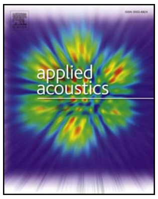
Applied Acoustics 133 (2018) 194–201

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Integrated acoustic echo and background noise suppression based on stacked deep neural networks

Hyeji Seo, Moa Lee, Joon-Hyuk Chang*

School of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea

ARTICLE INFO

Keywords:
Speech enhancement
Noise suppression
Acoustic echo suppression
Deep neural network

ABSTRACT

In this paper, a regression-based integrated acoustic echo and background noise suppression algorithm was proposed through the use of a deep neural network (DNN) with a multi-layer deep architecture. Motivated by an idea that DNNs are a superior hierarchical generative model for modeling the complex relationships between input features and desired target features through its multiple nonlinear hidden layers, a stacked DNN is developed in a sequential fashion such that the DNN for noise suppression is followed by the DNN for acoustic echo suppression. This algorithm is compared to a single DNN-based integrated system to simultaneously suppress acoustic echoes and noise. When developing the DNN-based regression technique using our approach, spectral envelop estimation is a crucial point for which log-power spectra (LPS) are used as features in order to determine the gain, which ensured nonlinear mapping from the LPS of the frames contaminated by echoes and noise to the LPS of the echo- and noise-free frames. This leads to the successful reduction of acoustic echoes and background noise without an additional double-talk detection algorithm. Additionally, an augmented feature technique is adopted to use additional knowledge derived from conventional noise and acoustic echo suppression techniques when designing the DNN architecture in our algorithm. The proposed DNN-based integrated system to suppress acoustic echoes and noise was evaluated in terms of objective measures and demonstrated a significant improvement over conventional integrated algorithms.

1. Introduction

As various Internet of Things (IoT) devices have introduced speech recognition, the importance of nonlinear acoustic echo suppression (AES) and background noise suppression (NS) has increased. If nonlinear acoustic echoes and background noise coexist, the AES and NS algorithms are treated as independent [1]. In this case, two algorithms are separately designed and combined in a serial fashion. However, the performance of the overall suppression algorithm is dependent on the structure of the integrated AES and NS algorithm [2]. For example, if AES is performed before NS, the noise estimation can be hindered through AES processing. Additionally, when AES is performed after NS, the performance of AES can be degraded due to the nonlinear operation of the NS algorithm. To address this issue, many studies have paid particular attention to the integrated acoustic echo and background noise suppression algorithm [3,4]. For example, an integrated system based on a statistical model [5] used the Wiener filter to estimate integrated suppression gain using the combined power of acoustic echoes and background noise based on soft decisions in the frequency domain, which is known to effectively suppress acoustic echoes and background noise and exhibit superior performance. However, the performance of this work in various real environments has not always been satisfactory.

Recently, deep neural networks (DNNs) have attracted considerable attention, particularly in the field of speech enhancement [6,7]. Recent insights into DNN-based regression have allowed the design of mapping functions from noisy speech to clean speech through multiple nonlinear hidden layers. For DNN training, features from noisy speech can be employed in the input layer for large training sets, ensuring the nonlinear mapping of frames from noisy speech to noise-free frames. Recent DNN-based regression methods have focused on how to improve generative abilities in not only de-noising tasks, but also de-reverberation tasks [8]. In [9], a de-noising auto-encoder was developed to reconstruct clean input features from its disrupted features. Also, a two-stage algorithm was derived to separately address de-noising and reverberation. Specifically, an ideal ratio mask (IRM) was estimated from the complementary features pointed out in [10] for the de-noising and de-reverberated log-power spectra (LPS) obtained from noisy speech input [11]. Then, two DNNs were concatenated and jointly trained, which demonstrates higher performance than a single DNN-based method for de-noising and de-reverberation.

* Corresponding author.
E-mail address: jchang@hanyang.ac.kr (J.-H. Chang).

https://doi.org/10.1016/j.apacoust.2017.12.031
Received 26 July 2017; Received in revised form 11 December 2017; Accepted 22 December 2017

DNNs have attracted considerable attention for nonlinear residual echo suppression (RES) in the work of Lee et al. [12], in which the DNN was employed to directly predict the optimal RES gain. The benefit of learning a DNN-based generative model was further magnified when the AES-algorithm was totally implemented by the DNN [13]. However, it is difficult to achieve promising performance with nonlinear AES using a simple DNN structure. Hence, additional knowledge on echoes, including the *a priori* and *a posteriori* signal-to-echo ratio (SER) levels independently derived from [13], were fed into the DNN input. Unlike the studies mentioned above, our DNN-based regression algorithm simultaneously deals with both acoustic echoes and noise, which is a challenging task that had not yet been addressed. For this, a basic DNN framework including a single DNN and stacked DNN is first designed, followed by the proposal of novel techniques to improve the baseline DNN system. In a single DNN, acoustic echoes and noise are suppressed in a single DNN framework for which the single DNN is trained to map the two inputs of noisy speech with echoes and far-end speech into clean speech features. Next, a stack of DNNs is designed, one for NS and the other for AES. The DNN is first trained to learn to map pairs of noisy speech to anechoic clean speech. Then, the output speech is fed into the DNN input designed for AES by stacking the DNN for AES on top of the DNN for NS. As a result, the top DNN is designed after the bottom DNN proved to be well-suited to the bottom DNN developed for NS. It has been reported that side information has been found to be beneficial for training the regression algorithm, and an augmented feature technique is employed when training the DNN for both NS and AES. Because of their multiple hidden layers and hidden units, DNN-based algorithms yield high computational complexity. Despite these shortcomings, DNN-based algorithms have shown the possibility on suppressing both noise and echo. The proposed DNN-based algorithms were evaluated in terms of perceptual evaluation of speech quality (PESQ), frequency weighted segmental SNR (fwSNRseg) and echo return loss enhancement (ERLE) and then the algorithms demonstrated a significant improvement compared to conventional single DNNs and statistical model-based integrated acoustic echo and background noise suppression algorithms [14].

The rest of the paper is organized as follows: Section 2 introduces the statistical model-based integrated suppression approach, Section 3 presents the DNN-based integrated suppression approaches, Section 4 presents the simulation results, and Section 5 presents the conclusions.

2. Statistical model-based integrated acoustic echo and background noise suppression

In this section, the baseline integrated acoustic echo and background noise suppression technique proposed in [5] is briefly reviewed, during which the combined power of acoustic echoes and background noise is estimated for the soft decision scheme. If the discrete Fourier transform (DFT) of a noise signal is $N(i,k)$ and the near end speech signal is $S(i,k)$ for the kth frequency bin at the ith frame, two hypotheses, H_0 and H_1, indicating the absence and presence of speech, respectively, are given as follows:

$$\begin{aligned} &H_0\text{: near–end speech absent: } Y(i,k) = E(i,k) + N(i,k) \\ &H_1\text{: near–end speech present: } Y(i,k) = S(i,k) + E(i,k) + N(i,k) \end{aligned} \tag{1}$$

where $E(i,k)$ and $Y(i,k)$ denote the DFTs of the echo and microphone input signals, respectively. Under the assumption that $N(i,k)$,$E(i,k)$, and $S(i,k)$ are statistically independent and characterized by zero-mean complex Gaussian distributions, the probability density functions (PDFs) of H_0 and H_1 can be given by [5]

$$p(Y(i,k)|H_0) = \frac{1}{\pi\{\lambda_e(i,k) + \lambda_n(i,k)\}}\exp\left[-\frac{|Y(i,k)|^2}{\lambda_e(i,k) + \lambda_n(i,k)}\right] \tag{2}$$

$$p(Y(i,k)|H_1) = \frac{1}{\pi\{\lambda_s(i,k) + \lambda_e(i,k) + \lambda_n(i,k)\}}\exp\left[-\frac{|Y(i,k)|^2}{\lambda_s(i,k)\lambda_e(i,k) + \lambda_n(i,k)}\right] \tag{3}$$

where $\lambda_e(i,k)$,$\lambda_n(i,k)$, and $\lambda_s(i,k)$ represent the variance of echoes, noise, and near-end speech, respectively. The near-end speech absence probability (NSAP) $p(H_0|Y(i,k))$ for each frequency bin can be represented using Bayes' rule such that [5]

$$\begin{aligned} p(H_0|Y(i,k)) &= \frac{p(Y(i,k)|H_0)p(H_0)}{p(Y(i,k)|H_0)p(H0) + p(Y(i,k)|H_1)p(H_1)} \\ &= \frac{1}{1 + q\Lambda(Y(i,k))} \end{aligned} \tag{4}$$

where $p(H_0)$(=1−$p(H_1)$) represent the *a priori* probability of near-end speech absence and $q = p(H_1)/p(H_0)$. Substituting Eqs. (2) and (3) into Eq. (4), the likelihood ratio $\Lambda(Y(i,k))$ can be shown as follows:

$$\Lambda(Y(i,k)) = \frac{p(Y(i,k)|H_1)}{p(Y(i,k)|H_0)} = \frac{1}{1 + \xi(i,k)}\exp\left[\frac{\gamma(i,k)\xi(i,k)}{1 + \xi(i,k)}\right] \tag{5}$$

where $\gamma(i,k)$ and $\xi(i,k)$ denote the *a posteriori* and *a priori* signal-to-combined power ratio (SCR) as defined by [5]:

$$\gamma(i,k) \equiv \frac{|Y(i,k)|^2}{\lambda_c(i,k)}, \tag{6}$$

$$\xi(i,k) \equiv \frac{\lambda_s(i,k)}{\lambda_c(i,k)} \tag{7}$$

where $\lambda_c(i,k)$ is the combined power of acoustic echoes and background noise to be estimated. Also, $\widehat{\xi}(i,k)$ can be estimated with the help of a well-known decision-directed (DD) approach as given by

$$\widehat{\xi}(i,k) = \alpha_D D\frac{|\widehat{S}(i-1,k)|^2}{\widehat{\lambda}_c(i-1,k)} + (1-\alpha_D D)P[\gamma(i,k)-1] \tag{8}$$

where $\widehat{S}$,$(i-1,k)$ is the kth frequency estimate of near-end speech in the previous frame, and $\widehat{\lambda}_c(i-1,k)$ is the long-term smoothed combined acoustic echo and background noise power. Also, $\alpha_D D$ is the smoothing parameter. The combined acoustic echo and background noise power $\lambda_c(i,k)$ can be estimated with the assumption that acoustic echoes and background noise are uncorrelated. Indeed, $\widehat{\lambda}_c(i-1,k)$ can be determined as follows:

$$\widehat{\lambda}_c(i,k) = \alpha_{\lambda_c}\widehat{\lambda}_c(i-1,k) + (1-\alpha_{\lambda_c})\{\widehat{\lambda}_e(i,k) + E[|N(i,k)|^2|Y(i,k)|]\} \tag{9}$$

where α_{λ_c} denotes the smoothing parameter and $\widehat{\lambda}_e$ is the echo power estimated during the near-end absence. Also, the noise power estimate $E[|N(i,k)|^2|Y(i,k)|]$ can be adaptively calculated during noise-only periods, as detected through a voice activity detection (VAD) algorithm. Then, the clean near-end speech estimate $\widehat{S}(i,k)$, acoustic echo, and noise suppressed spectra can be given as follows:

$$\widehat{S}(i,k) = (1-P(H_0|Y(i,k)))G(i,k)Y(i,k) = \widetilde{G}(i,k)Y(i,k) \tag{10}$$

where $P(H_0|Y(i,k))$,$G(i,k)$, and $\widetilde{G}(i,k)$ represent the NSAP defined in Eq. (4), the integrated suppression gain, and the overall suppression gain, respectively. Here, the integrated suppression gain $G(i,k)$ is derived from the Wiener filter. It is known that the overall suppression gain $\widetilde{G}(i,k)$ plays a role in preserving the quality of near-end speech by adapting the soft decision scheme.

3. Proposed DNN-based acoustic echo and background noise suppression

In this section, two systems are proposed for the simultaneous suppression of acoustic echoes and background noise in the integrated system through a single DNN and stacked DNNs. A method to improve the proposed system for which additional information can be plugged into the DNN training was subsequently devised.

3.1. Integrated system based on a single DNN

The first integrated system is devised, for which acoustic echoes and