논문 17-42-09-04 The Journal of Korean Institute of Communications and Information Sciences '17-09 Vol.42 No.09
https://doi.org/10.7840/kics.2017.42.9.1736

장단기 메모리 순환신경망 기반의 비침입적 음성 명료도 추정 방법

윤 덕 규◆, 이 한 나*, 최 승 호°

A Non-Intrusive Speech Intelligibility Estimation Method Based on Recurrent Neural Network with Long Short-Term Memory

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

요 약

본 논문은 기준 음성신호가 없는 비침입적 (non-intrusive) 음성 명료도 추정 방법에 관한 것이며, 장단기 메모리 (long short-term memory, LSTM) 구조의 순환신경망 (recurrent neural network, RNN) 기반의 방식을 제안한다. 기존의 표준 명료도 추정 방법인 P.563은 추정 성능이 미흡하고 특히, 다양한 잡음과 잔향 환경에서 일관성이 부족하다. 제안한 방법은 기준 음성신호가 있는 침입적 (intrusive) 표준 음성 명료도 평가 도구인 STOI를 활용하여 신경망의 모델 파라미터를 훈련한다. 신경망의 입력과 출력은 MFCC 벡터와 프레임별 STOI 값이다. 다양한 잡음과 잔향 환경에서의 실험을 통해, 제안한 명료도 추정 방법이 기존 표준 P.563에 비해 월등히 우수한 성능을 보임을 확인했다.

Key Words : Deep neural network (DNN), Recurrent neural network (RNN), Long short-term memory (LSTM), Non-Intrusive, Speech intelligibility estimation, STOI, P.563.

ABSTRACT

This paper proposes a non-intrusive speech intelligibility estimation method with no reference speech signal, which is based on recurrent neural network (RNN) with long short-term memory (LSTM) structure. Conventional standard estimation method P.563 has poor estimation performance and lack of consistency especially in various noise and reverberation environments. The proposed method trains the LSTM RNN model parameters by utilizing the STOI that is the standard intelligibility estimation method with reference speech signal. The input and output of the LSTM RNN are the MFCC vector and the frame-wise STOI values. Experimental results show that the proposed intelligibility estimation method outperforms the conventional standard P.563 in various noise and reverberation environments.

Ⅰ. 서 론

디지털 음성통신, 음성인식 등의 분야에서 음성의 명료도를 정확히 추정하는 기술이 필요하다. 음성 명료도의 추정 방법은 기준 음성신호 (reference speech signal)의 유무 여부에 따라 침입적 (intrusive)과 비침입적 (non-intrusive) 방법으로 나뉜다. 표준 침입적 음성 명료도 추정 방법인 STOI (Short-Time Objective Intelligibility measure)[1]는 주파수 영역에서 기준 신호와 왜곡된 신호의 상관도를 계산하는 방법이다. 그리고 대표적인 표준 비침입적 음성 명료도 추정 방법은 P.563[2]이다.

최근 들어, 전 세계적으로 심층신경망 (deep neural network, DNN)을 음성신호처리에 효과적으로 활용하는 연구가 활발히 진행 중이다. 본 논문에서는 DNN중에서 음성신호처리에 적합한 장단기 메모리 (long short-term memory, LSTM)[3] 구조의 순환신경망 (recurrent neural network, RNN)을 기반으로 하는 새로운 비침입적 음성 명료도 추정 방법을 제안한다.

※ 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, deokkyuyun@gmail.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 and IT Media Engineering, Seoul National University of Science and Technology
논문번호 : KICS2017-07-210, Received July 28, 2017; Revised August 14, 2017; Accepted August 16, 2017

Ⅱ. 장단기 메모리 순환신경망 기반 비침입적 음성 명료도 추정

그림 1과 같이 본 연구에서 음성 명료도 추정을 위해 사용한 장단기 메모리 순환신경망의 입력 노드의 개수는 MFCC 벡터의 차수와 같고 출력 노드의 개수는 명료도 추정 값인 1개이다. 신경망의 훈련 과정에서 출력 노드의 값은 프레임별 STOI 점수이고 테스트 과정에서는 프레임별 추정된 음성 명료도 값이다. 본 연구에서는 발성음 (utterance) 별로 한 개의 명료도 점수를 내는 기존 STOI를 수정하여, 프레임별 STOI 점수를 구하고 이를 신경망 훈련 시 출력 값으로 사용한다. 테스트 과정에서는 프레임별 MFCC 벡터를 입력받아 신경망의 출력 즉, 추정된 명료도 값을 구하고, 이를 평균하여 테스트 발성음에 대한 명료도 추정 값을 구한다.

장단기 메모리 모델은 음성신호와 같이 시간적 종속성을 지닌 데이터의 처리에 적합하며, 셀 구조는 그림 2와 같다. 본 연구에서 사용한 신경망의 활성 함수 (activation function)는 노드의 입력 x에 노드 출력이 max(0, x)인 ReLU (Rectified Linear Units)[4]이다. 그리고 신경망 파라미터 학습 시 통계적 최적화 알고리즘인 ADAM (ADAptive Moment estimation)[5]을 사용하였다.

그림 3은 전체적인 장단기 메모리 순환신경망의 훈련 방법을 나타낸다. 다양한 잡음 및 잔향 환경을 반영하기 위해 깨끗한 음성신호, 잡음 또는 잔향이 더해진 음성신호, 그리고 잡음과 잔향 모두 더해진 음성신호에서 구한 MFCC 벡터와 이에 대한 프레임별 STOI 점수를 신경망의 입력과 출력으로 사용한다.

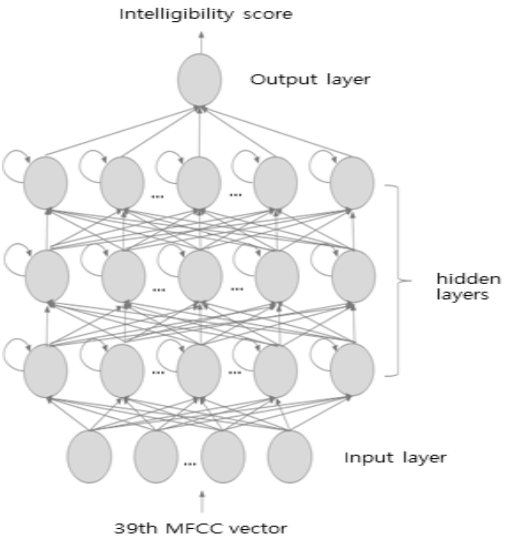

그림 1. 장단기 메모리 순환신경망 구조
Fig. 1. Architecture of LSTM RNN

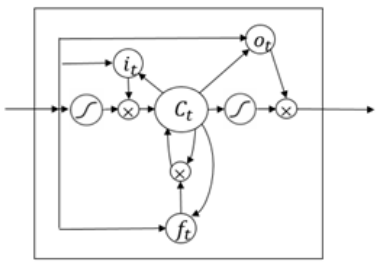

그림 2. 장단기 메모리 셀 구조
Fig. 2. Structure of LSTM cell

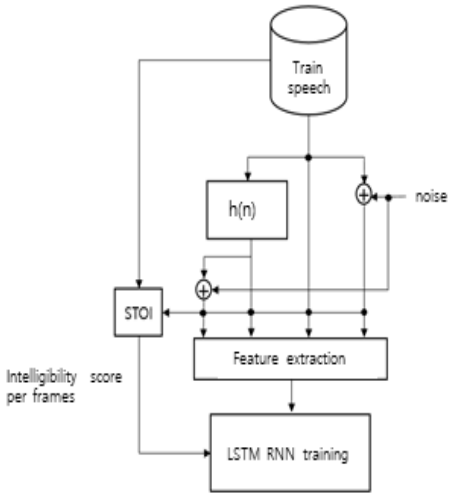

그림 3. 장단기 메모리 순환신경망 훈련 방법
Fig. 3. LSTM RNN training method

Ⅲ. 실험 및 결과

본 연구에서 사용한 신경망은 3개의 은닉층 (hidden layer)으로 구성되고 각 층별 노드의 개수는 256이다. 훈련과 테스트를 위해 TIMIT[6] 음성 데이터베이스를 사용하였으며, 깨끗한 환경, 잡음 또는 잔향 환경, 잡음 및 잔향환경 등 총 14가지의 다양한 환경으로 신경망을 훈련하였다. 테스트는 훈련 환경과 다른 37가지를 포함하여, 총 51가지 환경에서 수행하였다. 잡음 데이터는 일상생활에서 발생하는 휴대폰 소리, 벨소리, 계단 오르는 소리, 개 짖는 소리, TV 뉴스

소리 등이다. 잔향 데이터는 잔향시간이 1.4 초인 홀과 0.2 초인 교실의 실내 임펄스응답을 사용하였다.

표 1과 2에 나타낸 수치는 각각의 방법으로 추정한 명료도 값과 STOI 점수 간의 상관도이다. 각각 훈련된 환경과 훈련되지 않은 테스트 환경에서 잡음환경, 잔향환경, 잡음과 잔향이 함께 있는 환경으로 나눈 것이다. 실험 결과, 장단기 메모리 순환신경망 기반 비침입적 음성 명료도 추정 방법이 모든 환경에서 기존 표준 방법인 P.563과 비교해 우수한 성능을 보였다.

표 1. 훈련 환경과 같은 테스트 환경에서의 P.563과 제안 방법의 STOI와의 상관도
Table 1. Normalized correlation coefficients of P.563 and the proposed method with respect to STOI in known environments.

environ / tool	noise	reverb	noise& reverb	all
P.563	0.39	0.07	0.07	0.51
LSTM	0.97	0.77	0.84	0.97

표 2. 훈련 환경과 다른 테스트 환경에서의 P.563과 제안 방법의 STOI와의 상관도
Table 2. Normalized correlation coefficients of P.563 and the proposed method with respect to STOI in unknown environments.

environ / tool	noise	reverb	noise& reverb	all
P.563	0.68	0.17	0.10	0.69
LSTM	0.92	0.61	0.57	0.77

Ⅵ. 결 론

본 논문에서는 장단기 메모리 순환신경망을 기반으로 한 새로운 비침입적 음성 명료도 추정 방법을 제안하였다. 신경망의 입력은 MFCC 벡터이고 출력은 신경망 훈련 시는 프레임별 STOI 값이고 테스트 시는 프레임별 추정된 명료도 값이다. 다양한 잡음과 잔향 환경에서 신경망을 훈련하였고, 실험결과, 제안한 명료도 추정 방법이 기존 표준 P.563에 비해 우수한 성능을 보임을 확인했다. 따라서 제안한 방법은 비침입적 음성 명료도 추정에 성공적으로 사용될 수 있을 것으로 사료된다.

References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, 2011.

[2] L. Malfait, J. Berger, and M. Kastner, "P. 563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 1924-1934, 2006.

[3] S. Bae, C. M. Lee, I. Choi, and N. S. Kim, "Environmental sound classification using recurrent neural network with long short-term memory," in *Proc. KICS Symp.*, pp. 227-228, 2016.

[4] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Machine Learning*, pp. 807-814, 2010.

[5] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," *NIST*, 1993.