A review of multi-objective deep learning speech denoising methods

Arian Azarang*, Nasser Kehtarnavaz

Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX, USA

ARTICLE INFO

ABSTRACT

This paper presents a review of multi-objective deep learning methods that have been introduced in the literature for speech denoising. After stating an overview of conventional, single objective deep learning, and hybrid or combined conventional and deep learning methods, a review of the mathematical framework of the multi-objective deep learning methods for speech denoising is provided. A representative method from each speech denoising category, whose codes are publicly available, is selected and a comparison is carried out by considering the same public domain dataset and four widely used objective metrics. The comparison results indicate the effectiveness of the multi-objective method compared with the other methods, in particular when the signal-to-noise ratio is low. Possible future improvements that can be achieved are also mentioned.

1. Introduction

It is well known that noisy environments have a degrading effect on the quality and intelligibility of speech signals. Operating in noisy environments affect the performance of speech processing systems that are normally designed based on clean speech signals (Stone et al., 2012, Brown and Bacon, 2010, Dahl et al., 2011, Premananda and Uma, 2013, Wang et al., 2009, Boll and Pulsipher, 1980, Yang and Fu, 2005, Li et al., 2014, Sarampalis et al., 2009, Seok and Bae, 1997, April, Fakoor et al., 2018, April, Kemp and Waibel, 1999, Hughes et al., 1999, Tu et al., 2019, Besacier et al., 2000, June, Isabelle, 2000). Speech denoising involves the reduction or removal of the noisy part of the speech+noise mixture. Speech denoising has been extensively studied in the speech processing literature and have been deployed in a wide range of speech processing systems including speech recognizers (Bennett, 2010, Vu et al., 2012, March), voice activity detectors (Zhang and Wu, 2012, Ghosh et al., 2010, Ying et al., 2011), hearing aids (Cornelis et al., 2010, Ma et al., 2010) and voice over IP (VoIP) (Peláez-Moreno et al., 2001).

Initial attempts at speech denoising involved applying linear filters to the mixture to reduce or eliminate the noise portion, most prominently Wiener filtering (Lim and Oppenheim, 1979), by estimating the noise statistics. In this review, speech denoising methods are placed into two main categories (Wang and Chen, 2018, Hermus and Wambacq, 2006, Loizou, 2013, Chaudhari and Dhonde, 2015, January): conventional methods, including Wiener filtering, spectral subtraction, and Minimum Mean Square Error (MMSE) methods (Ephraim and Malah, 1984, Martin, 2002, May, Scalart, 1996, May, Cohen and Berdugo, 2002, Varga and Moore, 1990, April, Sameti et al., 1998, Mowlaee and

Saeidi, 2013, Dat et al., 2005, March, Kamath and Loizou, 2002, May, Paliwal, 1987, April), and more recent deep learning-based methods (Li et al., 2019, Lu et al., 2013, August, Xu et al., 2015, Xu et al., 2013, Xu et al., 2014, Sun et al., 2016, Liu et al., 2014, Gao et al., 2016, Donahue et al., 2018, April, Xia and Bao, 2014, Park and Lee, 2016, Mirsamadi and Tashev, 2016, September). Conventional methods attempt to estimate the statistical attributes in the mixture. A shortcoming of conventional methods is that their effectiveness significantly diminishes in nonstationary noisy environments (Sulong et al., 2017, Mirbagheri et al., 2013). More recently, deep learning methods have been introduced for speech denoising which are shown to perform better in nonstationary noisy environments compared with conventional methods. Deep learning methods attempt to model the nonlinear relationship between the mixture and the clean speech signals without knowing the noise statistics (Zhao et al., 2018, Qian et al., 2017, August, Saleem et al., 2018, June). There have also been speech denoising solutions where conventional and deep learning methods are combined (named hybrid methods) (Tu et al., 2018, April, Nie et al., 2018, Valin, 2018, August). In the hybrid methods, denoised speech signals are obtained by incorporating conventional methods into the deep learning framework. The latest deep learning methods have incorporated not a single objective or loss function to optimize the parameters of a deep neural network but rather a combination of a number of different objective or loss functions (named multi-objective deep learning methods) (Sun et al., 2017, March, Xu et al., 2017, Huang et al., 2015, Shivakumar and Georgiou, 2016, Zhao et al., 2018, April, Zhang et al., 2018, April, Han et al., 2016, June, Lai and Zheng, 2019, Williamson et al., 2016, March, Liu et al., 2018, Martín-Doñas et al., 2018, Fu et al., 2016,

---

September, Fu et al., 2017, September, Wang et al., 2018, Fu et al., 2016, Lee et al., 2018).

A number of reviews of speech denoising has previously appeared in (Wang and Chen, 2018, Hermus and Wambacq, 2006, Loizou, 2013, Chaudhari and Dhonde, 2015, January). This review differs from the previous reviews by focusing on the multi-objective deep learning methods, which are the state-of-the-art speech denoising methods at the time of this writing. For the sake of completeness and for the purpose of laying the mathematical foundation to review the multi-objective deep learning methods, an overview of conventional methods, single objective deep learning methods, and hybrid methods is first stated in Section II. Then, in Section III, the multi-objective deep learning methods are reviewed. A comparison of representative methods in the above speech denoising categories is reported in Section IV followed by mentioning possible future improvements in Section V. Finally, the paper is concluded in Section VI.

2. Speech denoising methods

2.1. Conventional methods

A widely used and one of the earliest conventional methods is Wiener filtering (Lim and Oppenheim, 1979). In general, conventional methods attempt to obtain an estimate of the noise statistics. Another conventional method is spectral subtraction, developed by Boll (Boll, 1979), where the noise in the frequency domain over non-speech activity segments is used to subtract it from the mixture during speech activity segments. MMSE-based (Ephraim and Malah, 1985) and its variations (Ephraim and Malah, 1984, Martin, 2002, May, Scalart, 1996, May, Cohen and Berdugo, 2002, Varga and Moore, 1990, April, Sameti et al., 1998, Mowlaee and Saeidi, 2013, Dat et al., 2005, March, Kamath and Loizou, 2002, May, Paliwal, 1987, April) is another popular conventional method. As stated earlier, a major shortcoming of these conventional methods is that their performance is adversely impacted in the presence of nonstationary noisy environments.

2.2. (Single objective) deep learning methods

After the initial introduction of deep learning networks by Hinton in 2006 (Hinton and Salakhutdinov, 2006), different deep learning configurations have been introduced such as Autoencoders (Vincent et al., 2010), Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012), and Recurrent Neural Networks (RNNs) (Graves et al., 2013, May). Autoencoders are designed to reconstruct the input at their output. They consist of encoding layers and decoding layers. Encoding layers abstract the input by removing redundant information in the input, while decoding layers reverse the process. CNNs first obtain high level representations of the input through several convolutional layers which are then distinguished by fully connected layers. RNNs are similar to CNNs but incorporate feedback connections in addition to forward connections.

Single objective deep learning methods can be placed into two major groups: (1) Time-frequency domain methods, e.g. Wang et al., 2017, Tu et al., 2019, May, Gao et al., 2018, April, Chai et al., 2017, September, Wang et al., 2014), in which generally LPS features are used as the input and the target of a deep neural network. Mapping-based (direct mapping) methods, e.g. (Mirsamadi and Tashev, 2016, September, Gao et al., 2018, April, Chai et al., 2017, September), in which the magnitude of LPS features are considered as training samples. Masking-based methods constitute another category in this group, e.g. (Hu and Wang, 2001, October, Hummersone et al., 2014, Narayanan and Wang, 2013, May), where the target of the network is considered to be an intermediate mask indicating the presence of clean and noise signals. This approach was first introduced in (Hu and Wang, 2001, October) by considering a binary mask that assigned 0 or 1, respectively, to the existence or absence of noise. In (Hummersone et al., 2014), a soft mask was considered instead of a binary mask (Hu and

Wang, 2001, October), which was shown to be more effective compared to the binary mask. Several improvements to masking-based methods have also appeared. For example, in (Erdogan et al., 2015, April), a phase-sensitive filter was proposed which incorporated the phase information into the estimated mask. In another category in this group, a set of complementary features as well as the masking and mapping-based targets developed for the training of deep neural networks (Wang et al., 2014) ((2) Time-domain methods, e.g. (Fu et al., 2017, December, Wang et al., 2018, Venkataramani et al., 2018, November, Venkataramani and Smaragdis, 2018, Pandey and Wang, 2019, May, Pascual et al., 2019, Kolbæk et al., 2019, Pandey and Wang, 2019, Longueira and Keene, 2018), in which raw data in the time domain is considered. One of the earliest works in this group was reported in (Fu et al., 2017, December), where a fully convolutional neural network was used to map noisy time domain signals to their clean version. Another example is the method discussed in (Wang et al., 2018) where an end-to-end speech denoising was performed by minimizing the reconstruction error in the time domain.

2.3. Hybrid methods

Tu et al. introduced a hybrid method in (Tu et al., 2018, April) by incorporating conventional methods into the deep learning framework. An LSTM Multi-Style (LSTM-MT) model was trained using LPS as the input and clean LPS and Ideal Ratio Mask (IRM) as the targets. In what follows, the equations for the hybrid methods are briefly mentioned in order to lay the mathematical foundation for the review of the multi-objective methods in the next section.

In the above hybrid method, the denoising process for the l-th audio frame is achieved in three steps. The first step, called Approximate Speech Signal Estimation (ASSE), preprocesses the noisy LPS $X(l)$ by considering a suppression rule, resulting in a clean speech estimate $Y(l)$. In the second step, the trained LSTM-MT network uses $Y(l)$ to generate estimates of IRM $M(l)$ and clean speech $\hat{S}(l)$. In the final step, the estimates of IRM $M(l)$ and clean speech $Y(l)$ are utilized to estimate the output speech signal $Z(l)$. The above steps are elaborated below.

The prior and posterior Signal-to-Noise Ratios (SNRs), denoted by $\gamma(k, l)$ and $\xi(k, l)$, respectively, are estimated using the following widely used equations

$$\gamma(k,l) \triangleq \frac{|X(k,l)|^2}{\sigma(k,l)}, \ \xi(k,l) \triangleq \frac{|S(k,l)|^2}{\sigma(k,l)} \tag{1}$$

where $\sigma(k, l)$ indicates the noise variance in time frame l and frequency bin k. Then, the noise reduction or suppression is achieved by the gain $G(k, l)$ as follows

$$G(k,l) = g(\gamma(k,l), \ \xi(k,l)) \tag{2}$$

where $g(.)$ is a function of prior and posterior SNRs described in (Han et al., 2016, June). ASSE is then computed using the following equation

$$Y(k,l) = \log\left[\delta M(k,l) + (1-\delta)G(k,l)\right] + X(k,l) \tag{3}$$

where δ denotes a weighting parameter.

Next, IRM is computed which is the ratio of the powers of the clean and noisy speech signals according to

$$M_{ref}(k,l) = \frac{|S(k,l)|^2}{|X(k,l)|^2} \tag{4}$$

The objective function for training is the combination of the terms noted below

$$E = \sum_{k,l} \left[\left(\hat{S}(k,l) - S(k,l)\right)^2 + \left(M(k,l) - M_{ref}(k,l)\right)^2 \right] \tag{5}$$

In another more recent hybrid method in (Nie et al., 2018), the speech signal $S(k, l)$ is recovered and the noise signal $V(k, l)$ is reduced by applying a linear filter transfer function $H(k, l)$, for example Wiener

filter, to the observation $X(k, l)$. The Wiener filter transfer function is given by

$$H(k,l) = \frac{\phi_{SS}(k,l)}{\phi_{SS}(k,l) + \phi_{VV}(k,l)} \qquad (6)$$

where $\phi_{VV}(k, l)$ denotes the Power Spectral Density (PSD) of the noise signal expressed as

$$\phi_{VV}(k,l) = E\{V(k,l)V^*(k,l)\} \qquad (7)$$

Likewise, the PSD of the desired signal is expressed as

$$\phi_{SS}(k,l) = E\{S(k,l)S^*(k,l)\} \\ = \phi_{XX}(k,l) - \phi_{VV}(k,l) \qquad (8)$$

where $\phi_{XX}(k,l) = E\{X(k,l)X^*(k,l)\}$ denotes the PSD of the observed signal. For real-time operation, at time frame l, the PSDs of the noise signal and the observed signal can be updated recursively as follows

$$\hat{\phi}_{VV}(k,l) = \hat{\alpha}_V(k,l)\hat{\phi}_{VV}(k,l-1) + (1-\hat{\alpha}_V(k,l))X(k,l)X^*(k,l) \qquad (9)$$

$$\hat{\phi}_{XX}(k,l) = \hat{\alpha}_X(k,l)\hat{\phi}_{XX}(k,l-1) + (1-\hat{\alpha}_X(k,l))X(k,l)X^*(k,l) \qquad (10)$$

where $\hat{\alpha}_V$ and $\hat{\alpha}_X$ denote smoothing weights for the PSDs of the noise and the input mixture, respectively.

The PSD of the noise signal can be estimated based on the past average of power spectral values in a time-varying smoothing manner using a so-called Speech Presence Probability (SPP). However, when noise is nonstationary, or the level of SNR is low, it is quite challenging to acquire an accurate SPP, which limits the tracking capability of the noise estimator in case of time-varying noise.

To address this limitation, a deep learning network is used to replace SPP. This way, conventional and deep learning modules are merged together as part of the same solution. Two-layers of a Gated Recurrent Unit (GRU) and a one-layer of a feedforward network are utilized to construct a deep noise tracking network. A sigmoid output layer follows the GRU generating a vector of 0 and 1 elements denoted by α_V, which is used as the SPP of the current frame enabling an adapting smoothing factor. After updating the network parameters, the final desired speech signal is obtained via the following equation

$$\hat{S}(k,l) = \frac{\phi_{XX}(k,l) - \phi_{VV}(k,l)}{\phi_{XX}(k,l)} X(k,l) \qquad (11)$$

The following MSE-based magnitude objective function is used to optimize the network parameters

$$J = \frac{1}{L}\sum_{l=0}^{L}\sum_{k=0}^{K}\left|\hat{S}(k,l) - S(k,l)\right|^2 \qquad (12)$$

where L and K correspond to the total number of frequency bins and time frames, respectively.

A real-time low-complexity hybrid speech denoising method was also discussed in (Valin, 2018, August) which uses a combination of a conventional method within a RNN network.

3. Multi-objective deep learning methods

A multi-objective deep learning method, named LSTM-RNN, was first proposed in (Sun et al., 2017, March). Initially, a Direct Mapping (DM) method is utilized based on a linear output layer and this Minimum Mean Square Error (MMSE) objective function

$$E_{DM} = \sum_{k,l}\left(\hat{S}(k,l) - S(k,l)\right)^2 \qquad (13)$$

in which $\hat{S}(k,l)$ and $S(k, l)$ are the estimated and the reference clean LPS features at the (Time-Frequency) T-F bin, respectively. LSTM-IRM is another deep learning method introduced in (Hummersone et al., 2014)

which is an extension of the Ideal Binary Mask (IBM) method. IRM has been shown to be effective as a soft mask defined as follows

$$M_{ref}(k,l) = \frac{S(k,l)}{S(k,l) + V(k,l)} \qquad (14)$$

The corresponding objective function is stated as

$$E_{IRM}(k,l) = \sum_{k,l}\left(\hat{M}(k,l) - M_{ref}(k,l)\right)^2 \qquad (15)$$

As noted earlier $\hat{M}(k,l)$ and $M_{ref}(k, l)$ are the estimated and reference IRM, respectively.

LSTM-DM would be the right option for speech denoising if the underlying clean speech can be perfectly reconstructed. However, due to data limitations and the local minima aspect of the LSTM-RNN optimization, it would be difficult to learn the relationship of the noisy and clean speech signals in LSTM-DM. As a result, one cannot say which approach is better, in particular in the presence of unseen noise signals, and across different SNR levels.

The idea here is to jointly learn the clean speech and IRM in one single LSTM-RNN using dual outputs, called Multi Task Learning (MTL), whose objective function is expressed as

$$E_{MTL} = \sum_{k,l}\left[\left(\hat{S}(k,l) - S(k,l)\right)^2 + \alpha_{MTL}\left(\hat{M}(k,l) - M_{ref}(k,l)\right)^2\right] \qquad (16)$$

where α_{MTL} denotes a weight for the dual outputs $\hat{S}(k,l)$ and $\hat{M}(k,l)$. During the denoising stage, the estimated clean speech and IRM can be combined via an averaging operation in the LPS domain as follows

$$\hat{Z}(k,l) = \frac{1}{2}\left[\hat{S}(k,l) + \log\hat{M}(k,l) + X(k,l)\right] \qquad (17)$$

Another multi-objective framework was proposed in (Xu et al., 2017) in order to optimize a joint objective function, involving errors not only for the primary clean LPS features but also errors in the secondary target for continuous features, such as MFCCs, and for categorical information, such as IBM. This joint optimization of different but related targets increases the DNN performance on the prediction of the primary target LPS which is then used to reconstruct the denoised output signal. When a DNN is considered to serve as a mapping function between the noisy and clean LPS features, no assumption is imposed during the training process. On the other hand, in masking-based methods, some constraints, such as the independency of noise and clean speech, need to be imposed.

The multi-objective learning approach considers the loss function of LPS features together with MFCC features and the IBM mask as noted below

$$E = \frac{1}{N}\sum_{l=1}^{N}\frac{\left\|\hat{S}_l\left(X_l, X_{l\pm\tau}^{cont}, W, b\right) - S_l\right\|_2^2}{\|X_n\|_2^2} + \alpha_{cont} *$$
$$\frac{1}{N}\sum_{l=1}^{N}\frac{\left\|\hat{S}_l^{cont}\left(X_{l\pm\tau}, X_{l\pm\tau}^{cont}, W, b\right) - S_l^{cont}\right\|_2^2}{\left\|X_l^{cont}\right\|_2^2}$$
$$+\alpha_{cate} * \frac{1}{N}\sum_{l=1}^{N}\left\|\hat{S}_l^{cate}\left(X_{l\pm\tau}, X_{l\pm\tau}^{cate}, W, b\right) - S_l^{cate}\right\|_2^2 \qquad (18)$$

where $\hat{S}^{cont}$ and S^{cont} correspond to the estimated and clean features, $\hat{S}^{cate}$ and S^{cate} denote the estimated and target meta category information, α_{cont} and α_{cate} are the weighting coefficients of the second and third terms, respectively. The prediction for the secondary feature is complementary to the primary LPS features. The IBM learning can also improve the clean speech estimation.

Although many improvements have been made by so called Denoising AutoEncoders (DAEs) in speech denoising applications, they often yield speech distortion due to over-smoothing and clipping clean speech using the MSE objective loss. This affects the perceptual speech

quality leading to a muffled sound. A perceptron optimized deep denoising autoencoder for single channel speech denoising appeared in (Shivakumar and Georgiou, 2016) which is described next.

The conventional DAE uses the MSE loss function to map the noisy to the clean signal. This loss or error is defined as

$$E = \frac{1}{2}\left\| S - \hat{S} \right\|_2^2 \tag{19}$$

where $\hat{S}$ is the output of the DAE in feedforward propagation and S is the clean target mentioned before. The backpropagation of the error is derived according to

$$\frac{\partial E}{\partial \hat{S}} = \frac{\partial}{\partial \hat{S}} \frac{1}{2}\left(S - \hat{S} \right)^2 = \hat{S} - S \tag{20}$$

The linearity in the computation of the above gradient often leads to the over-smoothing problem typically seen in DAE. This is due to the fact that the penalty for clipping speech is the same as clipping noise as long as the Euclidian distance from the clean target is the same. In this method, to maintain the perceptual quality, it is better to preserve speech segments with the residual error rather than clipping speech segments to remove noise. Hence, a new loss has been designed which considers high penalty against signal removal and preserves the same error for noise removal. This objective function is given by

$$E = \begin{cases} \frac{1}{2}\left\| S - \hat{S} \right\|_2^2 & if \ \hat{S} \geq S \\ \frac{1}{2}\left\| S - \hat{S} + p \right\|_2^2 & if \ \hat{S} < S \end{cases} \tag{21}$$

where p is a positive scalar denoting the penalty of speech clipping. The gradient is then computed as follows

$$\frac{\partial E}{\partial \hat{S}} = \begin{cases} \hat{S} - S & if \ \hat{S} \geq S \\ \hat{S} - S - p & if \ \hat{S} < S \end{cases} \tag{22}$$

This new objective function is equivalent to the conventional MSE loss when the penalty is set to zero ($p = 0$). The training phase of DAE leads to an estimate of the clean signal while avoiding the removal of the desired signal.

In another study, a perceptually guided speech denoising was proposed via a deep neural network (Zhao et al., 2018, April). From the hearing perception perspective, the MSE loss function is not optimal. It helps to incorporate the quality and intelligibility knowledge of speech into the loss function. In this approach, the short-time objective intelligibility measure (STOI) (Taal et al., 2011) is added to the objective function to optimize speech intelligibility. For this new loss function, the following steps are considered to derive the modified STOI function. Assuming a 16 kHz sampling rate, for each time frame, a 512-point Fast Fourier Transform (FFT) is taken yielding 25 frequency bins. Then, the frequency bins are grouped together to form one-third octave bands. Let $S(k, l)$ and $\hat{S}(k, l)$ denote the STFT representation of the clean and denoised speech signals, respectively. The corresponding frequency bins are grouped into 15 one-third octave bands. Then, the new T-F representation appears as follows

$$S_j(l) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} \| S(k, l) \|_2^2} \tag{23}$$

$$\hat{S}_j(l) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} \left\| \hat{S}(k, l) \right\|_2^2} \tag{24}$$

where j corresponds to the index of one-third octave bands, k_1 and k_2 denote the edges of these bands, and $\| \cdot \|_2$ indicates L_2 norm. As a result, the short-term temporal envelope of the clean and denoised signals can be expressed as

$$s_{l,j} = \left[S_j(l), S_j(l+1), \ldots, \ S_j(l+R-1) \right]^T \tag{25}$$

$$\hat{s}_{l,j} = \left[\hat{S}_j(l), \hat{S}_j(l+1), \ldots, \ \hat{S}_j(l+R-1) \right]^T \tag{26}$$

Algorithm 1. Gradient Approximation.

Initialization: Noise variance σ^2, learning rate α, initial parameters I_0
for $l = 1, 2, 3, \ldots, L$ **do**
Sample $\partial_1, \ldots, \partial_n \sim N(0, I)$
Calculate $H_i = h(w_l + \sigma_i) - h(w_l)$ for $i = 1, \ldots, N$
Do $w_{l+1} \leftarrow w_l - \alpha \frac{1}{N\sigma} \sum_{i=1}^{N} H_i / \partial_i$
end for

Noting that the analysis window is 384ms in length, R is set to 24. Based on the original STOI computation, the following equation is used to normalize and clip the denoised speech signal

$$\bar{s}_{l,j}(i) = \min\left(\frac{\left\| s_{l,j} \right\|_2}{\left\| \hat{s}_{l,j} \right\|_2} \hat{s}_{l,j}(i), \ \left(1 + 10^{-\frac{\beta}{20}} \right) s_{l,j}(i) \right) \tag{27}$$

where $i = 1, 2, \ldots, N$; β controls the lower bound of the Signal to Distortion Ratio (SDR).

The intermediate speech intelligibility measure is computed using the correlation coefficients between the vectors $s_{l,j}$ and $\bar{s}_{l,j}$, namely

$$d_{l,j} = \frac{\left(s_{l,j} - \mu_{s_{l,j}} \right)^T \left(\bar{s}_{l,j} - \mu_{\bar{s}_{l,j}} \right)}{\left\| s_{l,j} - \mu_{s_{l,j}} \right\|_2 \left\| \bar{s}_{l,j} - \mu_{\bar{s}_{l,j}} \right\|_2} \tag{28}$$

in which $\mu(.)$ indicates the mean vector. The modified STOI function is defined as follows

$$d_l = f\left(S_l^{24}, \hat{S}_l^{24} \right) = \frac{1}{J} \sum_j d_{l,j} \tag{29}$$

where S_l^{24} and $\hat{S}_l^{24}$ are the 24-frame magnitude spectrum starting from the time frame m of the clean and denoised speech signals, respectively, and J denotes the total number of the one-third octave bands.

To obtain the initial denoised speech signal, the IRM is first found. Then, the denoised speech obtained via the denoising module is fed into the modified STOI function. The following loss function is used to update the network parameters

$$(l) = \left(1 - f\left(S_l^{24}, \hat{S}_l^{24} \right) \right)^2 + \lambda \left\| S_l^{24} - \hat{S}_l^{24} \right\|_F / 24 \tag{30}$$

As discussed earlier, MSE as a loss function leads to over-smoothing speech trajectories and thus generating muffled sound. Also, MSE treats each element with equal importance which is not the case in realistic audio environments. In (Zhang et al., 2018, April), a supervised speech denoising was proposed to improve address the MSE shortcoming. The loss function used is based on the two widely used speech quality and speech intelligibility metrics, namely STOI and Perceptual Evaluation of Speech Quality (PESQ) (ITU-T P.862 2001).

The critical aspect regarding loss functions is their differentiability. Incorporating STOI and PESQ into the loss function is a challenging task since it cannot be differentiated using the standard algorithms such as gradient descent. A gradient approximation algorithm was introduced in the above approach whose steps are outlined below.

(Algorithm 1).

A deep neural network with perceptual connection weights was considered for monaural speech denoising in (Han et al., 2016, June). In this method, the auditory perception was incorporated into the reconstruction error to fine-tune the weights of the network. The frequency spectrum of the error was shaped in order to obtain good quality auditory masking performance. To do so, less emphasis was placed near the formant peaks, and more emphasis was placed on the spectral valleys. The following filter was considered in this method

$$P(z) = \frac{A\left(\frac{z}{\gamma_1} \right)}{A\left(\frac{z}{\gamma_2} \right)} = \frac{1 - \sum_{k=1}^{q} a_k \gamma_1^k z^{-k}}{1 - \sum_{k=1}^{q} a_k \gamma_2^k z^{-k}} \tag{31}$$

where $A(z)$ is the LPC polynomial, a_ks are the short-term linear prediction coefficients, γ_1 and γ_2 are the parameters controlling the energy error in the formant regions, and q is the prediction error.

Then, the magnitude of the clean speech S as well as the frequency response of the perceptual filter P were used as a joint objective function. Mathematically, the following equation was used

$$J_{MSE} = \left\| W_f \left(\hat{S} - S \right) \right\|^2 + \left\| \hat{P} - P \right\|^2 \tag{32}$$

where the second term indicates the error between the output $\hat{P}$ and the reference P. This term makes use of the clean speech for fine-tuning W_f as the perceptual weighting matrix which is defined as follows

$$W_f = \begin{bmatrix} P(0) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P\big((N-1)\omega_0\big) \end{bmatrix} \tag{33}$$

with $\omega_0 = \frac{2\pi}{T}$ and T being the FFT length.

To make use of the input information for a single channel speech denoising, a complex ratio masking for joint optimization of magnitude and phase was presented in (Williamson et al., 2016, March). The aim of this method was to derive a complex ratio mask that produces the STFT of the clean speech when applied to the STFT of the noisy speech. In other words, the following complex spectrum of the clean speech $S_{k,l}$ was considered

$$\hat{S}_{k,l}^c = \hat{M}_{k,l}^c * X_{k,l}^c \tag{34}$$

where "*" indicates complex multiplication, $\hat{M}_{k,l}^c$ is the complex IRM (cIRM), and $X_{k,l}^c$, $\hat{M}_{k,l}^c$ and $\hat{S}_{k,l}^c$ are complex numbers expressed as

$$X^c = X_r + iX_i \tag{35}$$

$$\hat{M}^c = \hat{M}_r + i\hat{M}_i \tag{36}$$

$$\hat{S}^c = \hat{S}_r + i\hat{S}_i \tag{37}$$

with the indices r and i denoting the real and imaginary components, respectively. Based on these definitions, Eq. (37) can be written as

$$\begin{aligned} \hat{S}_r + i\hat{S}_i &= \left(\hat{M}_r + i\hat{M}_i \right) * \left(X_r + iX_i \right) \\ &= \left(\hat{M}_r Y_r - \hat{M}_i X_i \right) + i\left(\hat{M}_r X_i + \hat{M}_i X_r \right) \end{aligned} \tag{38}$$

The real and imaginary parts of the clean speech are then given by

$$\hat{S}_r = \hat{M}_r X_r - \hat{M}_i X_i \tag{39}$$

$$\hat{S}_i = \hat{M}_r X_i + \hat{M}_i X_r \tag{40}$$

After solving for $\hat{M}_r$ and $\hat{M}_i$ by using Eq. (39) and Eq. (40), the complex ideal ratio mask $\hat{M}^c$ is computed as follows

$$\hat{M}^c = \frac{X_r \hat{S}_r + X_i \hat{S}_i}{X_r^2 + X_i^2} + i \frac{X_r \hat{S}_i - X_i \hat{S}_r}{X_r^2 + X_i^2} \tag{41}$$

The cIRM acts similar to the Wiener filter.

In (Liu et al., 2018), a Shifted Real Spectrum (SRS) mask was proposed for a single channel speech denoising. Given a time-domain signal x, one can decompose it into the time-frequency domain via Discrete Time Fourier Transform (DTFT). The result is a complex spectrum X^c with the real and imaginary parts of the signal denoted by X_r and X_i, respectively.

It is well known that a time-domain signal x can be represented in terms of its even and odd parts

$$x = x_{even} + x_{odd} \tag{42}$$

where $x_{even} = IDTFT(X_r)$ and $x_{odd} = IDTFT(jX_i)$. In SRS, the signal x is padded with zeros to make $x(l) = 0$ when $l \le 0$ with l being the time index. The decomposition using x_{even} and x_{odd} parts can be stated as

$$x_{even}(l) = x_{odd}(l) = \frac{1}{2} x(l) \text{ if } t > 0 \tag{43}$$

$$x_{even}(l) = -x_{odd}(t) \text{ if } t \le 0 \tag{44}$$

After obtaining the time-frequency representation, a time-frequency mask can be built. Based on the definition of cIRM and IRM, in SRS representations, two versions of SRS-mask can be defined: cIRM-like $cIRM_{srs}$ and IRM-like IRM_{srs}, in which X_{srs}, V_{srs}, and S_{srs} are representations of the time-domain noisy speech x, noise v, and clean speech s. Using the time-frequency SRS-mask $\hat{M}_{srs}$, the denoised speech signal can be obtained using $\hat{S} = \hat{M}_{srs} \otimes Y_{srs}$, where $\otimes$ denotes the Kronecker product of two matrices.

A multi-objective loss function based on the parameters in PESQ was proposed in (Martín-Doñas et al., 2018). It was shown that MSE could not follow the human auditory system properly. Hence, the PESQ algorithm, which is the most widely used metric for speech quality evaluation, was adopted as a loss function. In fact, the symmetric and asymmetric disturbance terms in PESQ was incorporated into the loss function for training. These terms were computed frame-by-frame based on the clean and noisy signals.

In the LPS domain, after mean and variance normalization, the commonly used MSE loss function can be expressed as follows

$$E = \frac{1}{K} \sum_k \left(\frac{\log |S_{k,l}|^2 - \mu_k}{\sigma_k} - \frac{\log |\hat{S}_{k,l}|^2 - \mu_k}{\sigma_k} \right)^2 = \frac{1}{K} \sum_k \frac{1}{\sigma^2} \left(log \frac{|S_{k,l}|^2}{|\hat{S}_{k,l}|^2} \right)^2 \tag{45}$$

As seen in this loss function, the loudness and threshold effects are not taken into consideration when optimizing a deep neural network.

To incorporate the perceptual features mentioned above, two disturbance terms, namely $D_t^{(s)}$ and $D_t^{(a)}$, were added to the loss function. The term $D_t^{(s)}$ considers the absolute difference between the denoised and clean loudness spectra, while the term $D_t^{(a)}$ is computed from the symmetrical disturbance but weighting positive and negative loudness differences differently. Thus, the final loss function is defined as

$$J = \frac{1}{L} \sum_l \left(E + \alpha_{D_s} D_t^{(s)} + \alpha_{D_a} D_t^{(a)} \right) \tag{46}$$

where α_{D_s} and α_{D_a} are the weighting factors and L is the number of frames. Eq. (46) can be viewed as a multi-objective optimization function in which the MSE and PESQ-based disturbance terms are minimized at the same time.

An SNR-aware CNN-based single channel speech denoising was presented in (Fu et al., 2016, September), where it was shown that the speech denoising performance is corrupted by the mismatch of training and testing for different noise types and SNRs (Qian et al., 2017, August). Several attempts have been made to classify noise types before feeding the noisy speech into a DNN. However, it would be difficult to generalize this capability due to the limitations associated with this classification problem. Noise Aware Training (NAT) has been proposed to incorporate the noise information into the input features. To cope with the noise type, two SNR-aware algorithms have been proposed allowing the denoising DNN model to achieve better speech denoising performance. First, an MTL framework is applied to approximate the clean speech signal together with the SNR level for a noisy input speech. By considering MTL, the trained CNN model is made aware of the SNR level. Second, a SNR Adaptive Denoising (SNR-AD) is introduced which includes an offline and an online mode. In the offline mode, several SNR-specific denoising models are prepared in which each model is trained by the noisy/clean pairs using the noisy speech within a particular range of SNR. Estimating the SNR is a regression problem while predicting the noise type is a classification problem. In the online mode, SNR-AD first predicts the SNR level and then selects the CNN model trained in that range.

To incorporate the ability of the SNR prediction into the denoising model, CNNs can jointly predict the primary LPS features and SNR level

of the noisy input. Mathematically, the loss function is modified to incorporate the SNR estimation capability as follows

$$J = \frac{1}{N}\left[\sum_{l=1}^{N}\left\|S_l - \hat{S}_l\right\|^2 + \lambda_{SNR}\sum_{l=1}^{N}\left(O_l - \hat{O}_l\right)^2\right] \tag{47}$$

where O_l and $\hat{O}_l$ represent the actual and estimated SNR levels of the noisy input frame at index l, respectively, and λ_{SNR} is a weighting factor.

In the second algorithm, the idea of using different denoising models is exploited based on the strength of noise. Before applying the proper denoising model, a decision is made by comparing the estimated SNR to some predefined thresholds. The thresholds are specified to be

$$m = \begin{cases} 1; & \hat{O}_l > \rho_1 \\ u; & \rho_{u-1} > \hat{O}_l > \rho_u, \ \forall U > l > 1 \\ U; & \rho_{u-1} > \hat{O}_l \end{cases} \tag{48}$$

where $\hat{O}_l$ is the estimated SNR level, U indicates the total number of denoising models, m corresponds to the m-th denoising model, and ρ_u is the l-th threshold.

In another MTL work in (Fu et al., 2017, September), a CNN-based complex spectrogram enhancement was proposed. It has been stated in (Paliwal, 1987, April) that by increasing the window overlap and length of the Fourier transform, the importance of phase grows. Also, as described in the joint optimization-based speech denoising methods, the phase information plays a critical rule in low SNRs. The reason for this is that signals in the Fourier domain are complex, which can be described either in terms of real-imaginary parts or magnitude-phase parts. Consider the real-imaginary parts of a noisy speech as N_r and N_i, respectively. The phase information can be expressed as

$$\arctan\frac{X_i}{X_r} = \arctan\frac{S_i + V_i}{S_r + V_r} \tag{49}$$

In case of high SNR values, i.e. $|S_i| \gg |V_i|$, $|S_r| \gg |V_r|$, Eq. (49) can be simplified to

$$\arctan\frac{X_i}{X_r} = \arctan\frac{S_i}{S_r} \tag{50}$$

This indicates that the importance of the phase information grows when dealing with low SNRs.

The real-imaginary parts of the spectrogram were used as an objective function as follows

$$\begin{aligned} \Theta &= \sum\left\|\hat{S}_i - S_i\right\|_2^2 + \left\|\hat{S}_r - S_r\right\|_2^2 \\ &= \sum\left\|\hat{S}_v - S_{v2}^2\right\| \end{aligned} \tag{51}$$

where $\hat{S}_v = [\hat{S}_r\ \hat{S}_i]^T$ and $S_v = [S_r\ S_i]^T$ denote the vertically cascaded vectors of the clean and enhanced real-imaginary spectrograms, respectively. Minimizing Eq. (51) is equivalent to maximizing Segmented SNR (SSNR) (Quackenbush et al., 1988) of the denoised signal. In this approach, it was proposed to incorporate the Log Spectral Distance (LSD) (Benesty et al., 2005) into the objective function to improve speech quality. Explicitly, the LSD was modeled as an LPS reconstruction term expressed as

$$LSD = \left\|\log\left(\hat{S}_i^2 + \hat{S}_r^2\right) - \log\left(S_i^2 + S_r^2\right)\right\|_2^2 \tag{52}$$

Then, Eq. (52) was combined with Eq. (51) to form the following formula

$$\Theta = \sum\alpha_{S_v}\left\|\hat{S}_v - S_v\right\|_2^2 + \alpha_{LSD}\left\|\log\left(\hat{S}_i^2 + \hat{S}_r^2\right) - \log\left(S_i^2 + S_r^2\right)\right\|_2^2 \tag{53}$$

with the weighting terms indicated by α_{S_v} and α_{LSD}. Because the first term of Eq. (53) reflects the maximization of SSNR while the second term reflects the minimization of LSD metric, this speech denoising is of the MTL type.

Since the phase information has unpredictable behavior, the deep learning models cannot learn the relationship between the clean and noisy phase. Williamson et al. (Ephraim and Malah, 1985) justified the similarity between the structures in real-imaginary spectrograms and magnitude spectrogram. Hence, instead of using the phase information directly, the spectrograms were used to compute the objective function (Fu et al., 2016). In this approach, an extension of Eq. (53) was considered. In fact, the raw waveforms were incorporated into the objective function to improve speech quality. Mathematically, Eq. (53) was modified as follows

$$\begin{aligned} \Theta &= \sum\alpha_{S_v}\left\|\hat{S}_v - S_v\right\|_2^2 + \alpha_{LSD}\left\|\log\left(\hat{S}_i^2 + \hat{S}_r^2\right) - \log\left(S_i^2 + S_r^2\right)\right\|_2^2 \\ &\quad + \alpha_{wave}\left\|\hat{w}_S - w_S\right\|_2^2 \end{aligned} \tag{54}$$

where $\hat{w}_S$ and w_S are the clean and enhanced signals, respectively, and α_{wave} is a weighting parameter for the last term. The last term in Eq. (54) can be expressed using the inverse discrete Fourier transform (IDFT) as a function of y and $\hat{y}$ as follows

$$\begin{aligned} \alpha_{wave}\left\|\hat{w}_S - w_S\right\|_2^2 &= \\ \alpha_{wave}\left\|\left(CU_1\hat{S}_r - DU_2\hat{S}_i\right) - \left(CU_1S_r - DU_2S_i\right)\right\|_2^2 \\ &= \alpha_{wave}\left\|F\hat{S}_v - FS_v\right\|_2^2 \end{aligned} \tag{55}$$

where U_1, U_2 are the matrices used for the recovery of the even symmetry of the real part, and the odd symmetry of the imaginary part, respectively, C and D are the cosine and sine matirces in the IDFT, and F is defined as

$$F = \begin{bmatrix} CU_1 - DU_2 \end{bmatrix} \tag{56}$$

Hence, the objective function is rewritten as

$$\begin{aligned} \Theta &= \sum\alpha_S\left\|\hat{S}_v - S_v\right\|_2^2 + \alpha_{LSD}\left\|\log\left(\hat{S}_i^2 + \hat{S}_r^2\right) - \log\left(S_i^2 + S_r^2\right)\right\|_2^2 \\ &\quad + \alpha_{wave}\left\|F\hat{S}_v - FS_v\right\|_2^2 \end{aligned} \tag{57}$$

It can be seen that all parts of Eq. (57) are directly related to the output vector S_v.

A shortcoming of the mapping-based and masking-based methods is that they do not consider the phase mismatch problem. To address this issue, a Phase Sensitive Mask (PSM) was proposed in (Gao et al., 2018, April). Here, it is worth mentioning that the PSM values are unbounded and thus the output values of the PSM estimation are normally truncated. In this method, a joint learning algorithm was proposed to model the PSM parameters explicitly. In comparison to the conventional algorithms which implicitly learn the T-F mask, this work focused on predicting the mask by an explicit framework of its variables.

Mathematically, the following approximation of PSM was first computed under the assumption that clean speech and noise are uncorrelated

$$M_{k,l}^{PSM} \approx \sqrt{\frac{|S_{k,l}|^2}{|S_{k,l}|^2 + |V_{k,l}|^2}}\cos\theta_{k,l}^{SX} \stackrel{\Delta}{=} M_{k,l}^{aPSM} \tag{58}$$

where θ^{SX} indicates the difference of the clean and noisy phase. The approximated PSM (aPSM) was described as a function of three parameters: (1) Clean speech magnitude, (2) noisy speech magnitude, (3) phase difference between clean and noisy speech spectra. A network was jointly optimized by the following loss functions

$$L_{aPSM} = \sum_{k,l}\left[\hat{m}_{k,l}^{aPSM}|X_{k,l}| - |S_{k,l}|\cos\theta_{k,l}^{SX}\right]^2 \tag{59}$$

$$L_{mag} = \sum_{k,l}\left[\left|\hat{S}_{k,l}\right| - |S_{k,l}|\right]^2 + \left[\left|\hat{V}_{k,l}\right| - |V_{k,l}|\right]^2 \tag{60}$$

$$L_{pha} = \sum_{k,l}\left[\cos\hat{\theta}_{k,l}^{XS} - \cos\theta_{k,l}^{XS}\right]^2 \tag{61}$$

Table 1
Average performance metrics of representative speech denoising methods in non-speech type noise (Factory – N_1) and speech type noise (Babble – N_2) at -2 dB SNR.

Metric Method	STOI (N_1)	STOI (N_2)	PESQ (N_1)	PESQ (N_2)	LSD (N_1)	LSD (N_2)	SSNR (N_1)	SSNR (N_2)
Noisy Signal	0.6235	0.6086	1.3057	1.3114	2.3968	1.9121	-4.4675	-3.8066
Wiener	0.5929	0.5450	1.4301	1.2305	1.5489	1.4658	0.2607	-0.6993
IBM	0.6894	0.6741	1.3694	1.3354	1.7839	1.5469	1.0894	1.1152
DNN – IBM	0.7366	0.7119	1.4209	1.3737	1.3953	1.2860	0.4422	0.6157
DNN – IRM	0.7577	0.7283	1.6558	1.4961	1.0895	1.0649	0.5055	0.3867
DNN – Hybrid	0.8010	0.7542	1.6850	1.5813	0.9157	0.8812	0.6024	0.7312
DNN – MO	0.8159	0.7614	1.7068	1.6573	0.8556	0.8331	0.8011	0.7556

Table 2
Average performance metrics of representative speech denoising methods in non-speech type type (Factory – N1) and speech type noise (Babble – N2) at 0 dB SNR.

Metric Method	STOI (N_1)	STOI (N_2)	PESQ (N_1)	PESQ (N_2)	LSD (N_1)	LSD (N_2)	SSNR (N_1)	SSNR (N_2)
Noisy Signal	0.6689	0.6585	1.3607	1.3689	2.3088	1.8318	-3.6128	-2.9568
Wiener Filter	0.6869	0.6040	1.5274	1.3041	1.4481	1.3931	0.8107	-0.1013
IBM	0.7031	0.6968	1.4052	1.4012	1.7642	1.5125	0.5024	0.6587
DNN – IBM	0.7831	0.7567	1.5740	1.4368	1.3538	1.3125	2.0030	1.8394
DNN – IRM	0.7991	0.7758	1.8233	1.6141	0.9668	0.9709	1.5943	1.5159
DNN – Hybrid	0.8312	0.8173	1.9547	1.8333	0.8123	0.8206	2.1415	1.8012
DNN – MO	0.8450	0.8234	2.1005	1.9673	0.7315	0.7519	1.9614	1.9684

Table 3
Average performance metrics of representative speech denoising methods in non-speech type noise (Factory – N1) and speech type noise (Babble – N2) at 5 dB SNR.

Metric Method	STOI (N_1)	STOI (N_2)	PESQ (N_1)	PESQ (N_2)	LSD (N_1)	LSD (N_2)	SSNR (N_1)	SSNR (N_2)
Noisy Signal	0.7837	0.7788	1.5424	1.5504	2.0065	1.5628	-0.6908	-0.0727
Wiener Filter	0.7878	0.7849	1.8624	1.5558	1.3534	1.2688	2.4677	1.8805
IBM	0.8046	0.7975	1.7056	1.6652	1.5193	1.4658	2.1259	2.0126
DNN – IBM	0.8657	0.8457	1.9607	1.7420	1.4568	1.3780	5.2782	5.0936
DNN – IRM	0.8841	0.8627	2.2386	2.0020	0.8084	0.7725	4.4368	4.2259
DNN – Hybrid	0.9111	0.8827	2.3071	2.1101	0.7130	0.6852	5.5115	5.3230
DNN – MO	0.9362	0.9135	2.5403	2.3114	0.6823	0.6533	6.4371	6.1123

In addition, since the MSE does not accurately match the human auditory perception as related to speech intelligibility, a nonlinear magnitude warping technique was considered in the estimation steps. Consequently, Eq. (60) was rewritten as

$$L_{mag}^{warp} = \sum_{t,f} \left[\left| \hat{S}_{k,l} \right|^\alpha - \left| S_{k,l} \right|^\alpha \right]^2 + \left[\left| \hat{V}_{k,l} \right|^\alpha - \left| V_{k,l} \right|^\alpha \right]^2 \tag{62}$$

To further improve performance, the sum of clean and noise estimates were minimized with respect to a given mixture, that is

$$\left| \hat{X}_{k,l} \right|^2 = \left| \hat{S}_{k,l} \right|^2 + \left| \hat{V}_{k,l} \right|^2 + 2 \cos \theta_{k,l}^{SV} \left| \hat{S}_{k,l} \right| \left| \hat{V}_{k,l} \right| \tag{63}$$

$$L_{add}^{warp} = \sum_{t,f} \left[\left| \hat{X}_{k,l} \right|^\alpha - \left| X_{k,l} \right|^\alpha \right]^2 \tag{64}$$

Hence, the final objective function was defined to be

$$L_{Fin} = \gamma_1 L_{aPSM} + \gamma_2 L_{mag}^{warp} + \gamma_3 L_{pha} + \gamma_4 L_{add}^{warp} \tag{65}$$

where γ_1, γ_2, γ_3 and γ_4 indicate the weight values.

4. Comparison results

In this section, experiments were conducted to compare the performance of representative methods of the speech denoising methods reviewed in this paper. These representative methods were chosen because their codes are publicly available. Four extensively used metrics of Short-Time Objective Intelligibility (STOI) (Taal et al., 2011), Perceptual Evaluation of Speech Quality (PESQ) (ITU-T P.862 2001), Segmented

Signal-to-Noise Ratio (SSNR) (Quackenbush et al., 1988), and Log Spectral Distance (LSD) (Benesty et al., 2005) were employed to compare the methods in an objective manner. The same public domain dataset of IEEE (McCloy et al., 2013) was used in the experiments. The IEEE Corpus includes 3600 speech wave (.wav) files by 20 speakers (10 females and 10 males) in which each file is about 2 seconds long. The speakers are from two regions of the Pacific Northwest (PN) and the Northern Cities (NC) reading the IEEE "Harvard" sentences. Noisy dataset was made by adding two types of noises (speech type noise (Babble) and non-speech type noise (Factory) noise signals) to the clean dataset. Three levels of SNR were made as -2, 0, and 5 dB. Both noise types were used for training, evaluation, and testing parts. The LPS features were considered as the input using 256-point FFT and 50% overlap. The dataset was randomly partitioned into three parts with no overlap: 70% for training, 20% for testing, and 10% for validation.

To provide a comprehensive comparison, the conventional methods of Wiener filtering (Lim and Oppenheim, 1979) and IBM (Tu et al., 2019, May), the single objective deep learning methods of DNN-IBM (Tu et al., 2019, May), DNN-IRM (Hummersone et al., 2014), the hybrid method of DNN-Hybrid (Nie et al., 2018), and the multi-objective method of DNN-MultiObjective (DNN-MO) (Fu et al., 2016) were examined considering the availability of their codes in the public domain.

Table 1 through Table 3 exhibit the performance of the denoised signals in terms of the above four metrics for three SNR levels. As can be seen from these tables, the DNN-Hybrid and DNN-MO methods provided superior outcomes. This was expected since they utilize more information to optimize a deep neural network. Another key observation from

these tables is that in lower SNR levels, the multi-objective method performed better than the DNN-Hybrid method. Basically, as the SNR level becomes higher, the clean speech characteristics can be obtained well using the conventional methods. In other words, at higher SNRs, the difference between the DNN-Hybrid and DNN-MO method becomes small or rather negligible.

In general, since the multi-objective deep learning methods use richer information to achieve speech denoising, their performance in comparison to the other methods is better. In fact, what plays a critical role in achieving better performance is the terms selected for the optimization. Another important issue that is worth noting here is that environmental noises cannot be modeled accurately by the conventional methods. However, the deep learning methods can model the nonlinear relationship between noisy speech and clean speech signals more effectively and thus their performance in general is better even in not previously encountered noisy environments.

5. Possible future directions

It is important to note that all the above deep learning methods are supervised methods, meaning that clean speech signals are assumed to be available for their training. In these methods, clean speech signals are used as the target. In practice, speech sentences spoken in the field differ from database speech sentences that are used for training. As a result, training on one set of speech signals does not guarantee the same performance when a different set of speech signals are spoken in the same noise environments. In real-world scenarios, clean speech signals are in fact not available. In (Alamdari et al., 2019), a speech denoising method was introduced which does not rely on the availability of clean speech signals to achieve training and it uses the mixture for both the input and the target of a deep neural network. This method eases the major assumption in the existing speech denoising methods and allows the training process to be conducted in an online or self-supervised manner without having access to clean speech signals. We envision that much progress can be made by developing different variations of the method described in (Alamdari et al., 2019).

Also, in the existing speech denoising methods, all noises are treated the same. We envision another future improvement can be achieved by designing deep neural networks for different types of noise and thus using a bank of deep neural networks depending on the noise characteristics encountered in the field. Since in practice or real-world scenarios, there are many different noise types, unsupervised noise classification approaches such as the one in (Alamdari and Kehtarnavaz, 2019, January) can be used to train these deep neural networks in an online manner depending on the noise type encountered in the field.

6. Conclusion

A review of multi-objective deep learning speech denoising methods has been covered in this paper. These recent methods denote the current state-of-the-art in speech denoising. To set the stage for this review, an overview of conventional, single objective deep learning, and hybrid methods was first presented. This overview was followed by a review of the mathematical framework of the existing multi-objective deep learning methods. Representative methods in the speech denoising categories whose codes are publicly available were then compared by considering the same public domain dataset and four widely used objective metrics. The comparison conducted has shown the effectiveness of the multi-objective deep learning methods, in particular in low SNRs as compared with the other existing methods.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2020.04.002.

References

Alamdari, N., Kehtarnavaz, N., 2019, January. A real-time smartphone app for unsupervised noise classification in realistic audio environments. In: 2019 IEEE International Conference on Consumer Electronics, pp. 1–5.

Alamdari, N., Azarang, A. and Kehtarnavaz, N., 2019. Self-supervised deep learning-based speech denoising. arXiv preprint arXiv:1904.12069.

Benesty, J., Makino, S., Chen, J., 2005. Speech Enhancement. Springer Science & Business Media.

Bennett, I.M., Phoenix Solutions Inc, 2010. *Network based interactive speech recognition system.* U.S. Patent 7,831,426.

Besacier, L., Grassi, S., Dufaux, A., Ansorge, M., Pellandini, F., 2000, June. GSM speech coding and speaker recognition. In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100*, pp. II1085-II1088.

Boll, S., Pulsipher, D., 1980. Suppression of acoustic noise in speech using two microphone adaptive noise cancellation. IEEE Trans. Acoust. Speech Signal Process. 28 (6), 752–753.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27 (2), 113–120.

Brown, C.A., Bacon, S.P., 2010. Fundamental frequency and speech intelligibility in background noise. Hear. Res. 266 (1–2), 52–59.

Chai, L., Du, J., Wang, Y.N., 2017, September. Gaussian density guided deep neural network for single-channel speech enhancement. In: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing, pp. 1–6.

Chaudhari, A., Dhonde, S.B., 2015, January. A review on speech enhancement techniques. In: 2015 International Conference on Pervasive Computing, pp. 1–3.

Cohen, I., Berdugo, B., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9 (1), 12–15.

Cornelis, B., Moonen, M., Wouters, J., 2010. Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. IEEE Transactions on. Audio, Speech, and Language Lang. ProcessingProcess. 19 (5), 1368–1381.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 20 (1), 30–42.

Dat, T.H., Takeda, K., Itakura, F., 2005, March. Generalized gamma modeling of speech and its online estimation for speech enhancement. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. iv–181.

Donahue, C., Li, B., Prabhavalkar, R., 2018, April. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5024–5028.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J., 2015, April. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 708–712.

Fakoor, R., He, X., Tashev, I., Zarar, S., 2018, April. Constrained convolutional-recurrent networks to improve speech quality with low impact on recognition accuracy. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3011–3015.

Fu, S.W., Tsao, Y., Lu, X., 2016, September. SNR-aware convolutional neural network modeling for speech enhancement. In: Interspeech, pp. 3768–3772.

Fu, S.W., Hu, T.Y., Tsao, Y. and Lu, X., 2016. Multi-metrics learning for speech enhancement.

Fu, S.W., Tsao, Y., Lu, X., Kawai, H., 2017, December. Raw waveform-based speech enhancement by fully convolutional networks. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 006–012.

Fu, S.W., Hu, T.Y., Tsao, Y., Lu, X., 2017, September. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: Proceedings IEEE 27th International Workshop on Machine Learning for Signal Processing, pp. 1–6.

Gao, T., Du, J., Dai, L.R., Lee, C.H., 2016. SNR-based progressive learning of deep neural network for speech enhancement. In: INTERSPEECH, pp. 3713–3717.

Gao, T., Du, J., Dai, L.R., Lee, C.H., 2018, April. Densely connected progressive learning for lstm-based speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5054–5058.

Ghosh, P.K., Tsiartas, A., Narayanan, S., 2010. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio, Speech, Lang. Process. 19 (3), 600–613.

Graves, A., Mohamed, A.R., Hinton, G., 2013, May. Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6645–6649.

Han, W., Zhang, X., Min, G., Zhou, X., Zhang, W., 2016, June. Perceptual weighting deep neural networks for single-channel speech enhancement. In: 2016 12th World Congress on Intelligent Control and Automation, pp. 446–450.

Hermus, K., Wambacq, P., 2006. A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Adv. Signal Process. 2007 (1), 045821.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hu, G., Wang, D., 2001, October. Speech segregation based on pitch tracking and amplitude modulation. In: Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575), pp. 79–82.

Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 23 (12), 2136–2147.

Hughes, T.B., Kim, H.S., DiBiase, J.H., Silverman, H.F., 1999. Performance of an HMM speech recognizer using a real-time tracking microphone array as input. IEEE Trans. Speech Audio Process. 7 (3), 346–349.

Hummersone, C., Stokes, T., Brookes, T., 2014. On the ideal ratio mask as the goal of computational auditory scene analysis. In: Blind Source Separation. Springer, Berlin, Heidelberg, pp. 349–368.

Isabelle, S.H., Verance Corp, 2000. Noise suppression for low bitrate speech coder. U.S. Patent 6,122,610.

ITU-T P.862, "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrowband telephone network and speech coders," 2001.

Kamath, S., Loizou, P., 2002, May. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 4 44164–44164.

Kemp, T., Waibel, A., 1999. Unsupervised training of a speech recognizer: Recent experiments. Sixth European Conference on Speech Communication and Technology.

Kolbæk, M., Tan, Z.H., Jensen, S.H. and Jensen, J., 2019. On loss functions for supervised monaural time-domain speech enhancement. arXiv preprint arXiv:1909.01019.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Adv. Neural Inf. Process. Syst. 1097–1105.

Lai, Y.H., Zheng, W.Z., 2019. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. Biomed. Signal Process. Control 48, 35–45.

Lee, J., Skoglund, J., Shabestary, T., Kang, H.G., 2018. Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. IEEE Signal Process Lett. 25 (8), 1276–1280.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Transactions on. Audio, Speech, and Language Lang. ProcessingProcess. 22 (4), 745–777.

Li, A., Zheng, C. and Li, X., 2019. Convolutional recurrent neural network based progressive learning for monaural speech enhancement. arXiv preprint arXiv:1908.10768.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Liu, D., Smaragdis, P., Kim, M., 2014. Experiments on deep learning for speech denoising. Fifteenth Annual Conference of the International Speech Communication Association.

Liu, Y., Zhang, H., Zhang, X., 2018. Using shifted real spectrum mask as training target for supervised speech separation. In: Interspeech, pp. 1151–1155.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC press.

Longueira, F. and Keene, S., 2018. A fully convolutional neural network approach to end-to-end speech enhancement. arXiv preprint arXiv:1807.07959.

Lu, X., Tsao, Y., Matsuda, S., Hori, C., 2013, August. Speech enhancement based on deep denoising autoencoder. In: Interspeech, pp. 436–440.

Ma, G., Gran, F., Jacobsen, F., Agerkvist, F.T., 2010. Adaptive feedback cancellation with band-limited LPC vocoder in digital hearing aids. IEEE Transactions on. Audio, Speech, and Language Lang. ProcessingProcess. 19 (4), 677–687.

Martín-Doñas, J.M., Gomez, A.M., Gonzalez, J.A., Peinado, A.M., 2018. A deep learning loss function based on the perceptual evaluation of the speech quality. IEEE Signal Process. Lett. 25 (11), 1680–1684.

Martin, R., 2002, May. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, pp. I–253.

D. R. McCloy, P. E. Souza, R. A. Wright, J. Haywood, N. Gehani, and S. Rudolph. (2013). The PN/NC Corpus. Version 1.0. [Online]. Available:https://depts.washington.edu/phonlab/resources/pnnc/pnnc1/.

Mirbagheri, M., Xu, Y., Akram, S., Shamma, S.A., 2013. Speech enhancement using convolutive nonnegative matrix factorization with cosparsity regularization. In: INTERSPEECH, pp. 456–459.

Mirsamadi, S., Tashev, I., 2016, September. Causal speech enhancement combining data–driven learning and suppression rule estimation. In: Interspeech, pp. 2870–2874.

Mowlaee, P., Saeidi, R., 2013. Iterative closed-loop phase-aware single-channel speech enhancement. IEEE Signal Process. Lett. 20 (12), 1235–1239.

Narayanan, A., Wang, D., 2013, May. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092–7096.

Nie, S., Liang, S., Liu, B., Zhang, Y., Liu, W., Tao, J., 2018. Deep noise tracking network: a hybrid signal processing/deep learning approach to speech enhancement. In: Interspeech, pp. 3219–3223.

Paliwal, K., 1987, April. Estimation of noise variance from the noisy AR signal and its application in speech enhancement. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 12, pp. 297–300.

Pandey, A., Wang, D., 2019, May. TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6875–6879.

Pandey, A., Wang, D., 2019. A New Framework for CNN-Based Speech Enhancement in the Time Domain. IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP) 27 (7), 1179–1188.

Park, S.R. and Lee, J., 2016. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

Pascual, S., Serrà, J., Bonafonte, A., 2019. Time-domain speech enhancement using generative adversarial networks. Speech Commun. 114, 10–21.

Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F., 2001. Recognizing voice over IP: a robust front-end for speech recognition on the World Wide Web. IEEE Trans. Multimedia 3 (2), 209–218.

Premananda, B.S., Uma, B.V., 2013. Speech enhancement algorithm to reduce the effect of background noise in mobile phones. Int. J. Wirel. Mob. Netw. 5 (1), 177.

Qian, K., Zhang, Y., Chang, S., Yang, X., Florêncio, D., Hasegawa-Johnson, M., 2017, August. Speech enhancement using bayesian wavenet. In: Interspeech, pp. 2013–2017.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Englewood Cliffs, NJ: PrenticeHall.

Saleem, N., Irfan, M., Chen, X., Ali, M., 2018, June. Deep neural network based supervised speech enhancement in speech-babble noise. In: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science, pp. 871–874.

Sameti, H., Sheikhzadeh, H., Deng, L., Brennan, R.L., 1998. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans. Audio Speech Process. 6 (5), 445–455.

Sarampalis, A., Kalluri, S., Edwards, B., Hafter, E., 2009. Objective measures of listening effort: effects of background noise and noise reduction. J. Speech Lang. Hear. Res..

Scalart, P., 1996, May. Speech enhancement based on a priori signal to noise estimation. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 2, pp. 629–632.

Seok, J.W., Bae, K.S., 1997, April. Speech enhancement with reduction of noise components in the wavelet domain. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1323–1326.

Shivakumar, P.G., Georgiou, P.G., 2016. Perception optimized deep denoising autoencoders for speech enhancement. In: Interspeech, pp. 3743–3747.

Stone, M.A., Füllgrabe, C., Moore, B.C., 2012. Notionally steady background noise acts primarily as a modulation masker of speech. J. Acoust. Soc. Am. 132 (1), 317–326.

Sulong, A., Gunawan, T.S., Khalifa, O.O., Kartiwi, M., Dao, H., 2017. Single channel speech enhancement using wiener filter and compressive sensing. Int. J. Electr. Comput. Eng. 7 (4), 1941.

Sun, M., Zhang, X., Zheng, T.F., 2016. Unseen noise estimation using separable deep auto encoder for speech enhancement. IEEE/ACM Trans. Audio, Speech, Lang. Process. 24 (1), 93–104.

Sun, L., Du, J., Dai, L.R., Lee, C.H., 2017, March. Multiple-target deep learning for LSTM-RNN based speech enhancement. In: 2017 Hands-free Speech Communications and Microphone Arrays, pp. 136–140.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on. Audio, Speech, and Language Lang. ProcessingProcess. 19 (7), 2125–2136.

Tu, Y.H., Tashev, I., Zarar, S., Lee, C.H., 2018, April. A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2531–2535.

Tu, Y.H., Du, J., Lee, C.H., 2019, May. DNN training based on classic gain function for single-channel speech enhancement and recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 910–914.

Tu, M., Tang, Y., Huang, J., He, X. and Zhou, B., 2019. Towards adversarial learning of speaker-invariant representation for speech emotion recognition. *arXiv preprint arXiv:1903.09606*.

Valin, J.M., 2018, August. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing, pp. 1–5.

Varga, A., Moore, R.K., 1990, April. Hidden Markov model decomposition of speech and noise. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 845–848.

Venkataramani, S. and Smaragdis, P., 2018. End-to-end networks for supervised single-channel speech separation. arXiv preprint arXiv:1810.02568.

Venkataramani, S., Higa, R., Smaragdis, P., 2018, November. Performance based cost functions for end-to-end speech separation. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 350–355.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11 (Dec), 3371–3408.

Vu, N.T., Lyu, D.C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.S., Schultz, T., Li, H., 2012, March. A first speech recognition system for Mandarin-in-English code-switch conversational speech. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4889–4892.

Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio, Speech, Lang. Process. 26 (10), 1702–1726.

Wang, D., Kjems, U., Pedersen, M.S., Boldt, J.B., Lunner, T., 2009. Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Am. 125 (4), 2336–2347.

Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (12), 1849–1858.

Wang, Y., Du, J., Dai, L.R., Lee, C.H., 2017. A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation. In: Interspeech, pp. 1178–1182.

Wang, Q., Du, J., Dai, L.R., Lee, C.H., 2018. A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures. IEEE/ACM Trans. Audio, Speech, Lang. Process. 26 (7), 1181–1193.

Wang, Z.Q., Roux, J.L., Wang, D. and Hershey, J.R., 2018. End-to-end speech separation with unfolded iterative phase reconstruction. arXiv preprint arXiv:1804.10204.

Williamson, D.S., Wang, Y., Wang, D., 2016, March. Complex ratio masking for joint enhancement of magnitude and phase. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5220–5224.

Xia, B., Bao, C., 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun. 60, 13–29.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2013. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett. 21 (1), 65–68.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2014. Dynamic noise aware training for speech enhancement based on deep neural networks. Fifteenth Annual Conference of the International Speech Communication Association.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio, Speech, Lang. Process. 23 (1), 7–19.

Xu, Y., Du, J., Huang, Z., Dai, L.R. and Lee, C.H., 2017. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement. *arXiv preprint arXiv:1703.07172*.

Yang, L.P., Fu, Q.J., 2005. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. J. Acoust. Soc. Am. 117 (3), 1001–1004.

Ying, D., Yan, Y., Dang, J., Soong, F.K., 2011. Voice activity detection based on an unsupervised learning framework. IEEE Transactions on. Audio, Speech, and Language Lang. ProcessingProcess. 19 (8), 2624–2633.

Zhang, X.L., Wu, J., 2012. Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. 21 (4), 697–710.

Zhang, H., Zhang, X., Gao, G., 2018, April. Training supervised speech separation system to improve stoi and pesq directly. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5374–5378.

Zhao, Y., Xu, B., Giri, R., Zhang, T., 2018, April. Perceptually guided speech enhancement using deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5074–5078.

Zhao, Y., Wang, Z.Q., Wang, D., 2018. Two-stage deep learning for noisy-reverberant speech enhancement. IEEE/ACM Trans. Audio, Speech, Lang. Process. 27 (1), 53–62.