 기존 보코더의 인코더에서 추출된 비트열을 입출력으로 하는 자기부호화기를 훈련하며, 훈련된 자기부호화기는 변형 보코더의 인코더와 디코더 모델로 나뉜다. 또한 자기부호화기의 병목특징은 스칼라 양자화 한 뒤 비트열로 송신된다. 객관적인 음성명료도 평가 실험을 통해, 제안한 방법이 기존 보코더와 성능이 유사함을 보인다.

Key Words : Bit-stream, Modified vocoder, MELP, Autoencoder, Bottleneck feature

ABSTRACT

In this paper, we propose a bit-stream-based modified MELP vocoder using autoencoder. This technique trains an autoencoder using the bit-streams extracted from the encoder of an existing vocoder, and the trained autoencoder is divided into an encoder and a decoder model of the modified vocoder. In addition, the bottleneck features of the autoencoder are scalar quantized and transmitted as bit-streams. Through objective speech intelligibility evaluation experiments, we show that the performance of the proposed method is similar to the existing vocoder.

Ⅰ. 서 론

디지털 음성통신을 위한 저 비트율 보코더에는 LPC-10e(linear predictive coding(enhanced) with 10 predictive coefficients)[1], MELP(mixed-exited linear prediction)[2], CELP(code-excited linear prediction)[3] 등이 존재한다. 이러한 보코더를 변형하는 방법에 대해서 많은 연구가 진행되어 왔다. 기존 연구에서는 개선된 CELP 합성기를 통해 복잡도를 감소시키는 방법[4], 음성, 비음성, 묵음 구간마다 각각 다른 비트 전송률을 가져 보코더의 평균 비트 전송률을 낮추는 방법[5], 그리고 암호비트를 이용하여 보코더의 코딩 내에 정보를 숨기는 방법[6] 등이 제안되었다. 최근 들어 전 세계적으로 심층신경망에 대한 관심이 높아지며 딥러닝 기반 변형 보코더[7-8]에 대한 연구도 활발히 진행 중이다.

본 논문에서는 MELP 보코더를 변형하기 위해 기존 MELP 보코더의 비트 열을 입출력으로 사용해 자기부호화기를 훈련한 후, 훈련된 자기부호화기를 인코더와 디코더로 나누고, 자기부호화기의 병목특징을 스칼라 양자화 한 뒤 비트열로 송신하는 방법을 제안한다. 서론에 이어 II장에서는 제안된 변형 MELP 보코더 방법이 서술한다. III장에서는 제안된 방법의 성능을 객관적 음성명료도 평가 방법으로 실험한 내용과 결과를 보이며 마지막으로 IV장에서 결론과 향후 연구방향을 제시한다.

Ⅱ. 자기부호화기 기반 변형된 MELP 보코더

자기부호화기를 훈련 시 기존 1.2 kbps MELP 인코더에서 출력된 양자화 된 비트열을 입출력으로 사용했다. 이 때 1.2 kbps MELP는 22.5 msec마다 1 프레임을 획득하며 특징 값은 총 27 비트이다. 훈련된 자기부호화기의 인코더와 디코더 모델은 각각 1.2 kbps MELP의 인코더, 디코더와 결합되어 변형보코

※ 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, hannha9656@naver.com, 학생(석사), 학생회원
° Corresponding Author : Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, shchoi@seoultech.ac.kr, 교수, 정회원
* Department of Electronic Engineering, Seoul National University of Science and Technology, deokkyuyun@gmail.com, 학생(박사)
논문번호 : 202002-030-A-LU, Received February 16, 2020; Revised February 21, 2020; Accepted February 24, 2020

더의 인코더, 디코더를 구성한다.

변형보코더의 인코더 상세 구조는 그림 1과 같다. 음성 샘플이 1.2 kbps MELP 인코더를 통과하면 프레임 당 양자화 된 27 비트를 얻는다. 이 비트열을 자기부호화기의 인코더에 입력하여 13차 병목특징을 추출한다. 본 연구에서는 실제 디지털 통신을 위해 추출된 병목특징을 선형적으로 4 비트 스칼라 양자화 하였다. 최종적으로 변형보코더의 인코더는 표준인 2.4 kbps MELP 보코더와 동일하게 프레임 당 54차 비트열을 전송한다.

해당 정보가 변형보코더의 디코더에 수신되었을 때의 동작은 그림 2와 같다. 프레임 당 수신되는 54차 비트열은 13차 벡터로 역 양자화 된 뒤 자기부호화기의 디코더에 입력된다. 자기부호화기의 디코더에서 출력된 27차 벡터는 추정된 비트열이다. 해당 벡터에서 특정 임계값보다 큰 값을 1, 작은 값을 0으로 치환한 뒤 변형보코더의 디코더 내부의 1.2 kbps MELP 디코더에 입력하여 합성된 음성을 얻는다.

본 실험에서 훈련된 자기부호화기는 인코더와 디코더가 각각 3개의 층으로 구성되었고 입력을 제외한 인코더 모델의 노드수는 256-128-13, 디코더 모델의 노드수는 128-256-27이다. 활성화 함수로 각 모델의 마지막 층에 softsign, 다른 층에는 hard sigmoid가 사용된다.

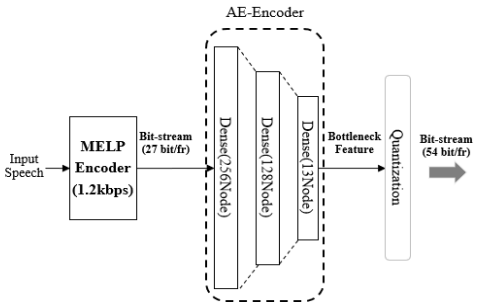

그림 1. 변형 인코더의 구조
Fig. 1. Structure of modified encoder

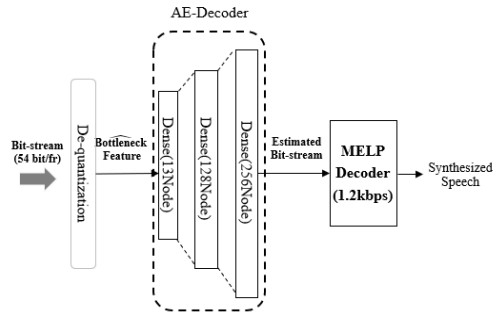

그림 2. 변형 디코더의 구조
Fig. 2. Structure of modified decoder

Ⅲ. 실험 및 결과

본 논문에서는 합성된 음성 신호를 평가하기 위해 객관적 음성명료도 평가 방법인 LSD(log spectral distance)와 표준 침입적(intrusive) 음성명료도 추정 방법인 STOI(short-term objective intelligibility measure)[9]를 사용하였다. LSD는 아래의 식 (1)과 같이 두 신호간의 스펙트럼 차이를 log scale로 계산한 것이다. 식 (1)에서 K는 이산푸리에변환의 차수이며 N은 총 프레임의 수이다.

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

실험에 사용한 음성 데이터베이스는 TIMIT[10] 데이터베이스와 NTT 한국어 발성음이다. 자기부호화기 훈련에 TIMIT 데이터베이스 중 총 6,334개의 문장을 사용하였으며 평가에 32개 한국어 문장을 사용하였다.

표 1은 기존 MELP 보코더로 합성한 음원과 변형보코더로 합성한 음원을 각각 원래의 음성샘플과 비교했을 때의 LSD와 STOI 결과이다. 인코더와 디코더가 모두 기존 2.4 kbps MELP일 때 음질이 가장 좋다. 변형 보코더는 2.4 kbps MELP 보코더와 비트 전송률이 동일하지만 음성 합성에 1.2 kbps MELP를 사용하기 때문에 성능이 감소하였다. 변형보코더로 합성된 음성의 음질을 1.2 kbps MELP와 비교하면 양자화 오류로 인한 약간의 명료도 저하가 발생하나, 비슷한 성능을 보임을 알 수 있다.

표 1. 합성된 음성과 원음의 LSD 및 STOI 결과
Table 1. LSD and STOI results between original speech and synthesized speech

Measure \ Vocoder	MELP 1.2 kbps Encoder& Decoder	MELP 2.4 kbps Encoder & Decoder	Modified MELP Encoder & Decoder
LSD	12.39	9.17	12.49
STOI	0.82	0.88	0.80

Ⅳ. 결론 및 향후 연구방향

본 논문에서는 자기부호화기를 이용하여 비트 열 기반으로 MELP 보코더를 변형하는 방법을 제안했다.

기존 보코더의 인코더에서 추출된 비트열을 입출력으로 하는 자기부호화기를 훈련했으며, 훈련된 자기부호화기는 인코더와 디코더로 나뉘었다. 또 비트열 송신을 위해 자기부호화기의 병목특징을 스칼라 양자화하였다. 객관적인 음성명료도 평가 실험을 통해, 제안한 방법이 기존 1.2kbps MELP 보코더와 유사한 성능을 보임을 확인했다. 향후 자기부호화기 병목특징의 차수를 변경하여 비트율을 가변화하는 가변 비트율 부호화(variable bit-rate coding)에 대한 연구를 할 계획이다. 또한, MELP 외에 다른 보코더로도 실험을 확장 할 계획이다.

References

[1] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29-32, Feb. 1988.

[2] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new federal standard at 2400 bps," *IEEE Int. Conf. Acoustics, Speech, and Sign. Process.*, pp. 1591-1594, Apr. 1997.

[3] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," *IEEE Int. Conf. Acoustics, Speech, and Sign. Process.*, pp. 937-940, Apr. 1985.

[4] L. M. de Silva and A. Alcaim, "A modified CELP model with computationally efficient adaptive codebook search," *IEEE Sign. Process. Lett.*, vol. 2, no. 3, pp. 44-45, Mar. 1995.

[5] L. Zhang, T. Wang, and V. Cuperman, "A CELP variable rate speech codec with low average rate," *IEEE Int. Conf. Acoustics, Speech, and Sign. Process.*, pp. 735-738, vol. 2, Apr. 1997.

[6] J. Liu, Z. Lu, and H. Luo, "A CELP-Speech information hiding algorithm based on vector quantization," *Fifth Int. Conf. Inf. Assurance and Secur.*, pp. 75-78, Aug. 2009.

[7] J. Skoglund and V. Jean-Marc, "Improving opus low bit rate quality with neural speech synthesis," *arXiv:1905.04628v2* [eess.AS] 21, Nov. 2019.

[8] J. H. Lee, H. K. Kim, and S. K. Kim, "Vocoder parameter reduction based on deep learning auto-encoder," in *The Korean Soc. Speech Sci.*, p. 131, Seoul, Korea, Jun. 2019.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *IEEE Int. Conf. Acoustics, Speech and Sign. Process.*, pp. 4214-4217, Mar. 2010.

[10] J. S Garofolo, et al., "*DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*," NASA STI/Recon Technical Report N, 93, Feb. 1993.