Digital Signal Processing 74 (2018) 102–110

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

DNN-based monaural speech enhancement with temporal and spectral variations equalization

Tae Gyoon Kang [a,1], Jong Won Shin [b,*], Nam Soo Kim [a]

[a] Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[b] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Republic of Korea

ARTICLE INFO

Article history:
Available online 20 December 2017

Keywords:
Monaural speech enhancement
Deep neural network
Objective function
Temporal and spectral variation equalization

ABSTRACT

Recently, deep neural networks (DNNs) were successfully introduced to the speech enhancement area. Conventional DNN-based algorithms generally produce over-smoothed output features which deteriorate the quality of the enhanced speech. In addition, their performance measures calculated in the linear frequency scale do not match the human auditory perception where the sensitivity follows the Mel-frequency scale. In this paper, we propose a novel objective function for DNN-based speech enhancement algorithm. In the proposed technique, a new objective function which consists of the Mel-scale weighted mean square error, and temporal and spectral variations similarities between the enhanced and clean speech is employed in the DNN training stage. The proposed objective function helps to compute the gradients based on a perceptually motivated non-linear frequency scale and alleviates the over-smoothness of the estimated speech. In the experiments, the performance of the proposed algorithm was compared to the conventional DNN-based speech enhancement algorithm in matched and mismatched noise conditions. From the experimental results, we can see that the proposed algorithm performs better than the conventional algorithm in terms of both the objective and subjective measures.

1. Introduction

For a few decades, monaural speech enhancement from additive noise signal has been widely studied to improve various communication and signal processing systems [1,2]. Though considerable performance improvements have been achieved by various approaches, speech enhancement in realistic noise environments still remains a challenging problem.

Early studies on monaural speech enhancement are mostly based on the minimum mean-square error (MMSE) criterion [3,4] which has improved the perceptual speech quality with an affordable amount of musical noise. The quality of the enhanced speech was further improved by adopting various techniques estimating minimum statistics of the acoustic environments minima controlled recursive averaging noise estimation [5–8]. However, algorithms based on this approach have difficulties in tracking non-stationary or speech-like noises which cause speech quality degradation in real-world applications.

In order to enhance the noisy speech in various noise environments, deep neural networks (DNNs) which can learn complicated inter-dependencies between the input variables [9–13] were successfully introduced to the speech enhancement area [14–16]. In these approaches, the DNN provides a mapping between consecutive noisy speech frames and the corresponding clean speech frame with its deep hidden structure. Furthermore, in [17], global variance (GV) equalization post-filter, dropout training, and noise-aware training techniques were incorporated to DNN-based speech enhancement to improve the speech quality in mismatched noise conditions.

Many studies have applied the DNN-based approaches to speech enhancement and target speaker separation with various new ideas. Huang et al. proposed a technique to jointly optimize all the sources with a discriminative objective function for DNN and recurrent neural network (RNN) [18]. Han et al. applied a DNN-based method for joint dereverberation and denoising followed by iterative signal reconstruction [19]. The training targets of the DNNs were studied in [20], and the phase-sensitive filter and complex ratio masking were also proposed [21,22]. Zhang et al. investigated the performance of the mapping- and masking-based training targets both theoretically and experimentally in [23] where they also proposed the multi-context stacking networks for

* Corresponding author.
E-mail address: jwshin@gist.ac.kr (J.W. Shin).
[1] This author is currently with the Samsung Electronics.

https://doi.org/10.1016/j.dsp.2017.12.002
1051-2004/

deep ensemble learning. The multi-objective learning scheme was adopted to utilize secondary targets in [24]. Finally, the divide and conquer strategy was applied by hierarchical DNN and SNR-based progressive learning algorithms [25,26].

Conventional DNN-based speech enhancement algorithms generally apply the objective functions which are related to the mean square error between the enhanced and clean speech features [15–26]. Since these measures compute the errors from various frequency bins with linear frequency scale, they do not align with the human auditory perception where the sensitivity follows the Mel-frequency scale. The perceptual quality of the enhanced speech would be improved if the cost function of DNN reflects relative importance of frequency components considering this non-linear frequency sensitivity.

In addition, it is well-known that the estimated speech trajectories obtained from the DNN-based algorithms are usually over-smoothed compared to those of the clean speech, since the conventional mean square error measures are derived from each time-frequency bin separately rather than from whole spectral trajectory [17]. The speech generated from these enhancement algorithms may result in muffled sound quality and decreased intelligibility [17,27,28]. Several studies have applied the element-wise weight function and the penalty term to the conventional mean square error [29,30]. However, their works were not closely related to the human auditory perception.

In this paper, we propose a novel DNN-based speech enhancement algorithm which computes the gradients based on a perceptually motivated non-linear frequency scale and alleviates the over-smoothness problem by equalizing temporal and spectral variations of the enhanced speech to match those of the clean speech. The main contributions of the proposed algorithm are summarized as follows.

First, we apply the Mel-scale weight to fit the objective function to the critical frequency bands of hearing. Similar to the human auditory perception, the network trained using the Mel-scaled gradients is more sensitive to the perceptually important frequency bins. The Mel-frequency scale was adopted to speech enhancement in [31] to smooth the gain function over spectral coefficients. In contrast, the Mel-scale is introduced to prioritize the gradients according to the perceptual importance in the proposed algorithm.

Second, the objective function for DNN training is modified to incorporate the temporal and spectral variation similarities between the enhanced and clean speech. By equalizing the temporal and spectral variations, the enhanced speech could have the spectral peaks and valleys distributed similarly to those of the clean speech. The proposed objective functions are motivated by the relation between the human intelligibility and short-time analysis on one-third octave band trajectory [32]. We adopt variation similarity over short-time trajectories and spectral coefficients into the DNN-based speech enhancement framework and analyze their effect on the naturalness and intelligibility of the enhanced speech. While the long short-term memory (LSTM) [33] and gated recurrent unit (GRU) [34] can learn the temporal relations of the consecutive input frames, they cannot directly compensate the lack of output feature structure as the proposed approach, and thus these approaches can be jointly applied [35].

The rest of this paper is organized as follows: an overview of the conventional DNN-based speech enhancement technique is given in Section 2 and the Mel-scale weighted mean square error and variation similarities are described in Section 3. Then, the performance evaluation of DNNs with various training algorithms are provided in Section 4. Finally, conclusions are drawn in Section 5.

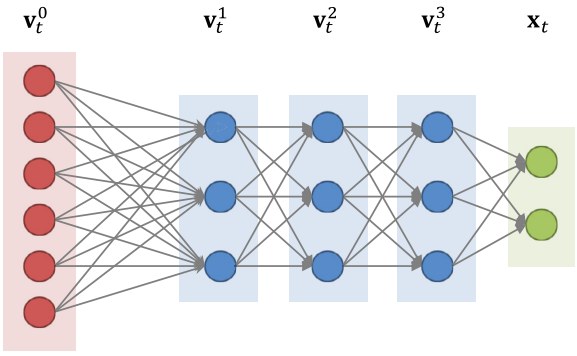

Fig. 1. Scheme of the DNN with three hidden layers.

2. Conventional DNN-based speech enhancement

The task of DNN-based speech enhancement can be divided into the training and test stages. In the training stage, the noisy speech features and the corresponding clean speech features are respectively fed to the input and output nodes of the DNN, and the network is optimized to minimize the mean square error between the enhanced and clean speech features. After the training stage, the clean speech features are estimated from the noisy speech features through the DNN and a GV equalization post-filter is applied to compensate the over-smoothed output trajectory. In this section, we present the feature structures and training scheme of the conventional DNN-based speech enhancement algorithm.

2.1. Training stage

In the training stage, the input and output features of the DNN are respectively extracted from the noisy speech utterances and corresponding clean speech utterances. The input and output features of the DNN are usually normalized to have zero mean and unit variance before being fed to the network.

For the input and output features, we extract log-power spectra of the noisy and clean speech as in [17,19,36]. Recent studies have compared the performance of the mapping-based method which directly estimates the clean speech to the masking-based method which produces the binary or ratio mask targets [20,23,36,37]. It is controversial which method results in better performance [23, 36]. Although this paper focuses on the mapping-based method, the proposed algorithm can also be applied to the masking-based method with slight modification to generate the masked clean speech features.

Let us denote F-dimensional normalized log-power spectra of the noisy speech and clean speech at the t-th frame as $\mathbf{z}_t$ and $\mathbf{y}_t$, respectively. Then, the input feature vector $\mathbf{v}_t^0$ is generally constructed as follows:

$$\mathbf{v}_t^0 = [\mathbf{z}_{t-K}^\dagger,\ \mathbf{z}_{t-K+1}^\dagger,\ \cdots,\ \mathbf{z}_{t+K}^\dagger]^\dagger \tag{1}$$

where K denotes an input context expansion parameter and $\mathbf{z}_t^\dagger$ denotes the transpose of a vector $\mathbf{z}_t$.

Fig. 1 shows the structure of a typical DNN with three hidden layers. The DNN consists of an input layer, a few hidden layers and an output layer which are fully connected to their adjacent layers. For the sake of notational simplicity, the number of hidden layers is assumed to be L and the input and output layers of the DNN are denoted as the 0-th and $(L+1)$-th layers of the DNN, respectively.

The number of nodes in the l-th layer is denoted by n_l. The n_l-dimensional activation vector $\mathbf{v}_t^l$ is generated as

$$\mathbf{v}_t^l = g(\mathbf{a}_t^l) = g(W^l \mathbf{v}_t^{l-1} + \mathbf{b}^l) \tag{2}$$