Original Article

Trends in Hearing
Volume 27: 1–52
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165231209913
journals.sagepub.com/home/tia
Sage

Sixty Years of Frequency-Domain Monaural Speech Enhancement: From Traditional to Deep Learning Methods

Chengshi Zheng[1,2,*], Huiyong Zhang[1,2], Wenzhe Liu[1,2], Xiaoxue Luo[1,2], Andong Li[1,2], Xiaodong Li[1,2] and Brian C. J. Moore[3,*]

Abstract

Frequency-domain monaural speech enhancement has been extensively studied for over 60 years, and a great number of methods have been proposed and applied to many devices. In the last decade, monaural speech enhancement has made tremendous progress with the advent and development of deep learning, and performance using such methods has been greatly improved relative to traditional methods. This survey paper first provides a comprehensive overview of traditional and deep-learning methods for monaural speech enhancement in the frequency domain. The fundamental assumptions of each approach are then summarized and analyzed to clarify their limitations and advantages. A comprehensive evaluation of some typical methods was conducted using the WSJ + Deep Noise Suppression (DNS) challenge and Voice Bank + DEMAND datasets to give an intuitive and unified comparison. The benefits of monaural speech enhancement methods using objective metrics relevant for normal-hearing and hearing-impaired listeners were evaluated. The objective test results showed that compression of the input features was important for simulated normal-hearing listeners but not for simulated hearing-impaired listeners. Potential future research and development topics in monaural speech enhancement are suggested.

Keywords

speech enhancement, speech dereverberation, multistage learning, noise estimation, deep complex network

Received 2 December 2022; accepted 9 October 2023

Introduction

The aim of monaural speech enhancement is to extract clean speech or to improve the speech-to-background ratio by removing noise and reverberation from a noisy-reverberant speech signal captured by a microphone (Lim, 1983; Benesty et al., 2006; Martin et al., 2008; Naylor & Gaubitch, 2010; Loizou, 2013). This is especially important for improving speech quality and/or intelligibility for digital speech communication devices, such as hearing aids and other assistive listening devices (Dillon, 2012; Popelka et al., 2016), audio-visual conference equipment, smartphones, and true wireless earphones. It is also important for automatic speech recognition systems (Hansen, 1996; Li et al., 2014), such as smart home appliances, in-car voice assistants, and speech transcription services. It should be noted that it is not always necessary to reconstruct the time-domain speech signal for the purpose of automatic speech recognition; enhancement of extracted features often leads to improve speech recognition performance (Schuller et al., 2009; Krueger & Haeb-Umbach, 2010). The present paper focuses on monaural speech enhancement for human hearing rather than for machine hearing, where the time-domain waveform of the clean speech needs to be reconstructed.

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Cambridge Hearing Group, Department of Psychology, University of Cambridge, Cambridge, UK

[*]Chengshi Zheng, Xiaoxue Luo, and Brian C. J. Moore contributed equally to this work as first authors.

Corresponding Author:
Chengshi Zheng, Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190, Beijing, China; University of Chinese Academy of Sciences, 100049, Beijing, China.
Email: cszheng@mail.ioa.ac.cn

Both time- and frequency-domain methods of monaural speech enhancement have been proposed and widely studied. For the former, the clean speech is estimated directly in the time domain without (short-term) spectral analysis and synthesis (Lee & Jung, 2000; Benesty & Chen, 2011; Luo & Mesgarani, 2018; Macartney & Weyde, 2018; Pandey & Wang, 2018, 2019b; Hao et al., 2019; Pandey & Wang, 2019a; Von Neumann et al., 2020; Zucatelli & Coelho, 2021; Pandey & Wang, 2022). For the latter, the short-term complex spectrum of the clean speech is estimated, the spectrum is converted back to a time-domain signal, and this process is repeated for a series of overlapping frames (time segments) to reconstruct the complete time-domain signal, using the overlap-add method (Allen, 1977; Boll, 1979; Ephraim & Malah, 1984; Griffin & Lim, 1984; Loizou, 2013; Wang & Chen, 2018). There are some hybrid methods, in which the appropriate gain for each of several frequency sub-bands is estimated in a first stage, and a time-domain enhancement filter is designed in a second stage to partially remove the noise and reverberation (Vary, 2006; Löllmann & Vary, 2007; Zheng et al., 2022). Among these methods, frequency-domain methods have been the most extensively studied, for the following reasons. First, short-term spectral analysis and synthesis can be efficiently performed by the fast Fourier transform (FFT) algorithm, although the Fourier decomposition may not be optimal for speech enhancement (Johnson et al., 2007; Benesty et al., 2009). Second, the spectral magnitude and phase of the speech are decoupled, and thus one can process the spectral magnitude alone or optimize the spectral magnitude and the phase step by step or jointly depending the hardware resources and the desired performance (Paliwal et al., 2011; Gerkmann et al., 2015; Li et al., 2022a). Third, the human ear works as a frequency analyzer (Plomp, 1964). In particular, different places on the basilar membrane within the cochlea respond to different frequency ranges in sound (Moore, 2013). Hence, it seems very natural to enhance speech in a way that mimics the operation of the normal healthy ear. Fourth, speech is sparse in the frequency domain, which facilitates removal of nonspeech components, while it is not sparse in the time domain (He et al., 2007). Fifth, frequency-domain methods can readily be used in combination with special applications, such as frequency-dependent amplification to compensate for hearing loss (Kates, 2008). Finally, frequency-domain monaural speech enhancement methods usually achieve higher objective and subjective scores than time-domain methods (Hu et al., 2020; Li et al., 2021a, 2021), although this may in part be the case because more effort has been put into frequency-domain methods than into time-domain methods. For the above reasons, this paper reviews only frequency-domain monaural speech enhancement methods, which have been extensively studied over the 60 years since Schroeder (1965, 1968) first proposed a noise suppression method using an analog implementation.

In the last ten years, monaural speech enhancement has made great progress with the development of deep learning methods (Hochreiter & Schmidhuber, 1997; Hinton & Salakhutdinov, 2006; LeCun et al., 2015). A large number of deep-learning-based frequency-domain methods have been proposed (Wang et al., 2014; Xu et al., 2014b, 2015), and they have been shown to surpass traditional frequency-domain methods in challenging conditions such as at low speech-to-noise ratios (SNRs), in high reverberation, and when the noise is nonstationary. Deep-learning methods can be categorized into two types: spectral magnitude-only enhancement and complex spectrum enhancement. The former estimate only the spectral magnitude of the clean speech. The noisy phase is used to reconstruct the time-domain speech signal. The latter estimate the real and imaginary parts of the complex spectrum of the clean speech directly, which has the potential to further improve speech quality. In general, deep-learning methods have much more computational complexity and greater storage requirements than traditional frequency-domain methods. Although many researchers have proposed ways of reducing the number of parameters and the complexity (Tan & Wang, 2021), it is still challenging to implement advanced deep-learning-based methods in resource-limited devices such as hearing aids. In contrast, traditional frequency-domain speech enhancement methods have been extensively applied in devices such as digital hearing aids, smartphones, and audio-visual communication systems. However, it is becoming more feasible to implement deep-learning methods in such devices because of the development of low-resource deep-learning methods in combination with increases in computing performance and memory capacity of signal-processing chips.

The purposes of this paper are fourfold. Firstly, the paper provides an overview of both traditional and deep-learning frequency-domain monaural speech enhancement methods that have been proposed over the last six decades. Although many review papers and books have been published on traditional methods (Loizou, 2013; Gerkmann et al., 2015) or deep-learning methods (Wang & Chen, 2018), the reviews do not give a deep insight into the advantages and disadvantages of the two categories of methods. The second purpose is to clearly reveal the fundamental assumptions underlying the different methods and the consequences of these in practical applications. The third purpose is to compare the objective performance of the two types of methods using the same speech corpus and noise corpus. To our knowledge, such a comprehensive evaluation has not been reported before. Finally, challenges for the future are formulated and some potential research topics are outlined.

The remainder of this paper is organized as follows. Firstly, the processing stages in frequency-domain monaural speech enhancement are formulated, and different estimation targets are presented. After that, traditional methods are described, including a complete flowchart of these methods and a review of each key module. Deep learning-based methods are then presented, and the reasons why deep-learning methods surpass traditional methods are discussed. The two types of methods are evaluated using a common set of objective measures. Finally, future research topics in monaural speech enhancement are discussed.

Signal Model and Problem Formulation

For monaural speech enhancement, only a single sensor, such as a microphone, is used to pick up a sound. Thus, the signal can be represented as

$$y(t) = s(t) * h(t) + v(t), \tag{1}$$

where $s(t)$ denotes the clean speech, $h(t)$ is the transfer function from the clean speech to the microphone, $v(t)$ is the noise with t the time index, and * denotes the convolution operation. When the microphone picks up the clean speech in an enclosure, the microphone signal includes the direct and early and late reflected speech components. It is usually assumed that early reflections make a positive contribution to intelligibility, while late reflections degrade speech quality and intelligibility (Bradley et al., 2003; Yoshioka et al., 2012; Hu & Kokkinakis, 2014).

With the short-time Fourier transform (STFT), the time-frequency (T-F) representation of Equation (1) can be written as

$$Y(k, l) = S(k, l)H(k, l) + V(k, l), \tag{2}$$

where $Y(k, l)$, $S(k, l)$, and $V(k, l)$ are the complex spectra of $y(t)$, $s(t)$, and $v(t)$, respectively, with k the frequency bin index and l the frame index. $H(k, l)$ is the frequency-domain representation of $h(t)$. $Y(k, l)$ can be computed as:

$$Y(k, l) = \sum_{\mu=0}^{K-1} y\big(lR + \mu\big)w\big(\mu\big)e^{-j\frac{2\pi k\mu}{K}}, \tag{3}$$

where $w(t)$ is a window, R is the frame shift and K is the frame length. When $y(t)$ is replaced by $s(t)$ and $v(t)$ in Equation (3), $S(k, l)$ and $V(k, l)$ can be obtained. For frequency-domain monaural speech enhancement, a speech dereverberation method is often designed to suppress late reflected speech components (Naylor & Gaubitch, 2010; Nakatani et al., 2010; Jukić et al., 2015), while a denoising method is designed to suppress $N(k, l)$ (Benesty et al., 2006; Loizou, 2013). When both late reflected speech components and noise are suppressed, this is denoted joint noise reduction and dereverberation (Doire et al., 2016; Li et al., 2021a; Reddy et al., 2021).

This paper focuses on reviewing and evaluating monaural speech denoising methods. It should be noted that many speech denoising methods have been extended to perform speech dereverberation (Lebart et al., 2001), based on the assumption that late reverberant speech components are uncorrelated with the direct and early-reflected speech components (see Braun et al., 2018, and references therein). This assumption is reasonable when the speech rate is reasonably rapid, so that a given speech sound at the microphone has been completed before the late reverberation from that sound arrives at the microphone. The assumption may also be reasonable if the impulse response $h(t)$ changes over time, as pointed out by Elko et al. (2003).

Without sound reflections, $h(t)$ is the discrete unit sample function $\delta(t)$, and Equation (1) becomes

$$y(t) = s(t) + v(t), \tag{4}$$

so the T-F representation of Equation (4) reduces to

$$Y(k, l) = S(k, l) + V(k, l) \tag{5}$$

because $H(k, l) \equiv 1$ in this case.

For frequency-domain monaural speech enhancement methods, the goal is to estimate the complex spectrum of the clean speech from $Y(k, l)$, so that the time-domain speech signal can be reconstructed using the inverse STFT and overlap-add method. In this paper, the estimated complex spectrum of $S(k, l)$ is denoted $\widehat{S}(k, l)$ and the estimated time-domain clean speech signal is denoted $\widehat{s}(t)$.

There are many ways to obtain $\widehat{S}(k, l)$, and these can be roughly categorized into two classes: indirect and direct. Almost all traditional methods and some deep-learning methods belong to the first class. For this indirect class, a spectral gain function $G(k, l)$ is first estimated or mapped by a network based on the noisy observations $Y(k, l)$, and this is then multiplied with $Y(k, l)$, i.e., $\widehat{S}(k, l) = G(k, l)Y(k, l)$. Commonly, these methods estimate only the clean speech spectral magnitude, such that $G(k, l)$ is a real value ranging from 0 to 1, while the noisy phase is unaltered and used for time-domain speech reconstruction.

Recent studies have confirmed that it is important to minimize the phase difference between the estimated speech and the clean speech (Paliwal et al., 2011), and phase processing for speech enhancement has attracted considerable attention in the last decade (Gerkmann, 2014; Krawczyk & Gerkmann, 2014; Gerkmann et al., 2015). Because it is difficult if not impossible to estimate the phase of the clean speech at low SNRs, traditional methods of phase processing do not greatly improve speech quality. For deep-learning speech enhancement methods, it has been shown that performance can be improved in terms of both objective and subjective metrics if the clean speech phase can be estimated before reconstructing the estimated time-domain speech signal (Williamson et al., 2016; Fu et al., 2017; Williamson & Wang, 2017; Tan & Wang, 2019, 2020; Yin et al., 2020).

In recent years, many direct methods have been proposed, where the magnitude of $\widehat{S}(k, l)$, $|\widehat{S}(k, l)|$, or the real and imaginary parts of $\widehat{S}(k, l)$, $\widehat{S}_r(k, l)$ and $\widehat{S}_i(k, l)$, are mapped directly by deep complex networks (Tan & Wang, 2020; Li et al., 2021c, 2021). In the next two sections, an overview of traditional methods and deep-learning methods is presented, and their crucial modules are reviewed and discussed.

Traditional Methods

This section reviews statistical signal-processing-based methods for monaural speech enhancement. These methods are usually based on the assumption that the speech and noise are independent, and that the speech or noise follows a specific distribution, such as Gaussian (McAulay & Malpass, 1980; Ephraim & Malah, 1984, 1985), Gamma (Martin, 2002), Laplacian (Chen & Loizou, 2007), or Super-Gaussian (Breithaupt & Martin, 2003). Hendriks et al. (2007) and McCallum & Guillemin (2013) assumed that speech was a combined stochastic-deterministic signal instead of a completely stochastic signal. Note that a great number of traditional methods do not make use of any stochastic model (Boll, 1979; Berouti et al., 1979; Gülzow et al., 2003; Li et al., 2008; Loizou, 2013). Rather, they are heuristic (rule-based) or the speech is assumed to be a deterministic signal. For example, spectral subtraction, for which the estimated spectrum of the noise is subtracted from the spectrum of the speech-plus-noise, is a typical heuristic method, and although it is not a perfect solution for monaural speech enhancement, it has been widely used in practical applications. For all of these traditional methods, five key modules are often used, namely estimation of: noise, *a priori* SNR, speech-presence probability, spectral gain, and phase. The *a priori* SNR is the true short-term power ratio between each spectral component of the clean speech and the noise. It can be contrasted with the *a posteriori* SNR, which is the short-term power ratio between each spectral component of the observed noisy speech and noise.

Figure 1 shows a flow diagram of typical traditional frequency-domain speech enhancement methods. In this figure, both a spectral magnitude enhancement module and a phase-processing module are included. Note that all five modules are not necessarily used in all traditional methods. For example, when Boll (1979) proposed spectral subtraction, which involves subtraction of the estimated noise spectrum from the speech+noise spectrum for each time-frequency bin, he used only two modules, noise estimation and spectral gain estimation. The estimate of the noise power spectral density (PSD) was updated based on segments estimated to contain only noise, via a voice activity detector (VAD). Ephraim & Malah (1984) proposed the *decision-directed* method to estimate the *a priori* SNR, and Cappé (1994) analyzed its importance in suppressing the well-known "musical noise" artifact. McAulay & Malpass (1980) first introduced a two-state model ($\mathcal{H}_0(k, l)$ and $\mathcal{H}_1(k, l)$ representing the speech-absent and speech-present states, respectively). The speech-presence probability (SPP) ($P(\mathcal{H}_1(k, l)|Y(k, l))$ was adjusted to give a good balance between noise suppression and speech distortion. Malah et al. (1999) and Cohen & Berdugo (2001) combined the SPP with the spectral gain estimation to further reduce speech distortion and increase noise reduction. Without phase processing, the spectral gain of the log-spectral amplitude (LSA) estimator derived by Cohen & Berdugo (2001) can be expressed by

$$G(k, l) = \left(G_{\mathcal{H}_1}(k, l)\right)^{p(k,l)} (G_{\min}(k, l))^{1-p(k,l)}, \qquad (6)$$

where $G_{\mathcal{H}_1}(k, l)$ is the spectral gain when speech is present and $G_{\min}(k, l)$ is often set to a constant value ranging from -30 dB to -12 dB. $G_{\min}(k, l)$ can be regarded as a predefined spectral gain when the speech is absent and its value can be different in each T-F bin depending on the noise characteristics (Gustafsson et al., 1998). $p(k, l)$ is the SPP, which is estimated from the *a posteriori* SNR, *a priori* SNR, and the *a priori* probability of speech absence (Cohen & Berdugo, 2001), given by

$$p(k, l) = \left(1 + \frac{q(k, l)}{1 - q(k, l)}(1 + \xi(k, l)) \exp(-\upsilon(k, l))\right)^{-1}, \qquad (7)$$

where $q(k, l) = P(\mathcal{H}_0(k, l))$ is the *a priori* probability of speech absence and $\xi(k, l) = E\{|S(k, l)|^2\}/\sigma_v^2(k, l)$ is the *a priori* SNR. $\upsilon(k, l)$ is defined as

$$\upsilon(k, l) = \frac{\xi(k, l)}{1 + \xi(k, l)}\gamma(k, l), \qquad (8)$$

where $\gamma(k, l) = |Y(k, l)|^2/\sigma_v^2(k, l)$ is the *a posteriori* SNR. For both the *a posteriori* SNR and *a priori* SNR, the noise PSD, $\sigma_v^2(k, l)$, needs to be estimated beforehand, and the accuracy with which this is done has a significant impact on the performance of traditional methods.

In the following five sections, the five modules are each overviewed. Then, the valid and invalid assumptions of traditional methods are described and limitations of the methods are discussed.

Noise Estimation

The noise-estimation module plays an important role for almost all traditional frequency-domain speech enhancement methods. Its performance has a direct effect on both noise reduction and speech distortion. When the noise PSD is underestimated, the amount of noise reduction is reduced, leading to speech-amplification distortion (Loizou & Kim, 2011).[1] This can explain why traditional methods do not improve the intelligibility of speech in nonstationary noises for normal-hearing listeners (Loizou & Kim, 2011), although they can improve the quality of speech in quasi-stationary noises for both hearing-impaired and normal-hearing listeners (Sang et al., 2014, 2015). In contrast, when the noise PSD is overestimated, this results in speech-attenuation distortion, even though the amount of noise reduction increases. Because of the importance of estimation of the noise PSD, many types of noise PSD estimators have been proposed.

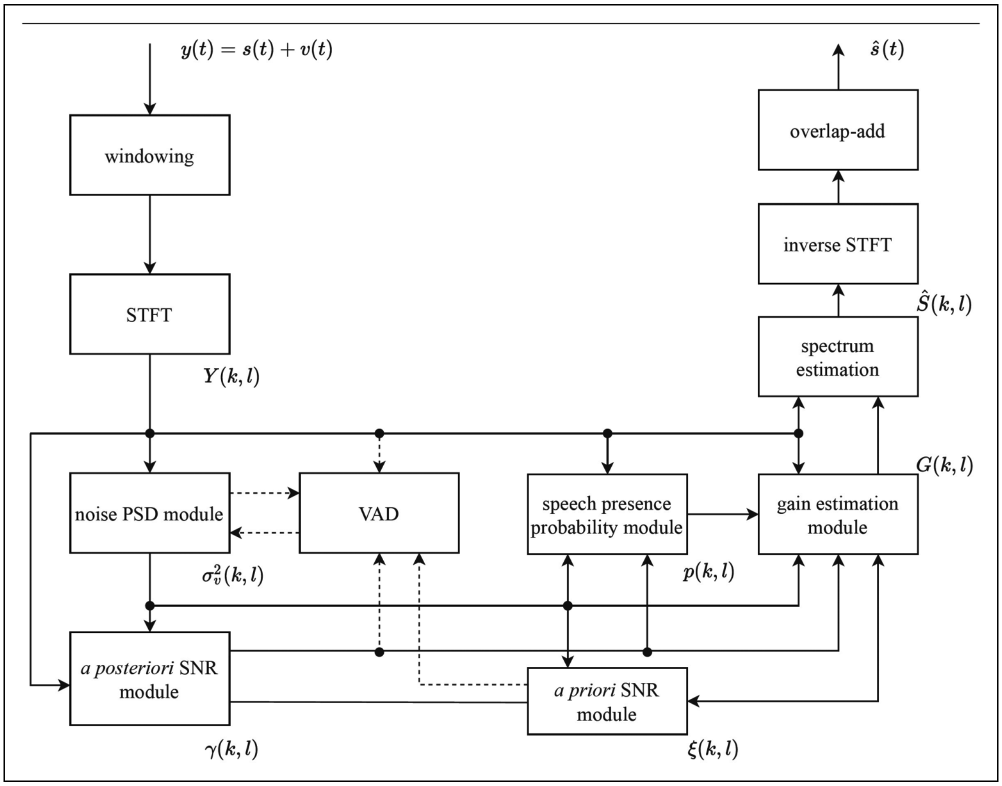

Figure 1. A flow-process diagram of traditional methods.

In early work on this topic, noise estimation was based on the use of a VAD (Lim et al., 1978; Boll, 1979; McAulay & Malpass, 1980; Ephraim & Malah, 1984, 1985). This exploits the fact that speech usually contains brief pauses, for example before or after a stop consonant. With this approach, each time frame was categorized into one of two states: speech absent and speech present. The noise PSD was updated in speech-absent frames and the estimate was maintained across subsequent speech-present frames. This is represented by

$$\sigma_v^2(k,l) = \begin{cases} \alpha_v \sigma_v^2(k,l-1) + (1-\alpha_v)|Y(k,l)|^2, & SA \\ \sigma_v^2(k,l-1), & SP \end{cases} \tag{9}$$

where $\alpha_v \in (0\ 1)$ is a smoothing factor, SA indicates speech absent, and SP indicates speech present.

There are two drawbacks of this VAD-based noise PSD estimator. Firstly, noise-estimation accuracy depends strongly on the performance of the VAD. Misclassification of speech-present frames as speech-absent frames leads to overestimation of the noise, resulting in speech attenuation distortion. Unfortunately, this misclassification problem cannot be avoided when the SNR is low (Sohn et al., 1999; Tan et al., 2020), especially when the noise is nonstationary. Secondly, since the noise PSD is not updated during frames classified as speech present, the sparsity of speech in the T-F domain is not fully exploited. This sparsity is readily apparent in spectrograms of clean speech from a single talker (Darwin, 2009); there are many T-F regions with very low energy.

If the noise PSD is estimated via use of a VAD, this implicitly assumes that the noise is quasi-stationary and that its PSD changes slowly over time. To avoid use of a VAD, one more assumption is necessary, namely that the PSD of each speech-plus-noise T-F bin is always larger than or equal to the noise PSD for the corresponding T-F bin, i.e., $E\{|Y(k,l)|^2\} \geq E\{|V(k,l)|^2\}$. This assumption is always true because the noisy PSD equals the sum of the speech and noise PSDs, i.e., $E\{|Y(k,l)|^2\} = E\{|S(k,l)|^2\} + E\{|V(k,l)|^2\}$. Accordingly, a recursive method of estimating the noise spectral magnitude without a VAD was proposed by Hirsch & Ehrlicher (1995). This method is expressed by

$$\sigma_v(k,l) = \begin{cases} \alpha_v \sigma_v(k,l-1) + (1-\alpha_v)|Y(k,l)|, \\ \mathit{if}\ |Y(k,l)| \leq \beta_v \sigma_v(k,l-1) \\ \sigma_v(k,l-1), \qquad \text{otherwise} \end{cases} \tag{10}$$

where β_v ranges from 1.5 to 2.5 in practical applications. Doblinger (1995) proposed a similar method for estimating the noise PSD recursively without a VAD. A histogram-based recursive noise estimation method was also presented by Hirsch & Ehrlicher (1995). With this method, each bin was roughly categorised as noise only if $|Y(k,l)| \leq \beta_v \sigma_v(k,l-1)$. Histograms of about 40 past noise bins in each subband (a subband corresponds to a limited frequency region) were created and the estimated noise level for a given subband was set to the value at the peak of the distribution for that noise band. Martin (1994) proposed a well-known noise estimation method, called minimum statistics, which can be expressed as

$$\sigma_v^2(k,\, l) = O_{\min} \min\left\{P_y(k,\, \ell)\big|_{\ell=l-D+1,\ldots,l}\right\}, \tag{11}$$

where $O_{\min}$ was introduced to reduce the estimation bias and D determines the number of past frames used to estimate the noise PSD. For a discussion of estimation bias and its compensation, see Martin (2006). $P_y(k,\, l)$ is a recursive average of $|Y(k,\, l)|^2$, given by $P_y(k,\, l) = \alpha_y P_y(k,\, l-1) + (1-\alpha_y)|Y(k,\, l)|^2$, with α_y the smoothing factor. Martin (2001) improved the noise estimation performance of the minimum statistics method using optimally recursive smoothing of the noisy PSD and bias compensation. More precisely, α_y was replaced by $\alpha_y(k,\, l)$, according to the extent to which the PSD fluctuated over time for each bin, and $O_{\min}$ became a T-F varying quantity, whose value depended on D and the smoothing factor $\alpha_y(k,\, l)$.

Cohen (2003) proposed an improved minima-controlled recursive averaging method for noise estimation. In this method, the noise PSD is first roughly estimated using the minimum statistics method, and the SPP is then estimated to control the smoothing factor used in estimating the noise PSD. This can be written as

$$\begin{aligned}\bar{\sigma}_v^2(k,\, l) &= \alpha_v(k,\, l)\bar{\sigma}_v^2(k,\, l-1) \\ &\quad + (1-\alpha_v(k,\, l))|Y(k,\, l)|^2,\end{aligned} \tag{12}$$

where $\alpha_v(k,\, l) = \alpha_v + (1-\alpha_v)p(k,\, l)$, with α_v a constant smoothing factor, and $\sigma_v^2(k,\, l) = O_v\bar{\sigma}_v^2(k,\, l)$, with O_v a bias compensation factor.

Hendriks et al. (2010) and Gerkmann & Hendriks (2012) suggested estimating the noise PSD based on the minimum mean-square error (MMSE) criterion[2] of the noise magnitude-squared STFT coefficients. The estimated noise PSD can then be calculated from the conditional expectation of $|N(k,\, l)|^2$ given $Y(k,\, l)$, i.e., $E\{|N(k,\, l)|^2|Y(k,\, l)\}$, which can be evaluated using Bayes' rule. The MMSE-based method achieved the best performance in terms of the mean and variance of the logarithmic difference between the true and estimated noise PSD among the traditional noise PSD estimators available at that time (Taghia et al., 2011). Because of its low complexity and low latency, this MMSE-based method has become one of the most popular methods for traditional monaural speech enhancement.

Although many researchers have attempted to improve noise-estimation accuracy when the noise is nonstationary (Cohen, 2003; Rangachari & Loizou, 2006; Hendriks et al., 2010; Gerkmann & Hendriks, 2012; Zhang et al., 2019), this remains a challenging task for highly nonstationary noises, such as babble noise, siren noise, and wind noise. As mentioned above, the performance degradation of traditional noise PSD estimators when the noise is nonstationary limits the performance of traditional methods. As a result, data-driven methods have been proposed to improve noise-tracking performance (Erkelens & Heusdens, 2008; Li et al., 2019; Liu et al., 2021).

A Priori *SNR estimation*

Before Cappé (1994) analyzed the operation of the *decision-directed* method in reducing musical-noise artifacts without sacrificing the quality of speech, the importance of the *a priori* SNR in speech enhancement had not been appreciated, although the maximum-likelihood method for estimation of the *a priori* SNR had been proposed a long time previously (Boll, 1979; McAulay & Malpass, 1980; Ephraim & Malah, 1984). Maximum likelihood and *decision-directed* estimation of the *a priori* SNR can be, respectively, given by

$$\xi_{\mathrm{ML}}(k,\, l) = \max\{\gamma(k,\, l) - \beta_\gamma,\, 0\}, \tag{13}$$

and

$$\begin{aligned}\xi_{\mathrm{DD}}(k,\, l) &= \alpha_\xi \frac{\left|\widehat{S}(k,\, l-1)\right|^2}{\sigma_v^2(k,\, l)} \\ &\quad + (1-\alpha_\xi)\max\{\gamma(k,\, l) - \beta_\gamma,\, 0\},\end{aligned} \tag{14}$$

where β_γ is a constant value, usually set to 1, $\widehat{S}(k,\, l-1)$ is the estimated complex speech spectrum for the previous frame, and the smoothing factor α_ξ is very close to one. $\xi_{\mathrm{DD}}(k,\, l)$ reduces to $\xi_{\mathrm{ML}}(k,\, l)$ when $\alpha_\xi = 0$, so $\xi_{\mathrm{DD}}(k,\, l)$ can be interpreted as a recursive average of $\xi_{\mathrm{ML}}(k,\, l)$, which can be written as

$$\begin{aligned}\xi_{\mathrm{DD}}(k,\, l) &= \alpha_\xi G_{\mathcal{H}_1}^2(k,\, l-1)\gamma(k,\, l-1) \\ &\quad + (1-\alpha_\xi)\max\{\gamma(k,\, l) - \beta_\gamma,\, 0\},\end{aligned} \tag{15}$$

where $G_{\mathcal{H}_1}(k,\, l)$ is the spectral gain when speech is present. In the work of Ephraim & Malah (1984, 1985) and Cohen & Berdugo (2001), $G_{\mathcal{H}_1}(k,\, l)$ was a function of both the *a posteriori* SNR and the *a priori* SNR. As shown by Cappé (1994), the *a priori* SNR is the key parameter in determining the spectral gain, and thus it is sufficient to use $G_{\mathcal{H}_1}(k,\, l) = \xi(k,\, l)/(1+\xi(k,\, l))$ in practical applications.

There are two drawbacks of the *decision-directed* method, as pointed out by Cohen (2005) and Plapous et al. (2006). Firstly, a delay of one frame is needed to track the *a priori* SNR at speech onsets, leading to speech distortion. Secondly, a delay of one frame is needed to track the *a posteriori* SNR at speech offsets (Cohen, 2005). By taking into account the correlation between adjacent speech frames, Cohen (2005) proposed causal and noncausal methods for *a priori* SNR estimation. The causal method can solve the first problem of the *decision-directed* method, reducing speech distortion, while the noncausal method can solve the two above-mentioned problems simultaneously at the expense of three additional frames delay. Plapous et al. (2004, 2006) proposed a two-stage SNR estimator, solving the two drawbacks of the *decision-directed* method simultaneously without introducing further algorithmic delay. The two-stage SNR estimator is given by

$$\xi_{\mathrm{TSNR}}(k, l) = G^2_{\mathcal{H}_1}(k, l)\gamma(k, l), \tag{16}$$

where $G_{\mathcal{H}_1}(k, l) = \xi_{\mathrm{DD}}(k, l)/(1 + \xi_{\mathrm{DD}}(k, l))$ was used by Plapous et al. (2006). In the first stage of the two-stage SNR estimator, the *a priori* SNR is roughly estimated with the *decision-directed* method, and this estimated *a priori* SNR is then used to calculate the Wiener filter gain to weight the *a posteriori* SNR in the second stage. This Wiener filter gain is derived by minimizing the mean-square error (MSE) between the estimated clean speech and the true speech. Note that $\xi_{\mathrm{TSNR}}(k, l)$ is determined by both $\xi_{\mathrm{DD}}(k, l)$ and $\gamma(k, l)$. For noise-only segments, $\xi_{\mathrm{DD}}(k, l)$ is reliable and close to zero, while $\xi_{\mathrm{TSNR}}(k, l)$ is also close to zero because $\gamma(k, l)$ is small. For speech-onset frames, although $\xi_{\mathrm{DD}}(k, l)$ is small because of the one-frame delay of the *decision-directed* method, $\gamma(k, l)$ is often not small, and thus the use of $\xi_{\mathrm{TSNR}}(k, l)$ solves the one-frame delay problem effectively. For speech-offset frames, $\xi_{\mathrm{DD}}(k, l)$ is not small because of the one-frame delay of the *decision-directed* method, while $\gamma(k, l)$ approaches zero, and thus $\xi_{\mathrm{TSNR}}(k, l)$ becomes much smaller than $\xi_{\mathrm{DD}}(k, l)$, as expected.

Except for the maximum-likelihood method, the above-mentioned methods for estimating the *a priori* SNR only exploit the correlation over time for each frequency bin, while the correlation over frequency is ignored. Breithaupt et al. (2008) exploited the latter to develop a novel *a priori* SNR estimator by selectively smoothing the maximum-likehood estimate of the speech power in the cepstral domain. This SNR estimator surpassed the *decision-directed* method in both stationary and nonstationary noise scenarios. With this new method, the residual noise sounded more natural than for the *decision-directed* method and the musical-noise problem was reduced.

All of the above-mentioned *a priori* SNR estimators need to make an initial estimate of the *a posteriori* SNR. If the *a posteriori* SNR is overestimated due to underestimation of the noise PSD, this often leads to an overestimate of the *a priori* SNR. There are two ways of solving this problem. One is to improve the accuracy of the noise PSD estimate using data-driven methods (Erkelens & Heusdens, 2008; Li et al., 2019; Liu et al., 2021). The other is to estimate the *a priori* SNR directly using data-driven methods without estimating the noise PSD (Nicolson & Paliwal, 2019, 2020; Zhang et al., 2020). Strictly speaking, these data-driven methods should not be classified as traditional methods, but since they can be used for parameter estimation, they are still mentioned in this section.

Speech Presence Probability Estimation

As shown in Equation (7), the SPP depends on three parameters: the *a posteriori* SNR, $\gamma(k, l)$, the *a priori* SNR, $\xi(k, l)$, and the *a priori* probability of speech absence $q(k, l)$. All three parameters need to be estimated for each T-F bin. Methods for estimating $\xi(k, l)$ are reviewed above. This section presents a brief survey of methods for estimating $q(k, l)$ and describes some improved SPP estimators that do not require estimation of the *a priori* SNR.

The *a priori* probability of speech absence is defined as $q(k, l) = P(\mathcal{H}_0(k, l))$. In general, $q(k, l)$ is different in each T-F bin. However, for initialization, $q(k, l)$ is often set to a constant value. For example, McAulay & Malpass (1980) set $q(k, l)$ to 0.5, which is equivalent to assuming that speech presence and absence are equally probable, i.e., $P(\mathcal{H}_0(k, l)) = P(\mathcal{H}_1(k, l)) = 0.5$. Ephraim & Malah (1984) analyzed the impact of $q(k, l)$ on the spectral gain, and showed that increasing $q(k, l)$ led to a decrease of the spectral gain as $\gamma(k, l)$ decreased when $\xi(k, l)$ was large. Malah et al. (1999) estimated $q(k, l)$ as

$$q(k, l) = \alpha_q q(k, l-1) + (1 - \alpha_q)I(k, l), \tag{17}$$

where α_q is a smoothing factor, and $I(k, l)$ is set to 1 if $\mathcal{H}_0(k, l)$ holds true while it is set to 0 if $\mathcal{H}_1(k, l)$ holds true. In the approach of Malah et al. (1999), $I(k, l)$ only depended on $\gamma(k, l)$, and $I(k, l)$ was set to 1 when $\gamma(k, l) \leq \gamma_{\mathrm{TH}}$ while it was set to 0 otherwise. Cohen (2003) did not smooth $q(k, l)$ over time. Its value depended only on the local smoothed *a posteriori* SNR values, with $q(k, l)$ ranging from 0 to 1. As in Malah et al. (1999), $q(k, l)$ was set to be close to 1 when the *a posteriori* SNR and its smoothed version were small, and $q(k, l)$ reduced as the *a posteriori* SNR and its smoothed version increased.

After estimating $\gamma(k, l)$, $\xi(k, l)$, and $q(k, l)$, the SPP can be calculated using Equation (7) (Malah et al., 1999; Cohen, 2003). To reduce the estimation variance, Malah et al. (1999) and Cohen (2003) introduced smoothing over time for the estimation of $q(k, l)$ and $\gamma(k, l)$. Gerkmann et al. (2008) proposed an improved SPP estimator using a fixed *a priori* probability of speech absence and a fixed *a priori* SNR. In this way, the SPP depended only on the *a posteriori* SNR. Instead of using only one T-F bin, smoothed *a posteriori* SNR values were used to determine the local and global SPP, the local SPP using many fewer frames and frequency bins than the global SPP. The final SPP value was a multiplicative combination of the local and global SPP. As pointed out by Gerkmann et al. (2008), by decoupling the SPP estimation from the *a priori* SNR estimation and the estimation of the *a priori* probability of speech absence, SPP estimation performance was improved. This happened because both $q(k, l)$ and $\xi(k, l)$ depend on the accuracy of the estimate of $\gamma(k, l)$. If $\gamma(k, l)$ is overestimated, leading to overestimates of both $q(k, l)$ and $\xi(k, l)$, the problems produced by overestimating the SPP will be increased.

Spectral Gain Estimation

Spectral gain estimation methods can be categorized into three groups: deterministic, stochastic, and stochastic-deterministic. In the following, each of these three groups is reviewed.

Deterministic Methods. The first group derives the spectral gain under the assumption that the speech and noise are independent of each other and that $E\{|Y(k,l)|^2\} = E\{|S(k,l)|^2\} + E\{|V(k,l)|^2\}$ always holds true. Because speech is highly nonstationary, leading to nonstationary characteristics of the corresponding noisy speech, only a very limited number of frames should be used to estimate $E\{|S(k,l)|^2\}$ and $E\{|Y(k,l)|^2\}$. On the other hand, the noise is often assumed to be stationary or quasi-stationary, so its PSD does not rapidly change over time, and more frames can be used to estimate $E\{|V(k,l)|^2\}$. This leads to the following approximation equation

$$|Y(k,l)|^{\alpha_g} \approx |S(k,l)|^{\alpha_g} + \sigma_v^{\alpha_g}(k,l), \tag{18}$$

where α_g is an exponent to which the magnitude of each STFT bin is raised (Sim et al., 1998). Using Equation (18), the speech spectral magnitude can be estimated by

$$|S(k,l)| = (\max\{|Y(k,l)|^{\alpha_g} - \beta_g \sigma_v^{\alpha_g}(k,l), 0\})^{1/\alpha_g}, \tag{19}$$

where β_g is the subtraction factor. When $\alpha_g = 1$, Equation (19) expresses spectral subtraction of magnitudes (Boll, 1979). When $\alpha_g = 2$, equation (19) expresses spectral subtraction of powers (Lim & Oppenheim, 1979). It is interesting that square-root spectral subtraction (setting α_g to 0.5) can give better performance in reducing the log kurtosis ratio[3] and cepstral distance while keeping the same amount of noise reduction as the magnitude and power spectral-subtraction methods (Inoue et al., 2010a, 2010). Evaluations with human listeners have confirmed the benefit of using $\alpha_g = 0.5$. Equation (19) can be re-written as

$$G(k,l) = \frac{\max\{\gamma^{\alpha_g/2}(k,l) - \beta_g, 0\}}{\gamma^{\alpha_g/2}(k,l)}, \tag{20}$$

which holds true because the noisy phase is used directly. Assuming that $\xi^{\alpha_g/2} = \max\{\gamma^{\alpha_g/2} - \beta_g, 0\}$, $G(k,l)$ in Equation (20) reduces to

$$G(k,l) = \frac{\xi^{\alpha_g/2}(k,l)}{\xi^{\alpha_g/2}(k,l) + \beta_g}, \tag{21}$$

and when $\alpha_g = 2$ and $\beta_g = 1$, Equation (21) reduces to the standard Wiener filtering method. Comparing Equation (20) with Equation (21), the former depends on the *a posteriori* SNR, while the latter depends on the *a priori* SNR. As mentioned above, except for the maximum-likelihood estimator for *a priori* SNR estimation, several existing estimators such as the *decision-directed* method can effectively reduce large fluctuations of the *a posteriori* SNR in noise-only segments and track the *a posteriori* SNR in noisy segments with high SNR values, leading to alleviation of the musical noise problem and reduced speech distortion.

When the spectral gain is determined only by the *a posteriori* SNR, as in Equation (20), the musical noise problem arises. There are many ways to reduce this problem. One is to increase the value of β_g (Boll, 1979) and reduce the value of α_g (Inoue et al., 2010b, 2010). A second method is to introduce a noise floor to mask the annoying musical noise (Lim & Oppenheim, 1979; Boll, 1979). A third method is to smooth the noise PSD over frequency (Hu & Loizou, 2004) and to smooth the spectral gain adaptively over time (Gustafsson et al., 2001). A fourth method is to exploit the masking properties of the human auditory system (Gustafsson et al., 1998; Virag, 1999). Virag (1999) calculated the masked threshold of the noise in the presence of the speech, and used this to determine the subtraction factor β_g and the residual noise floor $G_{\min}$ in each T-F bin. Gustafsson et al. (1998) proposed a novel spectral gain estimation method, where the spectral gain was designed to make the residual noise inaudible. With this method, the spectral gain can be written as

$$G(k,l) = \min\left\{\sqrt{\frac{R_T(k,l)}{\sigma_v^2(k,l)}} + G_{\min}, 1\right\}, \tag{22}$$

where $R_T(k,l)$ is the masked threshold of the noise. The method for calculating $R_T(k,l)$ can be found in Fastl & Zwicker (2007).

Equation (18) is only an approximation and the two cross terms, $S(k,l)V^*(k,l)$ and $S^*(k,l)V(k,l)$, cannot be ignored without the mathematical expectation operator. Lu & Loizou (2008) considered this approximation error and proposed a geometric approach for speech enhancement. In this method, the geometric relationship among the phases of the noisy, clean, and noise spectra was used to derive the suppression function. As shown by Lu & Loizou (2008), this geometric approach has similar properties to the MMSE method proposed by Ephraim & Malah (1984).

Stochastic Methods. McAulay & Malpass (1980) assumed that the speech and noise are two statistically independent Gaussian random processes and used a maximum-likelihood method to estimate the speech spectrum. Based on modeling the real and imaginary parts of the speech and noise complex spectra as statistically independent Gaussian random variables, Ephraim & Malah (1984) derived the MMSE short-time spectral amplitude (MMSE-STSA) estimator. Its spectral gain under the condition of speech presence is given by

$$G_{\mathcal{H}_1}(k,l) = \Gamma(1.5)\frac{\sqrt{\upsilon(k,l)}}{\gamma(k,l)} M(-0.5; 1, -\upsilon(k,l)), \tag{23}$$

where $\Gamma(\bullet)$ is the Gamma function and $M(\bullet;\bullet;\bullet)$ is the confluent hypergeometric function (Ephraim & Malah, 1984). Using the same stochastic assumptions, Ephraim & Malah (1985) derived the MMSE-LSA estimator whose spectral gain under the condition of speech presence is given by

$$G_{\mathcal{H}_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\upsilon(k,l)}^{+\infty} \frac{e^{-t}}{t} dt\right). \tag{24}$$

As demonstrated by Ephraim & Malah (1985), the MMSE-LSA estimator suppressed noise better than the MMSE-STSA estimator without introducing more noticeable speech distortion. There are two obvious differences between the MMSE-LSA estimator and the MMSE-STSA estimator: one is that the former attenuates more noise, and the other is that the former minimizes the MSE of the log power spectra while the latter minimizes the MSE of the magnitude spectra, which can, respectively, be given by

$$E\left\{\left(\log\left(\left|\widehat{S}(k,l)\right|\right) - \log(|S(k,l)|)\right)^2\right\}, \tag{25}$$

and

$$E\left\{\left(\left|\widehat{S}(k,l)\right| - |S(k,l)|\right)^2\right\}. \tag{26}$$

Taking the logarithm of the speech power spectrum reduces its dynamic range dramatically, which emphasizes estimation errors of the low-energy parts of the speech spectrum. Because $G_{\mathcal{H}_1}(k,l)$ is derived under the condition of speech presence, if residual noise exists when $S(k,l) = 0$, the estimation error of the log power spectrum is much larger than that of the power spectrum, and thus the residual noise tends to be more suppressed with the MMSE-LSA estimator. You et al. (2005) introduced the β-order MMSE-STSA estimator for which the MSE to be minimized was defined as

$$E\left\{\left(\left|\hat{S}(k,l)\right|^{\beta_s} - |S(k,l)|^{\beta_s}\right)^2\right\}, \tag{27}$$

where β_s is the compression factor. The spectral gain of the β-order MMSE-STSA estimator, as given by You et al. (2005), is

$$\begin{aligned} G_{\mathcal{H}_1}(k,l) &= \tfrac{\sqrt{\upsilon(k,l)}}{\gamma(k,l)} \times \\ &\left(\Gamma\left(\tfrac{\beta_s}{2}+1\right) M\left(-\tfrac{\beta_s}{2};1;-\upsilon(k,l)\right)\right)^{1/\beta_s}. \end{aligned} \tag{28}$$

As pointed out by You et al. (2005), Equation (28) reduces to the MMSE-STSA estimator when $\beta_s = 1$ and to the MMSE-LSA estimator when $\beta_s \to 0$. Breithaupt et al. (2008) extended the β-order MMSE-STSA estimator to a more general MMSE-STSA estimator, in which the Chi distribution with an arbitrary number of degrees of freedom was introduced to model the power spectrum. Note that the number of degrees of freedom is 2 for the MMSE-STSA estimator when using the Gaussian statistical model (Ephraim & Malah, 1984).

Loizou (2005) proposed a group of perceptually motivated Bayesian estimators of the STSA based on some perceptually related quantity that is to be minimized, such as the Itakura-Saito divergence (Itakura & Satio, 1968), weighted likelihood-ratio distortion, and weighted square estimation error. They showed that the best performance in terms of a specific objective measure was achieved if that same objective measure was chosen as the quantity to be minimized.

Although the Gaussian statistical model has been widely used for modeling speech and noise, it is well accepted that speech is non-Gaussian (Martin, 2002; Breithaupt & Martin, 2003; Cohen, 2005; Chen & Loizou, 2007) and that different types of noise follow different distributions (Davis et al., 2006). It is sometimes reasonable to assume that the noise has a Gaussian distribution, but a non-Gaussian statistical model may be more appropriate for speech. Using different statistical models of the speech, different MMSE-STSA estimators have been derived and have been shown to improve speech enhancement performance when compared with use of a Gaussian statistical model. In addition to modeling the speech and noise with the same statistical model, Martin (2002) derived two novel MMSE-STSA estimators using different distributions for the speech and noise. Both estimators were based on a Gamma distribution for the speech, but one was based on a Gaussian noise model and the other on a Laplacian noise model. Modeling both speech and noise with non-Gaussian statistical models led to increases in the amount of noise reduction, but only a marginal objective performance improvement was found.

Deterministic-stochastic Methods. In one well known speech-production model (Quatieri, 2006), speech can be linearly predicted as follows

$$s(t) = \sum_{t_0=1}^{t_P} a(t_0)s(t-t_0) + e(t), \tag{29}$$

where $a(t_0)$, with $t_0 = 1, \ldots, t_P$, are the linear prediction coefficients and $e(t)$ is the excitation signal. For voiced segments, $e(t)$ is a periodic pulse train or saw-tooth wave and is not stochastic (the waveform produced by vibration of the vocal folds is approximately saw-tooth shaped), while for unvoiced segments, $e(t)$ can be modeled as a Gaussian stochastic signal. Since $s(t)$ is not a fully stochastic signal, it may be better to use such a deterministic-stochastic model for the speech. Stylianou (2001) used the harmonic-plus-noise model in speech synthesis, and this was later extended to speech enhancement by Zavarehei et al. (2007). The harmonic-plus-noise model can be given by

$$\begin{aligned} s(t) &= s_h(t) + s_{nh}(t) \\ &= \sum_{k_h=-K_h(t)}^{K_h(t)} A_{k_h}(t) e^{jk_h\omega_0(t)t} + s_{nh}(t), \end{aligned} \tag{30}$$

where $s_h(t)$ and $s_{nh}(t)$ denote the harmonic and non-harmonic (noise) components, respectively. $\omega_0(t)$ denotes the fundamental frequency at time index t and $K_h(t)$ is the number of harmonics. The harmonic term $s_h(t)$ is deterministic and the noise term $s_{nh}(t)$ is stochastic. This deterministic-stochastic speech model gives a better characterization of the distribution of amplitudes in speech than deterministic-only and stochastic-only models.

Hendriks et al. (2007) proposed an MMSE-STSA estimator based on a stochastic-deterministic speech model. The estimated speech spectrum was a linear combination of the noisy spectrum filtered by a Wiener filter and the expectation of the noisy spectrum, where the linear combination factor was determined by the uncertainty of the speech presence. McCallum & Guillemin (2013) also proposed a stochastic-deterministic MMSE-STSA estimator that explicitly exploited the periodic structure of speech.

Evaluations using objective measures of speech quality/intelligibility have demonstrated the superiority of MMSE-STSA estimators based on stochastic-deterministic speech models over those based on stochastic models (Hendriks et al., 2007; McCallum & Guillemin, 2013). However, this superiority is marginal. Informal listening tests also show only a marginal benefit of stochastic-deterministic speech models (McCallum & Guillemin, 2013).

Some Remarks. All of the above-mentioned spectral gain estimation methods need to estimate the noise PSD. Some do this using only the *a posteriori SNR* or *a priori* SNR, while others use both types of SNR. Because the gain functions discussed above are all real and non-negative, they extract clean speech nonlinearly, especially when oversubtraction ($\beta_g > 1.0$ in Equation (19)) and other nonlinear operations are applied to spectral gain functions to alleviate "musical noise" (Hussain et al., 2007; Udrea et al., 2008; Parchami et al., 2016). For linear filtering of noisy speech, the complex-valued Wiener filter gain, defined as the ratio of the cross-PSD between the clean speech and noisy speech and the noisy PSD, should be used (Parchami et al., 2016). Under the assumption that the noise and speech are uncorrelated, the complex-valued Wiener filter gain reduces to the real-valued Wiener filter gain shown in Equation (21) with $\alpha_g = 2$ and $\beta_g = 1$. Estimation methods using stochastic and stochastic-deterministic speech models often show better performance than methods using deterministic models when evaluated using objective metrics, such as the segmental SNR, noise reduction, and PESQ (Rix et al., 2001). However, these objective measures are not always consistent with the results of listening tests. For example, as shown by Gustafsson et al. (1998), the MMSE-STSA estimator achieved better performance in terms of speech preservation and noise attenuation than a psychoacoustically motivated speech-enhancement method, but the former led to audible residual noise that led to markedly poorer scores in listening tests. For all these model-based methods, performance is poorer when the speech and noise are not consistent with the assumed model. This problem is difficult to solve because there are many types of noise with differing amplitude distributions, so no single model is appropriate for all types of noise. It is highly desirable for the spectral gain to be automatically chosen according to the noise characteristics, which would be expected to lead to good performance for a large variety of noise types. However, this is an extremely difficult task, if not impossible, without supervised methods. This issue is discussed later in this paper.

Phase Processing

Wang & Lim (1982) were among the first to consider whether or not phase estimation is necessary. They concluded that the phase was unimportant for monaural enhancement of speech in white Gaussian noise. Based on this finding, phase estimation was ignored for a long time. In any case, it is extremely difficult to estimate the phase of the clean speech directly from the noisy speech, especially when the SNR is low.

Instead of modifying the STSA of the noisy spectrum, Wojcicki et al. (2008) described a method of changing the noisy phase spectrum. This changed phase spectrum was combined with the noisy amplitude spectrum to reconstruct the complex spectrum of the enhanced speech. This also suppressed noise and preserved the speech, and Wojcicki et al. (2008) showed that changing the noisy phase spectrum outperformed the spectral subtraction method proposed by Boll (1979) and the MMSE-STSA estimator proposed by Ephraim & Malah (1984) in mean PESQ scores. Paliwal et al. (2011) conducted four experiments to assess the importance of phase estimation. They concluded that using the clean speech phase spectrum can greatly improve speech quality when using mismatched analysis windows for the spectral amplitude and phase estimation. In such approaches, windows with low dynamic range, such as Dolph-Chebyshev windows, are used for phase estimation and Hanning/Hamming windows are commonly used for spectral amplitude estimation. When the MMSE-STSA estimator was implemented with the phase spectrum compensation (PSC) method incorporated, better performance was found than when modifying only the amplitude spectrum or only the phase spectrum (Paliwal et al., 2011).

Gerkmann et al. (2015) give a comprehensive overview of phase-processing-based monaural speech enhancement methods. Interested readers are referred to that paper for details.

Only a few researchers have proposed enhancing speech in the complex-spectrum domain, based on the assumption that the real and imaginary parts of the complex spectrum are statistically independent (Martin, 2005; Erkelens et al., 2007; Zhang & Zhao, 2013; Schwerin & Paliwal, 2014). As stated by Parchami et al. (2016), this assumption is an alternative to the assumption that the magnitude and phase of the complex spectrum are independent; the assumptions are not the same (Martin, 2005). Without some sort of independence assumption, a closed-form estimator for speech enhancement cannot be derived. For deep-learning methods, such assumptions are not necessary, and the magnitude and phase or the real and imaginary parts of complex spectrum are often jointly optimized. We return to this issue later.

Discussion

For most traditional frequency-domain speech enhancement methods, there are four underlying assumptions. The first is that the speech and noise are statistically independent. The second is that the noise is much more stationary than the speech. The third is that each T-F bin is statistically independent of other bins when deriving the spectral gain function under a specific statistical model. The last is that the speech phase is not as important as the speech spectral amplitude. While the first assumption is reasonable, the other three are not, and this constrains the application scenarios and limits the performance of methods based on these assumptions.

The second assumption is fundamental to most existing noise PSD estimators, such as the VAD (Boll, 1979), minimum statistics (Martin, 2001), and MMSE (Gerkmann & Hendriks, 2012) methods. Without this assumption, the noise PSD cannot be estimated using noise-only segments or represented by the minimum value of the noisy PSDs of several past frames. This assumption also limits the application scenarios, and most noise PSD estimators only work well for quasi-stationary noises. The noise PSD is often underestimated for highly nonstationary noises, such as babble noise (Loizou & Kim, 2011), siren noise (Sherratt et al., 1999), and transient noises (Talmon et al., 2011). To improve the performance of speech enhancement in highly nonstationary noise scenarios, each highly nonstationary noise needs a specially designed method. To the best of our knowledge, a common framework for handling all types of highly nonstationary noise does not exist for traditional methods. It is highly desirable to be able to reduce both stationary and nonstationary noises using a unified framework.

The third assumption ignores the fact that the speech spectrogram has clear structure over time and frequency, as shown in Figure 2(b_1). For unvoiced speech segments, the energy is concentrated at high frequencies, while there are obvious harmonic components for voiced speech segments. Although the T-F independence assumption is critical in deriving the spectral gain using a specific statistical model, the time-correlation between adjacent speech spectral components has actually been exploited for speech enhancement over the last four decades. As an example, the *decision-directed* method proposed by Ephraim & Malah (1984) implicitly uses the time-correlation between successive speech spectral components, such that the estimated *a priori* SNR for the current frame is influenced by the estimated clean speech amplitude for the previous frame. Cohen (2005) explicitly exploited the correlation between successive frames and derived causal and noncausal *a priori* SNR estimators, which gave improved performance as assessed using objective measures when compared with the *decision-directed* method. Breithaupt et al. (2007, 2008) implicitly considered both the time-correlation and frequency-correlation of speech in the cepstral domain. Benesty & Huang (2011) first proposed a multiframe approach for single-channel speech enhancement, and this approach was further studied by Huang & Benesty (2012) and Schasse & Martin (2014). As shown by Huang & Benesty (2012), both narrowband and wideband actual SNRs can be improved when the interframe correlation of speech is taken into account. Making use of the interframe correlation led to residual noise being less artificial and more harmonic components being preserved (Huang & Benesty, 2012; Schasse & Martin, 2014). Some methods have explicitly exploited the harmonic structure of voiced speech to improve suppression of residual noise and/or to regenerate harmonic speech components (Lim et al., 1978; Plapous et al., 2006; Zavarehei et al., 2007; Jin et al., 2010; Hou & Zhu, 2011), leading to improvements in the amount of noise reduction in nonstationary noise scenarios. Speech has both short-term and long-term structure but it is difficult to exploit all forms of structure with traditional methods. It can be expected that monaural speech enhancement would be markedly improved if all of the forms of structure in speech were exploited.

The last assumption has had a significant impact on the progress of research on frequency-domain monaural speech enhancement. Most traditional methods do not take into account the importance of phase estimation in improving speech quality. As shown in Figure 2(c_1), the speech phase spectrogram is stochastic and has no clear time or frequency correlation, resulting in difficulty estimating the phase of clean speech[4] .

The unstructured nature of speech phase using typical analysis methods makes it difficult to estimate the phase of the clean speech directly from noisy observations. To tackle this problem, two-stage traditional methods have been proposed. In the first stage, the spectral amplitude of the clean speech is estimated. Iterative approaches for phase estimation are then applied to reconstruct the time-domain signal (Griffin & Lim, 1984). The spectral amplitude has been estimated using the MMSE-STSA estimator, and the phase has been separately estimated using a PSC algorithm (Wojcicki et al., 2008; Paliwal et al., 2011). Wojcicki et al. (2008) and Paliwal et al. (2011) showed that use of the PSC algorithm alone can suppress noise, and that higher PESQ scores were achieved when the PSC algorithm was combined with other speech enhancement methods such as MMST-STSA. Mowlaee & Saeidi (2013) proposed a method for joint optimization of spectral amplitude and phase in an iterative manner. This method can be regarded as an iterative version of two-stage traditional methods. Pruša et al. (2017) proposed a noniterative method, namely

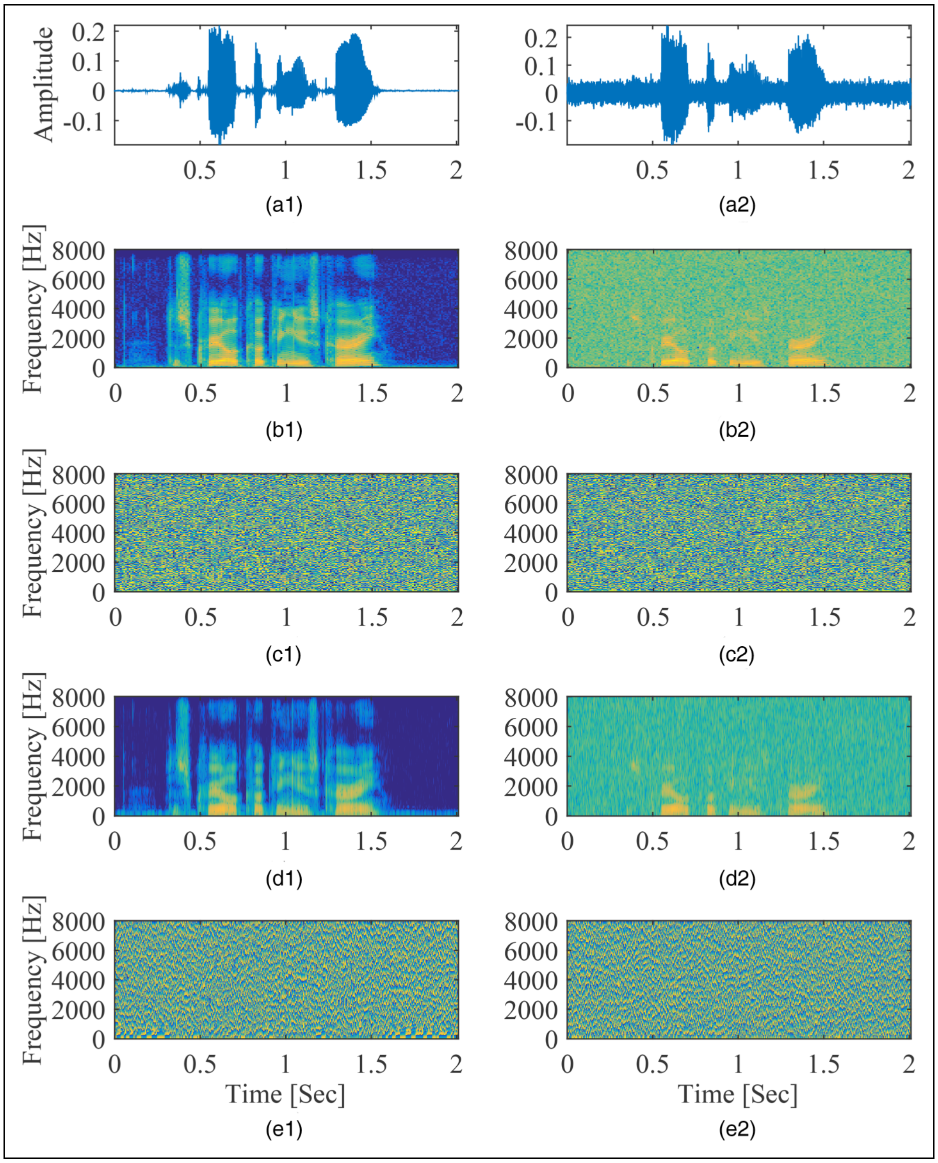

Figure 2. Time-domain clean speech and noisy speech, and their corresponding magnitude and phase spectrograms. (a_1) time-domain clean speech; (b_1) magnitude spectrogram of clean speech with frame length 320 and frame shift 160; (c_1) phase spectrogram of clean speech with the same frame length and shift as (b_1); (d_1) magnitude spectrogram of clean speech with frame length 64 and frame shift 16; (e_1) phase spectrogram of clean speech with the same frame length and shift as (d_1). (a_2) to (e_2) correspond to (a_1) to (e_1) with noisy speech at SNR = 5 dB.

phase gradient heap iteration (PGHI), for phase reconstruction from the STFT magnitude. PGHI is based on the simple relationship between the partial derivatives of the phase and log-spectral magnitude when a Gaussian window is used when computing the STFT coefficients (Portnoff, 1979). This method was computationally efficient, and achieved competitive performance when compared with many state-of-the-art algorithms. As described earlier, a higher PESQ score can be achieved when both spectral amplitude and phase are estimated. Unfortunately, because the spectral amplitude of clean speech is not estimated accurately using traditional methods, the accuracy of the phase estimation is limited, which in turn limits the contribution of phase estimation in improving speech quality.

As discussed above, the use of the last three assumptions largely explains why traditional methods do not work well in nonstationary noise scenarios. The performance degradation in low SNR environments can largely be explained by use of the third assumption. Because each T-F spectral amplitude bin is assumed to be independent of other bins, the SNR for each T-F bin is the only physical quantity used to determine whether or not that T-F bin contains speech. It is difficult to estimate the *a priori* SNR in low SNR environments. Most methods are biased towards underestimation in order to reduce musical noise. Without exploiting the structure of clean speech in time and frequency, the low SNR T-F bins containing speech cannot be recovered. In the next section, deep learning methods are reviewed. These methods implicitly relax the unrealistic assumptions, thus improving speech enhancement performance.

Before deep learning methods became the most popular methods for speech enhancement, researchers proposed many methods for solving the problems of traditional methods, including nonnegative matrix factorization (NMF) (Lee & Seung, 1999; Wilson et al., 2008; Mohammadiha et al., 2013; Sun et al., 2015) and K-means singular value decomposition (K-SVD) (Aharon et al., 2006), which is a method for factoring a matrix. For supervised NMF approaches, the speech and noise basis matrices[5] are respectively learned from the speech and noise datasets in a first stage. The noisy NMF coefficients are then obtained from the noisy speech magnitude and the two basis matrices. Finally, speech denoising is performed using the two basis matrices and their corresponding coefficients (Mohammadiha et al., 2013). For unsupervised NMF approaches, the noise basis matrix is learned from the noisy speech directly without use of the noise dataset. Mohammadiha et al. (2013), using objective assessment metrics, showed that both supervised and unsupervised approaches outperformed traditional methods such as Wiener filtering and the MMSE estimator of speech discrete-time Fourier transform coefficients based on a generalized Gamma distribution (Erkelens et al., 2007). For K-SVD-based approaches, noise dictionaries are trained using K-SVD, and they are then applied to speech denoising (Aharon et al., 2006). K-SVD-based approaches often work better than traditional methods in terms of noise attenuation.

Shallow neural networks (Tamura, 1989) and codebook-based methods were applied to speech denoising some time ago (Zavarehei et al., 2007; Suhadi et al., 2011). For example, Zavarehei et al. (2007) proposed pretraining the harmonic-noise model (HNM) codebook using a dataset comprising only clean speech. A codebook-mapping algorithm was then developed to create the estimated clean speech with the pretrained HNM codebook. However, while these methods improved speech enhancement in some specific scenarios, their ability to generalize to other scenarios, such as different types of background noise, was limited. This limitation mainly comes from the limited modeling ability of these methods due to the limited maximum number of dictionaries/basis matrices or the limited number of hidden layers and the difficulty of training a model with the limited computational resources and limited storage available at that time. Moreover, training models with greater numbers of hidden layers and more hidden units per layer was challenging before an effective initialization method was proposed by Hinton & Salakhutdinov (2006). In the next section, we review deep learning methods to clarify how these problems have been alleviated.

Deep Learning Methods

Over the last 15 years, deep learning methods have become pervasive, and they have been successfully applied to computer vision (He et al., 2016; Krizhevsky et al., 2017; Huang et al., 2017), speech processing (Yu & Deng, 2011; Hinton et al., 2012; Dahl et al., 2012; Abdel-Hamid et al., 2014), and other practical applications because of their powerful high-dimensional nonlinear modeling capability. About ten years ago, deep learning was extended to monaural speech enhancement. Many effective network architectures were proposed for denoising (Wang & Wang, 2012, 2013) and dereverberation (Han et al., 2014). Figure 3 shows a flow-chart of typical deep learning-based frequency-domain methods. These methods involve two stages: training and testing. For the training stage, there are four modules: feature extraction, network architecture, learning target, and loss function. For the testing stage, there are also four modules. The feature extraction and network architecture modules are the same as for the training stage. The target spectrum reconstruction and time-domain speech reconstruction modules are used to generate the processed time-domain speech signal. The last two modules of the testing stage are straightforward to realize. Therefore, in the following, only the four modules of the training phase are reviewed and discussed.

Feature Extraction

Feature extraction is the first step for deep learning methods. A good feature set can improve the discrimination of speech from noise. Chen et al. (2014) extracted a range of acoustic features from speech in noise at low SNRs and evaluated their influence on the classification of T-F bins as speech-dominated or noise-dominated. Classification accuracy was assessed in terms of hits (the proportion of correctly identified speech-dominated T-F bins) and false alarms (the proportion of noise-dominated T-F bins that were incorrectly classified as speech dominated). It was concluded that features derived using a gammatone filterbank, which is intended to represent the frequency analysis that takes place in the human auditory system (Moore, 2013), achieved better performance than other types of features, such as perceptual linear prediction (PLP) (Hermansky, 1990), power normalized cepstral coefficients (Kim & Stern, 2016), and Gabor filterbank features. Delfarah & Wang (2017) evaluated several acoustic features of speech in noise, including the amplitude modulation spectrogram as used by Kim et al. (2009), the relative spectral transform-PLP (RASTA) as used by Hermansky & Morgan (1994), gammatone frequency cepstral coefficients, Mel-Frequency Cepstral Coefficients (MFCCs) as used by Xu et al. (2017), log-amplitude spectral features (LOG-AMP) as used by Han et al. (2015), and fundamental-frequency-based features as used by Hu (2006), using the short-time objective

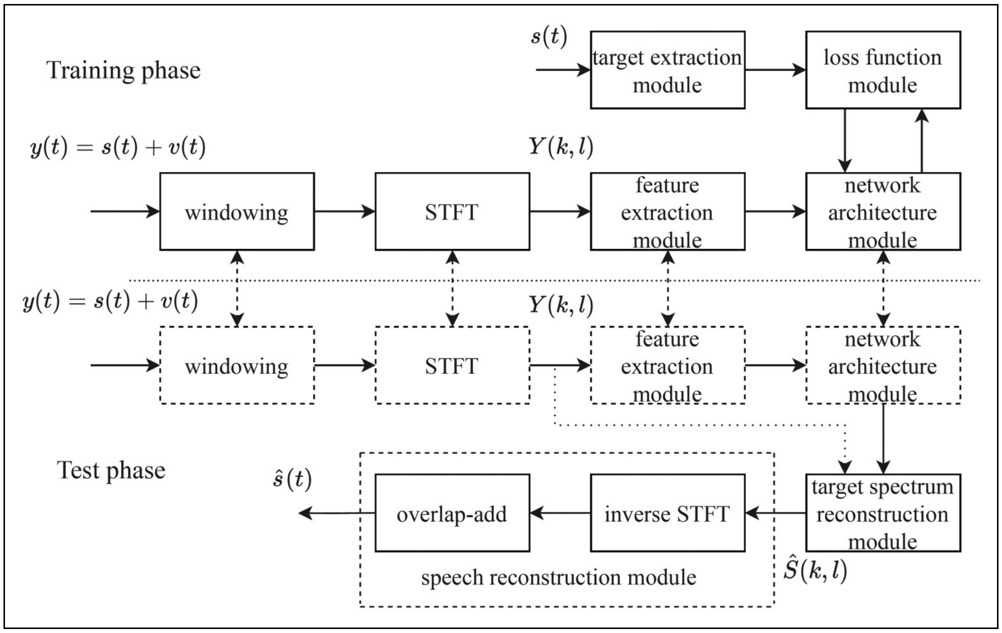

Figure 3. A generic flow-process diagram of deep learning methods.

intelligibility (STOI) score as the objective performance metric. The time required to extract each feature was also determined. Gammatone-domain features led to higher STOI scores than LOG-AMP, log-mel filterbank features and MFCC. However, the gammatone-domain features required more extraction time than the other features. Perhaps because of this, gammatone-domain features have not been used as widely as MFCC and LOG-AMP features for applications in resource-limited devices and real-time systems.

Various easily extracted features of noisy speech have been used in deep neural network (DNN) models designed to extract clean speech from noisy speech, including the LOG-AMP, the log-power spectrum (Xu et al., 2014b, 2015), spectral amplitudes (Tan & Wang, 2018) and the spectral amplitudes raised to a power less than 1 (Zhao et al., 2020), which represents a form of amplitude compression. The cube-root of the spectral amplitudes generally led to the best performance, perhaps because taking the cube-root reduces the dynamic range of the speech, facilitating the training process (Luo et al., 2022). Tan & Wang (2020) extracted the real and imaginary parts of the complex spectrum of noisy speech as input features. The mapping targets were the corresponding real and imaginary parts of the complex spectrum of the clean speech. Better performance was achieved than with networks that only mapped spectral magnitudes, because of the implicit phase recovery (Tan & Wang, 2020). It was also shown that compression of the complex spectrum improved speech dereverberation performance based on both objective metric scores and subjective preference scores (Li et al., 2021d). Better performance with compression in the joint denoising and dereveberation task was shown in the 3rd deep noise suppression (DNS) Challenge (Li et al., 2021a).

Extraction of STFT-domain features is computationally efficient, so their computational load is often small relative to the load of the DNN itself. However, the number of features increases with increasing frame length, and fine frequency resolution, which facilitates the discrimination of speech and noise, requires a large frame length. It is generally accepted that the frequency resolution of the human auditory system can be characterized using the ERB-Number frequency scale, for which the bandwidths of the auditory filters in Hz increase with increasing center frequency (Moore, 2013). The Bark scale (Fastl & Zwicker, 2007) has similar properties, but differs from the ERB-Number scale in resolution at low frequencies. These psychoacoustically based scales have been applied to reduce the number of STFT-domain features. For example, Valin (2018) extracted 22 Bark-frequency cepstral coefficients (BFCCs), the first and second derivatives of the first six BFCCs, and the strength of the dominant periodicity for the first six bands, which is related to the fundamental frequency (F0), and used them as input features for real-time speech enhancement. This reduced the number of features per frame from 481 for the STSAs (when the sampling rate was 48,000 Hz and the window length was 20 ms) to 42. Valin (2018) also computed and used at input features the period (1/F0) and a spectral nonstationary metric that measured the spectrum variation over time. To improve performance, Valin et al. (2020) used 70 input features. To extract these features, the audible frequency range was split into 34 bands based roughly on the ERB-number scale, and the magnitude of each band for the third future frame and the pitch correlation of each band for the current frame were computed and used to obtain 68 input features. The other two input features were the period (1/F0) and an estimate of the correlation of the period across frames. The computational load was about 40 million floating point operations per second when 42 features per frame were used (Valin, 2018) and about 800 million multiply-accumulate operations per second (Valin et al., 2020) when 70 features were used. Valin et al. (2020) showed that the quality of the enhanced speech was significantly improved in terms of mean opinion score (MOS) and PESQ when compared with the previous work by Valin (2018).

Another advantage of using perceptually based features occurs when full-band[6] speech is to be enhanced. For full-band speech, the number of the input features increases markedly when features are extracted on a linear frequency scale, resulting in much greater computational complexity. The number of input features can be reduced by extracting them on a logarithmic frequency scale, or by using ERB-based or Bark-based filter banks to extract input features (Schröter et al., 2022).

F0-based features have been widely used in computational auditory scene analysis to separate speech from simultaneous talkers, but their effectiveness in enhancing speech in noise is poor at low SNRs because of inaccurate estimation of F0-based features (Wang & Chen, 2018). Furthermore, as shown by Delfarah & Wang (2017), the extraction of F0-based features with a 64-channel gammatone filterbank followed by F0 estimation for each channel takes longer than the extraction of other features. Some researchers have used F0 information only implicitly. For example, Valin (2018) extracted the correlation of F0 estimates over time for the first six bands as supplementary features. Wang et al. (2022) extracted F0 information as an additional feature for predicting a mask as a postprocessing filter, further suppressing residual noise between harmonics of voiced speech and reducing speech distortion.

Comparison of the effectiveness of input features is often conducted using a specific DNN. A specific feature set may yield improved performance when used with a given DNN but may not with another DNN. In recent years, the importance of magnitude compression and phase has been confirmed using several different DNNs (Williamson & Wang, 2017; Zhao et al., 2020; Tan & Wang, 2020). As a result, the compressed complex spectrum has become popular. There are many benefits of magnitude compression in practical applications (Li et al., 2021d). The first is that while noisy speech is typically quantized with 16-bit resolution, only 8 bits are needed to represent the compressed magnitude if the compression exponent is 0.5. This leads to reduced computational complexity and memory requirements, which are important for practical applications of DNNs. The second benefit is that speech distortion and noise reduction can be better balanced when compared with the raw magnitude without compression, resulting in improving speech quality. This may be the case because compression reduces the dynamic range of the magnitude values, facilitating the training process (Luo et al., 2022). Compression of the magnitude of the noisy spectrum can be expressed by:

$$|Y_{\mathrm{cp}}(k, l)| = |Y(k, l)|^{\alpha_{\mathrm{cp}}}, \tag{31}$$

where $\alpha_{\mathrm{cp}} \in (0\ 1]$ is the compression factor. In Equation (18), α_g ranges from 0 to 2 (Sim et al., 1998; Loizou, 2005; Breithaupt et al., 2008), while α_{cp} here is often set to $1/2$ or $1/3$ for deep learning methods (Zhao et al., 2020; Li et al., 2021d). It is interesting that both deep learning and traditional methods, such as spectral subtraction, usually achieve better performance when the spectral magnitude is compressed (Inoue et al., 2011; Zhao et al., 2020; Li et al., 2021d). The compressed input complex spectrum can be expressed as:

$$\begin{aligned} Y_{\mathrm{cp}}(k, l) &= |Y_{\mathrm{cp}}(k, l)| \exp(j\angle Y(k, l)) \\ &= Y_{\mathrm{cp,r}}(k, l) + jY_{\mathrm{cp,i}}(k, l), \end{aligned} \tag{32}$$

where $Y_{\mathrm{cp}}(k, l)$ is the compressed complex spectrum of the noisy speech with real part $Y_{\mathrm{cp,r}}(k, l)$ and imaginary part $Y_{\mathrm{cp,i}}(k, l)$. These compressed spectra are used as the input features in some of the evaluations presented later in this paper.

Network Architecture

Early DNN-based speech enhancement methods operated in the T-F domain by enhancing the magnitude spectrum but leaving the phase spectrum unaltered. These methods focussed on capturing information in changes over time and in patterns across frequency. To capture information related to changes over time, Xu et al. (2014b) concatenated up to 11 frames and used the concatenated frames as input features to a fully connected DNN in which each node of the previous layer was connected to all nodes of the current layer. However, the use of such long features increases the number of input features and limits the generalization capabilities of fully connected DNNs. Recurrent neural networks (RNNs), which are naturally suitable for temporal modeling, were then proposed for speech enhancement. The main difference between fully connected DNNs and RNNs is that although both have connections between nodes, only the latter allow the output from some nodes to influence subsequent input on the same nodes. Returning to the speech production model specified by Equation (29), each time sample of speech can be recursively estimated from its previous t_P time samples. To capture the temporal characteristic of speech over time scales from samples to seconds, long short-term memory (LSTM)-based models, first proposed by Hochreiter & Schmidhuber (1997), have been applied to both masking-based (Chen & Wang, 2017) and mapping-based (Sun et al., 2017) approaches (the difference between the two types of approach is discussed later). Multiple-target learning, which jointly learns the clean spectrum and the mask, has also been used to develop an

LSTM-based speech enhancement method (Sun et al., 2017). Both objective and subjective results showed that exploiting the temporal characteristic of speech with RNNs or LSTMs led to better performance (Chen et al., 2016; Healy et al., 2021). Valin (2018) proposed an RNN called RNNoise as a low-complexity speech enhancement method that uses a gate recurrent unit (GRU[7])-based model with many fewer parameters than the LSTM-based model.

Recently, deep learning-based speech enhancement has benefited immensely from the use of convolutional neural networks (CNNs), which were inspired from biology and first proposed by Fukushima (1980) and developed extensively by LeCun et al. (1989). For CNNs, at least one convolutional layer is included in the architecture, performing a dot product of a convolution kernel with the input of this layer. The number of parameters of CNNs is markedly lower than for fully connected DNNs, although this does not necessarily reduce computational complexity. Park & Lee (2017) utilized a convolutional encoder-decoder network for spectral mapping, achieving comparable performance to an RNN-based model, while having many fewer trainable parameters. Convolutional recurrent networks (CRNs) that combine CNNs and RNNs or LSTMs into one model were first proposed by Pinheiro & Collobert (2014), and they have become one of the most popular architectures for image and speech processing, because they combine the feature-extraction capability of CNNs and the temporal modeling capability of RNNs or LSTMs. Naithani et al. (2017) developed a CRN by successively stacking convolutional layers, LSTM layers and fullyconnected layers. Takahashi et al. (2018) developed a CRN that combines convolutional layers and recurrent layers at multiple scales. The two CRN models proposed by Naithani et al. (2017) and Takahashi et al. (2018) achieved higher signal-to-distortion ratios (SDRs) than the simple combination of CNN and RNN models.

The approaches described above enhanced the noisy speech only in the magnitude domain. However, considerable improvements in both objective and subjective measures of speech quality can be achieved by recovering the phase of the clean speech (Paliwal et al., 2011). To this end, Williamson et al. (2016) employed stacked fully connected layers to estimate a complex ideal ratio mask (cIRM), which is applied to the real and imaginary parts of the spectrum to recover the magnitude and phase of the clean speech (the definition of cIRM is presented later). Subsequently, Fu et al. (2017) proposed a CNN for estimating the real and imaginary parts of the clean complex spectrum from the noisy features. Tan & Wang (2019) first applied CRNs to complex spectrum mapping-based speech enhancement, an approach they called GCRN. Two CRNs were used to separately estimate the real and imaginary parts of the clean complex spectrum and a gate mechanism was employed for both the encoder and decoder. The RNN utilized in GCRN did not effectively model extremely long sequences because of the problem of vanishing gradients and exploding gradients. These refer to situations where the error that is to be minimized changes hardly at all with changes in model parameters (vanishing gradients) or changes considerably with small changes in model parameters (exploding gradients). Vanishing gradients prevent tuning of the model parameters, while exploding gradients cause very large changes in the model parameters, resulting in the training being unstable and divergent. To deal with this, Le et al. (2021) proposed DPCRN, in which the RNN was replaced by a dual-path RNN. This dual-path RNN contained an intra-chunk RNN that was used to model the spectrum of a single time frame over frequency and an inter-chunk RNN that was used to model the variation of the spectrum over time.

Although the above-mentioned networks generated the complex spectrum of the clean speech or the cIRM, they themselves applied real-valued multiplication and addition operations. Choi et al. (2019) first introduced complex convolutional layers into the real-valued U-Net, which is a convolutional network architecture, proposed by Ronneberger et al. (2015) for speech enhancement. State-of-the-art performance was achieved in terms of many objective metrics, such as CSIG, CBAK and COVL (Hu & Loizou, 2008). Hu et al. (2020) proposed a complex CRN, called DCCRN, and reported significant performance improvements in both objective metric scores and subjective listening scores over the LSTM and CRN models, as well as over the baseline model (NSNet) provided by the DNS Challenge organizer (Xia et al., 2020). The DCCRN model was ranked first in the first DNS Challenge organized by Microsoft (Reddy et al., 2020).

Recently, approaches that decompose the speech enhancement task into several progressive tasks have been developed. These are called progressive task-oriented training approaches. The processing in earlier stages can improve the optimization of later stages. Experiments conducted using the Voice Bank + DEMAND dataset showed a large objective improvement relative to earlier methods without progressive learning, such as MMSE-GAN (Soni et al., 2018). Gao et al. (2016) first introduced the progressive learning concept into speech enhancement. They decomposed the mapping from noisy to clean speech into multiple stages so as to increase the SNR progressively. Inspired by Gao et al. (2016), a subband decomposition-based progressive learning framework was proposed by Li et al. (2021e). The progressive learning approaches proposed by Gao et al. (2016) and Li et al. (2021e) improved PESQ and STOI scores for SNR values ≤ 0 dB relative to comparable deep learning frameworks without progressive learning.

A related but somewhat different approach is to break down the speech-enhancement task into separate goals, and to optimize these goals sequentially or in parallel. For example, Yin et al. (2020) proposed a network called PHASEN for recovering magnitude and phase in a parallel processing topology. Fu et al. (2022) proposed Uformer, in which a U-Net based dilated complex-and-real dual-path conformer network was designed to improve performance in the complex-valued and magnitude domains simultaneously. Some other networks use a sequential processing structure to decompose the speech-enhancement task. For example, Strake et al. (2019) used a first task of noise suppression and a second task of restoring the parts of the speech that had been removed during noise suppression. A similar task-splitting method was used by Hao et al. (2020), who called the two tasks "masking and inpainting".

Another task-decoupling approach is to enhance spectral magnitude in the frequency domain in a first stage and then to further reduce noise in the time domain. For example, Westhausen & Meyer (2020) combined two signal transformation methods, i.e., the STFT and a learnable analysis/synthesis method, achieving competitive results in the Interspeech2020 DNS-Challenge (Reddy et al., 2021). Note that this learnable analysis/synthesis method was introduced for feature extraction with a data-driven approach; the idea had already been proposed by Luo & Mesgarani (2019) for speech separation. This decoupling method was independently proposed by Du et al. (2020). These authors designed a cascade framework including a Mel-domain denoising autoencoder[8] for magnitude recovery and a generative vocoder for waveform synthesis. The denoising autoencoder was first proposed by Vincent et al. (2008) for image processing to improve the robustness of feature extraction. It is often used to extract features for model training. Using the denoising autoencoder for speech enhancement led to improved model generalizability across different noisy conditions (Lu et al., 2013; Yu et al., 2020). Many other autoencoders have been introduced and applied to speech enhancement (Leglaive et al., 2018, 2020; Bie et al., 2022).

Different from the above decoupling approaches, Li et al. (2021b) proposed a complex spectral refinement network. A magnitude spectral estimation network was used to recover phase implicitly. This framework, called SDDNet, achieved state-of-the-art performance in the ICASSP2021 DNS-Challenge. Subsequently, Wang et al. (2021) pointed out that the estimated spectral magnitude and phase are related, and showed that it is better to estimate the spectral magnitude and the residual complex spectral component. It can be inferred that estimating spectral magnitudes first and then recovering phase is not necessarily the best choice. Wang et al. (2022) introduced a magnitude refinement module after estimation of the complex spectrum of the clean speech. Li et al. (2022b) and Fu et al. (2022) adopted parallel structures to optimize both the complex spectrum and the magnitude. Other examples of decoupling-style frameworks include decomposing the speech-enhancement task into noise estimation and speech recovery, and performing these tasks using parallel and serial structures, respectively. Zheng et al. (2021), and Liu et al. (2021) decomposed the speech enhancement process into two stages: in the first, the magnitude of the noise was estimated and used as *a priori* information for the second stage, which estimated the complex spectrum of the clean speech. Zhang et al. (2021) proposed a dual-branch framework for spectrum and waveform modeling. All of the above-mentioned decoupling methods achieved better performance in terms of PESQ and STOI or Extended STOI (ESTOI) scores than methods mapping learning targets directly.

Training Target

The training target plays an important role in deep learning methods. A well-defined training target is important for obtaining good speech intelligibility and quality. The training target should be suitable for supervised learning. Many training targets have been developed in the T-F domain. They can be divided into two main groups i.e., masking-based and mapping-based.

One masking-based target is the ideal binary mask (IBM) (Wang & Wang, 2013). For each T-F bin, the value of the IBM is either 1 or 0. A value of 1 means that the estimated SNR for this bin is larger than a predefined threshold value, while a value of 0 means that the estimated SNR is smaller than this threshold. The IBM is applied to the T-F matrix for that frame, effectively selecting the T-F bins that are to be preserved. The IBM can be expressed as

$$IBM(k,l) = \begin{cases} 1, \ |S(k,l)| \geq \theta_{\text{th}}|V(k,l)| \\ 0, \text{ otherwise} \end{cases} \tag{33}$$

where θ_{th} is the threshold, which typically has a value ranging from 0.5 to 1 in amplitude units (corresponding to $-6\,\text{dB}$ to $0\,\text{dB}$). The IBM labels every T-F unit as either target-dominated or noise-dominated, and the speech enhancement task is treated as a supervised classification problem. Use of the IBM results in good speech intelligibility but only moderate speech quality (Wang et al., 2014). Instead of using a hard threshold for each T-F unit, the ideal ratio mask (IRM) proposed by Narayanan & Wang (2013) applies an attenuation to each T-F bin that increases as the estimated SNR for that bin decreases. The IRM is defined as

$$IRM(k,l) = \left(\frac{|S(k,l)|^{\alpha_{\text{irm}}}}{|S(k,l)|^{\alpha_{\text{irm}}} + |V(k,l)|^{\alpha_{\text{irm}}}}\right)^{\beta_{\text{irm}}}, \tag{34}$$

where $\alpha_{\text{irm}} \geq 0$ and $\beta_{\text{irm}} \geq 0$ are tunable parameters. β_{irm} is often set to 0.5. The IRM leads to better speech quality than the IBM. The IBM and IRM operate in the magnitude domain; the phase is not taken into consideration (Narayanan & Wang, 2013). Erdogan et al. (2015) first proposed a phase-sensitive mask (PSM) that takes phase into account. The PSM is defined as

$$PSM(k,l) = \frac{|S(k,l)|}{|Y(k,l)|}\cos\Phi_{\Delta}, \tag{35}$$

where Φ_{Δ} denotes the phase difference between the clean speech and noisy speech within a given T-F unit. The introduction of the phase difference in the training target led to a higher SNR of the enhanced speech (Erdogan et al., 2015). Later, Williamson et al. (2016) noted that both the real and imaginary parts of the complex spectrum of speech have a clear structure, while when the magnitude and phase are treated separately a clear structure exists only in the magnitude spectrum. Accordingly, Williamson et al. (2016) proposed a cIRM, which can be regarded as an extension of the IRM to the complex-valued domain. The cIRM is defined as

$$cIRM(k,l) = M_r(k,l) + iM_i(k,l), \tag{36}$$

where $M_r(k, l)$ and $M_i(k, l)$ are, respectively, given by:

$$M_r(k,l) = \frac{Y_r(k,l)S_r(k,l) + Y_i(k,l)S_i(k,l)}{Y_r^2(k,l) + Y_i^2(k,l)}, \tag{37}$$

$$M_i(k,l) = \frac{Y_r(k,l)S_i(k,l) - Y_i(k,l)S_r(k,l)}{Y_r^2(k,l) + Y_i^2(k,l)}. \tag{38}$$

As can be seen from Equation (36), the cIRM has real and imaginary parts, like the complex spectrum of speech, and thus it should be able to recover the phase of the clean speech, unlike the IBM and IRM.

For mapping-based targets, a spectral representation of the clean speech is mapped directly using a pretrained deep-learning model. In much early work, mapping-based supervised speech-enhancement methods focussed on mapping from the noisy magnitude spectrum to the clean magnitude spectrum while leaving the phase unaltered. Because the dynamic range of the magnitude is large, the log-power spectrum was initially proposed as the mapping target (Xu et al., 2015, 2014b). However, Zhao et al. (2020) showed that power compression of the magnitude led to better performance than logarithmic compression. More recently, motivated by the fact that appropriate use of phase information could notably improve speech quality, especially at low SNRs (Paliwal et al., 2011), complex-valued spectrum mapping has become mainstream. Li et al. (2021d) proposed combining the noisy phase with the power-compressed magnitude. The compressed complex spectrum $S_{\text{cp}}(k, l)$ in this case can be expressed by

$$\begin{aligned} S_{\text{cp}}(k,l) &= |S_{\text{cp}}(k,l)| \exp(j\angle S(k,l)) \\ &= S_{\text{cp,r}}(k,l) + jS_{\text{cp,i}}(k,l), \end{aligned} \tag{39}$$

where $|S_{\text{cp}}(k, l)| = |S(k, l)|^{\alpha_{\text{cp}}}$. $S_{\text{cp,r}}(k, l)$ and $S_{\text{cp,i}}(k, l)$ are the real and imaginary parts of $S_{\text{cp}}(k, l)$, respectively. This compressed complex spectrum has become popular as a training target because it leads to better performance for both objective and subjective measures than the uncompressed complex spectrum or magnitude-based training targets.

Some researchers have used mapping targets in addition to the spectrum or mask (Xu et al., 2017; Fang et al., 2023). Xu et al. (2017) proposed a DNN that learned the magnitude spectrum as the primary target and MFCC as the secondary target. The additional MFCC estimation imposed constraints that were not applicable in the prediction of the magnitude spectrum alone, improving the prediction performance of the primary target. Fang et al. (2023) proposed a framework for jointly modeling random uncertainties and uncertainties due to insufficient training data for deep-learning-based Wiener filter estimation for speech enhancement. The involvement of modeling uncertainties increased the robustness of the estimator, and it was shown that this method preserved more speech at the cost of decreasing the amount of noise reduction slightly.

Loss Function

For deep-learning approaches, the loss function, also known as the cost function, evaluates how well a model is currently performing on a specific dataset, and tunes the model parameters based on the gradient of performance with respect to these parameters (Yu & Deng, 2011; Hinton et al., 2012). The training loss function is one of the key components for deep learning methods, and it is commonly used to assess how well the trained model fits the training data. An appropriate loss function can lead to a good balance between performance on the one hand and storage memory and/or computational complexity on the other hand. The commonly used loss functions for speech enhancement can be divided into three categories: frequency-domain, time-domain, and perceptually motivated.

In the frequency domain, early work used a neural network to map the IBM, and used binary cross entropy to supervise the model training. As mentioned above, for each T-F bin the value of the IBM is either 1 or 0, and thus this binary cross entropy measures the dissimilarity between the predicted and true labels of a dataset. Later, with the increasing complexity of training targets, the speech-enhancement task began to be regarded as an approximation or regression problem rather than a classification problem, and commonly used loss functions were the Kullback-Leibler divergence proposed by Kullback (1997), the Itakura-Saito distance proposed by Itakura & Satio (1968), and the MSE. Liu et al. (2014) showed that both the Kullback-Leibler divergence and the Itakura-Saito distance led to worse performance than the MSE, and since then the MSE has become a very popular loss function for monaural speech enhancement. The models are often trained by minimizing the MSE between the magnitude/complex spectrum of the clean speech and the estimated spectrum, a method denoted signal approximation (SA), rather than by minimizing the MSE between the true mask and the estimated mask (Huang et al., 2014; Weninger et al., 2014; Wang & Chen, 2018). Better objective metric scores have been obtained with SA (Weninger et al., 2014) than when the MSE between the estimated mask and the true one was minimized, provided that the same network architecture was used. Spectral magnitude-based MSE (Weninger et al., 2015), phase-sensitive MSE (Erdogan et al., 2015) and complex spectrum-based MSE (Fu et al., 2017) have all been proposed. The spectral magnitude-based MSE and complex spectrum-based MSE loss functions can be expressed as:

$$\mathcal{L}_{Mag} = \| |S| - |\widehat{S}| \|_F^2, \tag{40}$$

and

$$\mathcal{L}_{RI} = \|S_r - \widehat{S}_r\|_F^2 + \|S_i - \widehat{S}_i\|_F^2, \tag{41}$$

where $|\cdot|$, $(\cdot)_r$ and $(\cdot)_i$ extract the spectral magnitude, and the real and imaginary parts of the complex spectrum, respectively. $\| \bullet \|_F$ represents the Frobenius norm, defined as the square root of the sum of the squares of all the elements of a matrix or a vector. When the spectral magnitude-based MSE loss function is used, spectral magnitude distortion is minimized and the phase is unaltered. When the complex spectrum-based MSE loss function is used, the phase estimation error is reduced but spectral magnitude distortion increases. The trade-off between spectral magnitude distortion and phase recovery has been called the "compensation effect" (Wang et al., 2021; Luo et al., 2022). To reduce both magnitude and phase distortion, a combined loss function has been proposed, which is formulated as:

$$\mathcal{L}_{RI+Mag} = \alpha_{\text{com}}\mathcal{L}_{Mag} + (1 - \alpha_{\text{com}})\mathcal{L}_{RI}, \tag{42}$$

where α_{com} is a linear combination coefficient. A comprehensive evaluation of different loss functions was conducted by Braun & Tashev (2021). They showed that combining the magnitude and phase-aware losses led to performance improvement, even when only the noisy phase was used to reconstruct the time-domain speech, indicating the importance of including the phase-aware loss in the loss function in the training stage. Moreover, using compressed spectrum loss functions yielded further significant performance improvement.

The mean absolute error (MAE) between the estimated and clean speech magnitudes (Qi et al., 2020) has also been explored as a complex spectral distance metric, based on the assumption that the real and imaginary parts of each STFT bin follow a Laplacian distribution rather than a Gaussian distribution (Braun & Tashev, 2021). Tu et al. (2018) used a log-spectral MSE loss function for magnitude estimation, because human perception of auditory magnitudes roughly corresponds to a logarithmic scale. More recently, a power-law-compressed spectral MSE loss function (Yin et al., 2020) has been shown to be effective in denoising. This method can be formulated as:

$$\begin{aligned}\mathcal{L}_{cprs} = (1 - \alpha)(\| |S|^c \cos(\phi_S) - |\widehat{S}|^c \cos(\phi_{\widehat{S}}) \|_F^2 \\ + \| |S|^c \sin(\phi_S) - |\widehat{S}|^c \sin(\phi_{\widehat{S}}) \|_F^2) \\ + \alpha \| |S|^c - |\widehat{S}|^c \|_F^2.\end{aligned} \tag{43}$$

It should be noted that the time-domain loss functions commonly used for separating simultaneous talkers have been extended to speech enhancement, and have been shown to be effective (Rethage et al., 2018; Choi et al., 2019; Hu et al., 2020). In addition to time-domain MAE (or MSE) loss functions (Rethage et al., 2018), signal energy-based loss functions have become popular, such as the SDR and scale invariant SDR loss functions (Choi et al., 2019; Hu et al., 2020). In addition, restricting the SNR loss to a certain range was proposed for use with a time-domain progressive speech enhancement approach (Nian et al., 2022). The restricted SNR loss was introduced to preserve the relative values of the clean speech target, intermediate speech target, and input noisy speech.

Perceptually motivated loss functions have been developed to optimize DNNs based on objective metrics such as PESQ, cepstral distance, and STOI (Zhao et al., 2019). As described earlier, when used singly, these loss functions generally achieve better performance when assessed with the metric used for optimization (Fu et al., 2020), but not for other metrics, while combining these loss functions often results in improvement of multiple objective metric scores simultaneously (Fu et al., 2017).

For deep learning approaches, "overfitting" can occur when the number of training parameters is too large relative to the size of the training dataset. When this happens, performance evaluated using the training dataset is much better than performance evaluated using an independent but similar dataset. Overfitting can be avoided by stopping the training early (Wang & Chen, 2018). In addition, a learnable loss mix-up approach, in which two loss functions were combined together with a learnable parameter, has been proposed to improve the generalization of DNN-based speech-enhancement models (Chang et al., 2021). The purpose of introducing the learnable parameter was to automatically optimize the weighting factor of the two loss functions with the training dataset in the training stage.

Discussion

As described above, unrealistic assumptions limit the performance of traditional frequency-domain monaural speech enhancement methods. In contrast, deep learning methods do not depend on any specific assumptions about the properties of the speech or noise. For example, it is well known that deep learning methods can handle nonstationary noises better than traditional methods, because the former are not based on assumptions about the statistics of the noise, while the latter are often based on the assumption that the noise is stationary or quasi-stationary. Also, deep learning methods can utilize speech contextual information and the speech spectral structure to reconstruct clean speech or to separate speech from noise. The earliest deep learning speech enhancement methods estimated only the spectral magnitude of the clean speech, ignoring the importance of speech phase for speech quality and speech intelligibility. This was done because it was thought that speech phase is unstructured, and thus cannot be learned or mapped by deep learning methods.

However, the real and imaginary parts of the complex spectrum both have structure.

The speech phase can be implicitly recovered in two ways: one is to estimate the cIRM (Williamson et al., 2016) and the other is to estimate the real and imaginary parts of the complex speech spectrum (Wang & Chen, 2018). Note that these two ways have some similarities: the former often maps the cIRM directly and uses the MSE between the true and estimated cIRM as the loss function (Williamson et al., 2016), while the latter often maps the real and imaginary parts of the complex speech spectrum and uses the MSE between the true and estimated complex spectrum as the loss function. The loss function is often based on the distance between the mapped target and the true one. For example, Tan & Wang (2020) used the MSE between the true and estimated complex spectrum as the loss function for training, while the mapped target was the cIRM. However, this is not a mandatory requirement. For example, Choi et al. (2019) proposed mapping the cIRM, but a time-domain loss function that considers both the speech prediction error and noise prediction error, namely the weighted SDR, was used to train the model.

Although most deep learning methods train models using parallel noisy and clean speech, this is not always necessary. Some researchers have trained models using nonparallel noisy and clean data and achieved moderate performance in terms of both objective and subjective measures (Xiang & Bao, 2020; Xiang et al., 2020; Yu et al., 2022; Bie et al., 2022), although performance did not surpass that of models trained with a large amount of parallel data. Deep learning methods trained without parallel data are often referred to as *unsupervised methods*, while those trained with parallel data are referred to as *supervised methods*. Bie et al. (2022) categorized unsupervised methods into two types, noise-dependent and noise-agnostic. The latter methods require only clean speech, and thus in principle they can handle any type of noise (Bando et al., 2018; Bie et al., 2022), while the former methods also require noisy speech or noise although the noisy and clean speech signals are not necessarily parallel. Unsupervised noise-agnostic methods (Bando et al., 2018; Bie et al., 2022) learn the characteristics of the clean speech from the clean speech dataset in the training stage, and the noise signal is only modeled in the test stage. Some NMF methods also learn the speech basis matrix using clean speech and the noise basis matrix is estimated in the denoising stage using information from the speech basis matrix (Mohammadiha et al., 2013).

One interesting question is: what information do deep learning methods use to learn the mapping from noisy to clean speech. It is nearly always the case that the speech and noise have different T-F characteristics, so these differences can be used to distinguish speech from noise or to estimate the SNR for each bin. For monaural speech enhancement, important differences are: the noise is often more stationary than the speech; the speech has a more sparse spectral structure than the noise; in voiced segments, the F0 does not change rapidly, resulting in a correlation over time, and the harmonic structure shows a patterning in frequency; and voiced and unvoiced segments tend to alternate, unvoiced segments having most of their energy at mid and high frequencies and voiced segments having most of their energy at low and mid frequencies. These differences are presumably exploited by deep-learning methods.

Environmental noise can be roughly divided into four categories: industrial, construction, traffic, and noise resulting from social activities (Han et al., 2018). These noise types have different T-F characteristics, and, unlike speech, noise cannot be treated in a unified way and described by a simple model. Hence, it is relatively difficult for DNNs to learn the characteristics of noise. However, when speech is mixed with noise, if the speech can be reconstructed from its corresponding noisy version, the noise can then be estimated by subtracting the estimated complex spectrum of the speech from that for the noisy speech. This information can in turn be used to improve the speech enhancement process.

Zhang et al. (2020) estimated the *a priori* SNR using the Deep Xi model proposed by Nicolson & Paliwal (2019). The noise PSD was then estimated using the MMSE-based approach proposed by Gerkmann & Hendriks (2012). Xu et al. (2014a) have shown the benefits of noise-aware training for DNN-based speech enhancement. In the noise-aware training, information about the noise, such as its short-term spectrum, is used to provide supplementary features. This leads to better mapping of the clean speech spectrum. Liu et al. (2021) and Liu et al. (2022) further showed that estimating the noise spectrum in a first stage improved the performance of single and multimicrophone denoising and dereverberation methods.

A special case is when the background sound is competing speech from one or more talkers. In this case, the problem reduces to speech separation, which is outside the scope of this paper. However, speech from more than one talker often occurs in social interactions, and this is challenging for speech enhancement. A recent study showed that a DNN designed for single-talker speech enhancement performed poorly when it was required that speech from two talkers was to be preserved simultaneously in noisy and reverberant environments (Zheng et al., 2023). This happened because the speech from two talkers is very different from the speech of a single talker. For example, for two overlapping voiced segments there are two simultaneous F0s, and the spectral pattern is much more complex than for a single talker. Increasing the number of talkers generally results in worse performance of DNN-based methods. It seems that most current DNN methods are mainly applicable to single-talker situations. To reduce speech distortion for multiple desired speakers, the training dataset should contain multiple simultaneous talkers and the mixed speech should be regarded as the target when training the DNN (Zheng et al., 2023).

Traditional and deep learning methods differ in several ways in how they reconstruct clean speech. Firstly, the estimated short-term SNR for each T-F bin determines the IRM for mask-based deep learning methods, while the *a priori* SNR is the key parameter determining the gain function for traditional methods. For the latter, “musical noise” is a serious problem due to the large variability in the estimate of the spectrum for a given frame when time-averaging or frequency-averaging are not used. If the short-term SNR instead of the expected SNR can be accurately estimated for each T-F bin, “musical noise” can be largely eliminated. Unfortunately, it is extremely difficult if not impossible to estimate the short-term SNR using traditional methods, making it difficult to achieve a good balance between the suppression of “musical noise” and the reduction of speech distortion. A second difference between traditional and deep learning methods is that traditional methods estimate the spectral magnitude and phase to reconstruct the time-domain speech signal, while most deep-learning methods estimate the real and imaginary parts of the complex clean speech or estimate the spectral magnitude in a first stage and then estimate the residual complex component of the clean speech in a second stage. In other words, deep learning methods often recover the phase implicitly and indirectly, while traditional methods estimate the phase explicitly and directly. Note that phase reconstruction from the STFT magnitude using DNN approaches has been proposed by Oyamada et al. (2018), Masuyama et al. (2019), Takamichi et al. (2020), Masuyama et al. (2021), and Peer et al. (2022). The effectiveness of DNN-based phase reconstruction has been shown by the reconstruction of clean speech/ audio phase without knowledge of the phase spectrogram. However, although DNN-based phase reconstruction has been applied to speech separation (Wang et al., 2018), only a few researchers have explicitly estimated the clean speech phase for the purpose of enhancing speech in noise (Peer & Gerkmann, 2022).

Finally, while deep learning methods suffer less from the musical noise problem than traditional methods, the former can introduce artificial noise components that degrade speech quality. This artificial noise needs to be suppressed to achieve good speech perceptual quality. Li et al. (2021a) showed that a low-complexity spectral subtraction method used in a postprocessing stage to suppress the artificial noise improved subjective preference scores, although PESQ scores were somewhat decreased. It is interesting that DNSMOS scores (Reddy et al., 2021) correctly reflected subjective preference scores, supporting the effectiveness of DNSMOS in estimating subjective speech quality.

Hybrid Methods

As mentioned above, traditional methods cannot suppress nonstationary noise completely because the noise PSD cannot be accurately tracked, especially when the noise PSD increases rapidly or fluctuates strongly over time. One straightforward way to alleviate this problem is to improve noise tracking using data-driven methods (Erkelens & Heusdens, 2008). Another way is to jointly estimate the speech PSD and the noise PSD, so that Wiener-type filtering can be applied to the noisy speech (Zavarehei et al., 2007; Suhadi et al., 2011; Mohammadiha et al., 2013). Although nonstationary noise is better removed using these methods, the speech distortion for low SNR T-F bins still occurs, thus limiting the quality of speech recovery.

Although traditional and deep-learning methods are quite different, the latter have been considerably influenced by the former. Also, there are some hybrid methods for which key parameters extracted using traditional methods are mapped using DNN methods. For example, mapping the *a priori* SNR using the Deep Xi DNN proposed by Nicolson & Paliwal (2019, 2020) and integrating it into the MMSE-STSA estimator led to good speech quality. Another approach is to combine DNNs with NMF (see Bando et al., 2018, Bie et al., 2022, and references therein). The results showed better performance in terms of objective metrics than for both supervised and unsupervised noise-dependent models for unseen noise scenarios. A complete survey of hybrid models is outside the scope of this paper.

Evaluation of Different Methods

Datasets

To evaluate the performance of the different types of speech enhancement methods, we used two datasets, namely WSJ + DNS and Voice Bank + Demand (Valentini-Botinhao et al., 2016). The WSJ + DNS training dataset was generated using the WSJ0-SI84 and Interspeech 2020 DNS-Challenge noise datasets. They contain 150,000 mixtures of speech and noise together with the corresponding clean speech, amounting to about 300 hours in total. The clean utterances were selected from the WSJ0-SI84 dataset (Paul & Baker, 1992), which comprises 7138 utterances spoken by 83 speakers (42 males and 41 females). 5428 and 957 utterances spoken by 77 speakers were selected for model training and model validation, respectively. 150 utterances spoken by the remaining 6 speakers (3 males and 3 females) were used for testing. This was done to assess the generalization capability of the models for different speakers. The noise clips, which include about 20,000 noise types, were selected from the noise set of the Interspeech 2020 DNS-Challenge dataset (Reddy et al., 2021). Their total duration was about 55 hours. To create a training mixture, a randomly selected training utterance was mixed with a randomly cut segment from the noises at a given SNR. The SNR range for training was −5 to 0 dB in steps of 1 dB. Testing used a stationary white Gaussian noise and two highly nonstationary noises, namely babble and factory1, all taken from NOISEX92

(Varga & Steeneken, 1993). SNRs of −5, 0, 5, and 10 dB were used for the test set. 150 noisy-clean pairs were generated for each noise type and each SNR.

Voice Bank + Demand includes 11,572 utterances for training. 824 utterances were used for testing. The clean speech set was a selection of 30 speakers taken from the Voice Bank corpus (Veaux et al., 2013): 28 speakers were used for training and the remaining 2 speakers were used for testing. To create the noisy training set, 40 conditions were used: 10 types of noise (2 artificial noises and 8 noises taken from the Demand database (Thiemann et al., 2013)), each mixed with clean speech at SNRs of 15, 10, 5, and 0 dB. There were about 10 sentences in each condition for each training speaker. For the test set, 20 conditions were used: 5 types of noise (all from the Demand database) with 4 SNRs (17.5, 12.5, 7.5, and 2.5 dB). There were about 20 sentences in each condition for each test speaker. Note that the test set was totally unseen because it used different speakers, noises, and SNRs from the training set.

Parameter Values

All of the utterances were sampled at 16 kHz. For all of the models except the four models marked with asterisks in tables, the window duration and hop size were 20 ms and 10 ms, respectively, and thus the FFT length was 320. Pytorch was used to train the models, based on the Adam optimizer (β_1 = 0.9, β_2 = 0.999) (Kingma & Ba, 2014). The initial learning rate was set to 0.001, and it was divided by 2 if three consecutive validation loss[9] increments occurred. Training was stopped early if five consecutive validation loss increments occurred. 60 epochs were used for network training, and the batch size was set to 16 at the utterance level. For magnitude-spectrum mapping DNNs, two training targets were studied: one was the uncompressed magnitude and the other was the compressed magnitude with $\alpha_{cp} = 0.5$. For complex-spectrum DNNs, there were also two training targets. One was the real and imaginary parts of the uncompressed complex spectrum and the other was the real and imaginary parts of the compressed complex spectrum $S_{cp}(k, l)$, as presented in Equation (39), with the compression coefficient $\alpha_{\text{cp}} = 0.5$.

Methods Evaluated

The DNNs that were evaluated can be categorized into three groups: magnitude-spectrum based, complex-spectrum based, and decoupling style. For each group, the STFT spectrum or compressed spectrum of the noisy speech were used as the input features, and the training targets were the corresponding STFT spectrum or compressed spectrum of the clean speech. For all models, spectral MSE loss functions were utilized. All of the models were consistent with the best configurations reported by the authors of those models.

For completeness, six representative traditional methods were evaluated, including the MMSE-STSA estimator (Ephraim & Malah, 1984), the MMSE-LSA estimator (Ephraim & Malah, 1985), the β-order MMSE-STSA estimator (with $\beta = 0.5$) (You et al., 2005), the magnitude spectral subtraction (MSS) (Boll, 1979), power spectral subtraction (PSS) (Lim & Oppenheim, 1979), and square root of MSS (SQ-MSS) models (Inoue et al., 2010b). Note that the MMSE-STSA and MMSE-LSA estimators are two special cases of the β-order MMSE-STSA estimator: β-order MMSE-STSA becomes MMSE-STSA when $\beta = 1$, while it reduces to MMSE-LSA when $\beta \to 0$. In the following, the β-order MMSE-STSA estimator is referred to as MMSE-STSA(β) for brievity. These methods required estimation of the noise PSD. The MMSE-based method proposed by Gerkmann & Hendriks (2012) was used for noise PSD estimation. The subtraction factor β_g described in Equation (19) was set to different values for different spectral subtraction methods. Its value was adjusted so as to give a noise reduction of 17 dB, which led to a good balance between musical noise and noise reduction. In previous studies the amount of noise reduction was set to range from 12 to 25 dB for normal-hearing listeners (Cohen, 2003; Inoue et al., 2011) and from 14 to 20 dB for hearing-impaired listeners who wore hearing aids (Wong et al., 2018). The value of 17 dB used here was chosen to fall in the middle of these ranges. The six traditional methods are denoted group 1.

Two hybrid methods were chosen, namely DeepXi-LSA and DeepXi-STSA. These are denoted group 2.

1. DeepXi-LSA: The *a priori* SNR is estimated using the Deep Xi framework proposed by Nicolson & Paliwal (2019), and the *a posteriori* SNR is computed as the estimated *a priori* SNR plus one. With the estimated *a priori* SNR and *a posteriori* SNR, the spectral gain can be computed using Equation (24).
2. DeepXi-STSA: As for DeepXi-LSA, the *a priori* SNR and *a posteriori* SNR are estimated, and the spectral gain is then computed using Equation (23). Note that DeepXi-LSA and DeepXi-STSA use the same network architecture to estimate the *a priori* SNR. The only difference is in the method of determining spectral gain from the estimated *a priori* SNR and *a posteriori* SNR. The window duration and hop size of DeepXi-LSA and DeepXI-STSA were set to 32 ms and 16 ms, respectively, so as to be consistent with the parameter settings of the model proposed by Nicolson & Paliwal (2019).

For magnitude-spectrum mapping DNNs, three one-stage causal models were selected and denoted group 3, namely LSTM (Sun et al., 2017), FullSubNet (Hao et al., 2021), and CRN (Tan & Wang, 2018):

1. LSTM: the LSTM-based model contains four LSTM layers, each with 1024 units. The output layer is a 161-unit fully connected layer. The input and output are the noisy and estimated clean speech magnitudes, respectively.
2. FullSubNet: FullSubNet is a full-band and sub-band fusion model, which was ranked 11th in the ICASSP 2021 DNS-Challenge. Full-band and sub-band refer to models that input full-band and sub-band noisy spectral features and output full-band and sub-band speech targets, respectively. The window duration and hop size were set to 32 ms and 16 ms, respectively, so as to be consistent with the parameter settings of the model proposed by Hao et al. (2021).
3. CRN: CRN contains an encoder and a decoder with two LSTM layers. The input and output are the magnitude of the noisy and estimated speech, respectively.

Five single-stage complex spectrum mapping DNN-based causal models were chosen as group 4, namely GCRN (Tan & Wang, 2020), DPCRN (Le et al., 2021), Uformer (Fu et al., 2022), and DCCRN (Hu et al., 2020):

1. GCRN: GCRN is a complex spectral mapping network based on CRN, where both the spectral magnitude and phase are estimated. GCRN comprises one encoder and two decoders, and both the real and imaginary components of the clean speech complex spectrum are estimated.
2. DPCRN: DPCRN was ranked second in the Interspeech 2021 DNS-Challenge. The RNNs in the CRN were replaced by dual-path RNN modules, where the intra-chunk RNNs were used to model the spectral pattern of a single frame and the inter-chunk RNNs were used to model the inter-dependence of successive frames.
3. Uformer: Uformer is a U-Net based dilated complex and real dual-path conformer network, which processes speech in both complex and magnitude domains. The window duration and hop size were set to 25 ms and 10 ms, respectively, and the FFT length was 512, so as to be consistent with the parameter settings of the model proposed by Fu et al. (2022).
4. DCCRN: DCCRN was ranked first in the Interspeech 2020 DNS-Challenge. Both convolutional encoder-decoder and LSTM structures are handled by complex-valued operations. The window duration and hop size were set to 32 ms and 8 ms, respectively, so as to be consistent with the parameter settings of the model proposed by Hu et al. (2020).
5. DCCRN(SNR): DCCRN(SNR) has the same network architecture, input features and learning targets as DCCRN. The only difference is in their loss functions. DCCRN(SNR) used the SNR loss instead of the MSE loss. The SNR loss was used because initial evaluations showed that a DCCRN model trained using the MSE loss performed much worse than a DCCRN model trained with the SNR loss for the Voice Bank + DEMAND dataset, although the two models had comparable performance for the WSJ + DNS dataset.

Three decoupling-style multi-stage causal frameworks were chosen as group 5: i.e., CTSNet (Li et al., 2021b), G2Net (Li et al., 2022b), and TaylorSENet (Li et al., 2022a):

1. CTSNet: CTSNet is a two-stage pipeline that initially generates a coarse estimate of the magnitude spectrum and then predicts a complex residual component that is used to refine the coarse estimate. CTSNet was ranked first in the ICASSP 2021 DNS-Challenge.
2. GaGNet: GaGNet is an improved version of CTSNet. It uses a parallel structure for coarse magnitude and refined complex spectrum estimation, enabling bifurcate and joint optimization to estimate the complex spectrum of the clean speech.
3. TaylorSENet: TaylorSENet is a decoupling-style framework based on Taylor's approximation theory. The complex target spectrum is reconstructed by the superimposition of 0th-order and higher-order network estimations following Taylor's formula.

The models listed above were chosen to be representative of a wide range of models of different types. They were not chosen because they gave the best performance for a given type of processing. The models were trained and evaluated using the same datasets and using the same evaluation metrics, thus allowing a fair comparison across methods for both simulated normal-hearing and simulated hearing-impaired listeners. This allowed a fair assessment of the improvement achieved by each advance in signal-processing method.

Evaluation Metrics

Four objective metrics were chosen to estimate speech quality for normal-hearing listeners, namely PESQ (Rix et al., 2001), the extended STOI (ESTOI) (Jensen & Taal, 2016), the SDR (Vincent et al., 2007), and DNSMOS (Reddy et al., 2021). Note that the first three metrics are intrusive, requiring knowledge of the clean speech to compute their scores. The DNSMOS is nonintrusive, and it is used to evaluate and rank different DNS methods. Its predictions correspond well with MOS values (Reddy et al., 2021). The hearing-aid speech quality index (HASQI) version 2 proposed by Kates & Arehart (2014) and hearing-aid speech perception index (HASPI) version 2 proposed by Kates & Arehart (2021) were chosen as two objective measures to evaluate speech quality and intelligibility, respectively, for both normal-hearing listeners and hearing-impaired listeners.

PESQ scores are usually highly correlated with human estimates of speech quality. Both the narrow-band version and the wideband version recommended in ITU-T P.862 were used, and the two versions are denoted NB-PESQ and WB-PESQ, respectively. Note that NB-PESQ uses the raw scores provided by P.862 directly, while WB-PESQ maps raw scores to the MOS-listening quality objective (MOS-LQO) domain. NB-PESQ scores and WB-PESQ scores therefore have different ranges. NB-PESQ scores range from −0.5 to 4.5, while WB-PESQ scores range from 1.0 to 4.5.

The ESTOI is another measure that is widely used to evaluate speech intelligibility. Its value ranges from 0 to 1 (Jensen & Taal, 2016; Zhao et al., 2018). Only raw ESTOI scores are provided in this paper. In other words, we did not map the ESTOI scores to intelligibility because the parameters of this mapping depend on many factors, such as the test material and the test paradigm (Jensen & Taal, 2016).

The SDR is a time-domain metric that is widely used in blind speech separation and can also be applied for speech quality evaluation. DNSMOS is a robust speech quality metric serving as a proxy for subjective scores, which consists of three sub-metrics, i.e., DNS-OVL, DNS-SIG, and DNS-BAK, evaluating overall signal quality, signal distortion, and background distortion, respectively. The values of the three DNSMOS metrics range from 1 to 5. For all these metrics, a higher score indicates better performance.

The HASQI and HASPI can be used to evaluate speech quality and speech intelligibility, respectively, for both simulated normal-hearing and hearing-impaired listeners when the audiometric thresholds of the simulated listener are given (Kates & Arehart, 2014, 2021). The two metrics use a model of the auditory periphery. For normal-hearing listeners, the audiometric thresholds were set to 0 dB HL at all frequencies required by HASQI and HASPI (250, 500, 1000, 2000, 4000, and 6000 Hz). For hearing-impaired listeners, the audiometric thresholds were specified as greater than 0 dB HL and the auditory model was modified so as to take into account some of the typical consequences of hearing loss, such as reduced frequency selectivity (Glasberg & Moore, 1986) and reduced compression in the cochlea (Moore et al., 1996). In this paper, only mild and moderate hearing losses were considered. Bisgaard et al. (2010) divided standard audiograms into two groups: A flat and moderately sloping group, and a steeply sloping group. The first group included seven audiograms characterizing different degrees of hearing loss, while the second group had three audiograms with different degrees of hearing loss. Here, two standard audiograms, N2 and N3, with mild and moderate sloping hearing losses, respectively, were used, as shown in Figure 4. Both the HASQI and HASPI provide the option of applying frequency-dependent gain to compensate for the reduced audibility produced by the hearing loss, based on the NAL-R method (Byrne & Dillon, 1986). That option was used here. The software packages for calculating HASQI and HASPI scores were used with the default setting that a signal with a root-mean-square (RMS) value of 1 has a level of 65 dB SPL. In all tests, each signal was normalized to have an RMS value of 1, so the effective input level was 65 dB SPL. The NAL-R recommended gains are suitable for this level. The values of the HASQI and HASPI, expressed as a percentage, range from 0% to 100%, where higher scores indicate better performance.

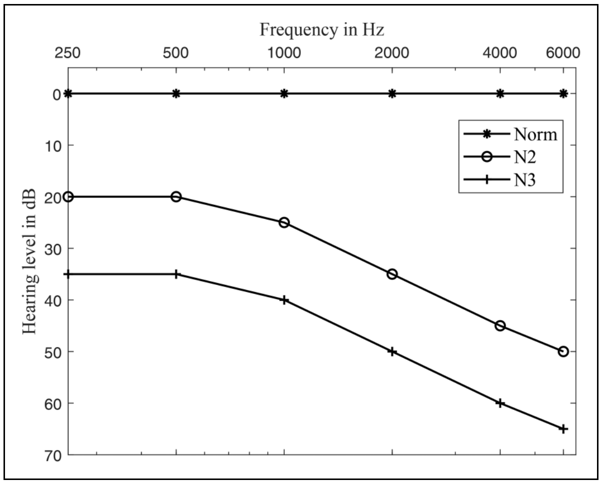

Figure 4. Standard audiograms for two of the audiograms defined by Bisgaard et al. (2010): Norm: normal, N2: mild sensorineural loss, N3: moderate sensorineural loss.

Results and Analysis

Results for the WSJ + DNS Dataset

Simulated Normal-hearing Listeners. Tables 1–6 give the results for the five groups of speech enhancement methods. Several observations can be made. First, all deep-learning methods and hybrid methods achieved much better performance than traditional methods, regardless of whether or not the input feature, magnitude or complex spectrum, was compressed. Second, almost all deep-learning methods performed better when the compressed input features were used. Third, when the phase was implicitly estimated by deep-learning methods, better performance was achieved than when only spectral magnitudes were processed, and this gap became larger when compressed input features were used. Fourth, when the magnitude was mapped in the first stage and the residual complex spectrum of the clean speech was recovered in the second stage, performance was further improved. Fifth, a more recent method did not always work better than earlier ones when the input feature changed from the compressed complex spectrum to the uncompressed complex spectrum. For example, TaylorSENet worked better than CTSNet when the input features were compressed, while CTSNet worked better than TaylorSENet when no compression was applied. A similar effect was observed for GCRN and DCCRN; their performance was comparable when the input

Table 1. Scores for the objective measures WB-PESQ and NB-PESQ for the different methods using the WSJ + DNS dataset.

		WB-PESQ					NB-PESQ				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	Noisy	1.09	1.22	1.47	1.84	1.41	1.42	1.74	2.10	2.47	1.93
	MMSE-LSA	1.09	1.22	1.47	1.84	**1.41**	1.70	2.12	2.50	2.86	**2.30**
	MMSE-STSA (1)	1.09	1.22	1.47	1.84	**1.41**	1.70	2.12	2.50	2.87	**2.30**
	MMSE-STSA (0.5)	1.09	1.22	1.47	1.82	1.40	1.70	2.12	2.49	2.84	2.29
	MSS	1.07	1.19	1.42	1.85	1.38	1.65	2.07	2.47	2.85	2.26
	PSS	1.08	1.19	1.40	1.78	1.36	1.65	2.06	2.44	2.82	2.24
	SQ-MSS	1.07	1.17	1.42	1.88	1.39	1.63	2.04	2.45	2.83	2.24
	DeepXi-LSA*	1.28	1.61	2.03	2.50	**1.85**	2.06	2.56	2.92	3.25	**2.70**
	DeepXi-STSA*	1.26	1.58	1.98	2.45	1.82	2.05	2.53	2.88	3.20	2.66
	LSTM	1.23	1.49	1.84	2.28	1.71	1.98	2.42	2.77	3.08	2.56
	FullSubNet*	1.33	1.68	2.12	2.60	**1.93**	2.24	2.65	3.00	3.30	**2.80**
	CRN	1.22	1.47	1.81	2.20	1.67	1.96	2.41	2.78	3.11	2.57
	GCRN	1.35	1.72	2.16	2.61	1.96	2.22	2.70	3.03	3.33	2.82
	DPCRN	1.25	1.55	1.95	2.41	1.79	2.03	2.52	2.89	3.22	2.66
	Uformer*	1.42	1.82	2.27	2.69	**2.05**	2.32	2.77	3.12	3.39	**2.90**
	DCCRN*	1.35	1.73	2.20	2.69	1.99	2.19	2.69	3.06	3.37	2.83
	DCCRN*(SNR)	1.28	1.62	2.10	2.72	1.93	2.14	2.64	3.02	3.37	2.80
	CTSNet	1.49	1.96	2.44	2.90	**2.20**	2.40	2.86	3.20	3.46	**2.98**
	G2Net	1.45	1.88	2.31	2.74	2.09	2.34	2.81	3.14	3.41	2.92
	TaylorSENet	1.44	1.88	2.34	2.79	2.11	2.31	2.78	3.13	3.42	2.91
Cafe	noisy	1.09	1.18	1.37	1.65	1.32	1.54	1.86	2.20	2.53	2.03
	MMSE-LSA	1.09	1.18	1.37	1.65	1.32	1.58	1.98	2.36	2.72	2.16
	MMSE-STSA (1)	1.08	1.18	1.37	1.65	1.32	1.59	1.98	2.36	2.72	2.16
	MMSE-STSA (0.5)	1.08	1.18	1.36	1.64	1.32	1.58	1.97	2.35	2.71	2.15
	MSS	1.08	1.17	1.37	1.67	1.32	1.59	1.98	2.36	2.73	2.17
	PSS	1.08	1.17	1.34	1.62	1.30	1.57	1.95	2.33	2.69	2.13
	SQ-MSS	1.08	1.17	1.38	1.71	**1.33**	1.61	2.00	2.38	2.75	**2.19**
	DeepXi-LSA*	1.29	1.63	2.05	2.51	**1.87**	2.09	2.58	2.96	3.29	**2.73**
	DeepXi-STSA*	1.28	1.60	2.00	2.45	1.83	2.07	2.55	2.92	3.24	2.70
	LSTM	1.26	1.51	1.87	2.28	1.73	2.02	2.46	2.83	3.12	2.61
	FullSubNet*	1.33	1.70	2.15	2.63	**1.95**	2.21	2.66	3.01	3.32	**2.80**
	CRN	1.23	1.48	1.80	2.17	1.67	1.99	2.46	2.83	3.15	2.61
	GCRN	1.42	1.82	2.26	2.65	2.04	2.31	2.75	3.09	3.36	2.88
	DPCRN	1.27	1.57	2.96	2.39	1.80	2.04	2.55	2.92	3.24	2.69
	Uformer*	1.44	1.86	2.29	2.71	**2.08**	2.35	2.82	3.16	3.43	**2.94**
	DCCRN*	1.38	1.80	2.27	2.74	2.05	2.20	2.74	3.11	3.43	2.87
	DCCRN*(SNR)	1.36	1.76	2.26	2.78	2.04	2.21	2.71	3.10	3.43	2.86
	CTSNet	1.58	2.09	2.58	2.98	**2.31**	2.48	2.95	3.27	3.51	**3.05**
	G2Net	1.56	2.00	2.43	2.81	2.20	2.45	2.90	3.21	3.45	3.00
	TaylorSENet	1.54	2.00	2.45	2.83	2.20	2.42	2.87	3.21	3.47	2.99
Babble	noisy	1.11	1.23	1.45	1.73	1.38	1.52	1.83	2.16	2.49	2.00
	MMSE-LSA	1.11	1.23	1.45	1.73	**1.38**	1.62	2.02	2.38	2.74	2.19
	MMSE-STSA (1)	1.11	1.23	1.44	1.73	**1.38**	1.62	2.02	2.38	2.74	2.19

(continued)

Table 1. Continued.

		WB-PESQ					NB-PESQ				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
	MMSE-STSA (0.5)	1.11	1.23	1.44	1.72	**1.38**	1.62	2.02	2.37	2.73	2.19
	MSS	1.09	1.20	1.42	1.74	1.36	1.60	2.00	2.37	2.73	2.17
	PSS	1.09	1.20	1.40	1.69	1.34	1.57	1.96	2.33	2.69	2.14
	SQ-MSS	1.09	1.19	1.41	1.76	1.36	1.63	2.02	2.39	2.74	**2.20**
	DeepXi-LSA*	1.22	1.52	1.95	2.45	**1.79**	1.85	2.41	2.85	3.20	**2.58**
	DeepXi-STSA*	1.21	1.50	1.91	2.39	1.75	1.83	2.39	2.81	3.15	2.54
	LSTM	1.19	1.43	1.78	2.20	1.65	1.81	2.31	2.71	3.03	2.46
	FullSubNet*	1.24	1.56	2.03	2.54	**1.84**	1.97	2.47	2.87	3.21	**2.63**
	CRN	1.18	1.42	1.74	2.11	1.61	1.78	2.32	2.73	3.06	2.47
	GCRN	1.32	1.69	2.16	2.57	1.93	2.03	2.58	2.99	3.29	2.72
	DPCRN	1.21	1.48	1.89	2.34	1.73	1.78	2.38	2.82	3.15	2.53
	Uformer*	1.33	1.75	2.21	2.64	**1.98**	2.07	2.66	3.06	3.35	**2.79**
	DCCRN*	1.29	1.65	2.12	2.60	1.91	1.95	2.55	2.99	3.32	2.70
	DCCRN*(SNR)	1.27	1.64	2.15	2.69	1.94	1.96	2.54	2.98	3.32	2.70
	CTSNet	1.44	2.00	2.54	2.95	**2.23**	2.22	2.83	3.21	3.45	**2.93**
	G2Net	1.39	1.84	2.35	2.79	2.10	2.17	2.73	3.12	3.39	2.85
	TaylorSENet	1.38	1.85	2.35	2.78	2.09	2.15	2.70	3.10	3.39	2.83

For all deep learning methods, the uncompressed spectrum was used. Bold font indicates the best average score in each group. DNS = deep noise suppression; SNR = speech-to-noise ratio; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction; LSTM = long short-term memory; CRN = convolutional recurrent network.

features were uncompressed, but DCCRN performed better when the input features were compressed.

Hybrid methods obtained comparable performance to single-stage deep learning based magnitude-only mapping methods in terms of PESQ, ESTOI, and SDR scores, confirming that the *a priori* SNR is a key parameter for magnitude estimation in the speech-enhancement task. The hybrid methods performed more poorly than the deep-learning methods of groups 2, futher confirming the importance of phase recovery for speech enhancement. As can be seen from Table 5, hybrid methods yielded much lower DNS-MOS scores than the deep-learning methods of groups 3–5, especially for low-SNR scenarios. On the other hand, the hybrid methods worked better for relatively high SNR scenarios, and this finding was confirmed by the objective test results using the Voice Bank + DEMAND dataset, as described below. The poorer performance of hybrid methods for low-SNR scenarios may have occurred because speech distortion is inevitable using traditional spectral gain functions when many bins have low *a priori* SNR values.

The HASQI scores for simulated normal-hearing listeners denoted by "Normal" and shown in Tables 7 and 8 were consistent with the scores in Tables 1–6. Also, the HASPI scores for simulated normal-hearing listeners shown in Tables 9 and 10 were consistent with the HASQI scores for simulated normal-hearing listeners. Over the SNR range from −5 to 0 dB, the HASPI scores were significantly improved by deep learning methods for all three types of noise.

Simulated Hearing-impaired Listeners. Tables 7 and 8 present the HASQI scores for the different methods for all three simulated listener groups. The results differ in some interesting ways from those for the simulated normal-hearing listeners. First, MSS and SQ-MSS achieved better performance than MMSE-LSA. Second, all deep-learning methods gave lower or comparable HASQI scores when the input features were compressed than when they were not, which is opposite to the results for the simulated normal-hearing listeners. Third, for all of the single-stage deep-learning methods, the HASQI scores were not improved when the phase was also implicitly estimated, in contrast to the metrics for simulated normal-hearing listeners. Fourth, the most recently proposed methods did not always achieve the best performance. Fifth, for relatively high SNRs, the HASQI score for the enhanced speech was sometimes lower than that for the noisy speech, especially for the traditional methods. For the deep-learning methods, the benefit of speech enhancement reduced dramatically for high SNRs and only a small improvement was obtained when the SNR was 10 dB. Finally, the HASQI scores increased with increasing hearing loss, perhaps because the speech degradation produced by signal processing has a smaller perceptual effect for listeners with more severe hearing loss (Kates & Arehart, 2022).

Table 2. Scores for the objective measures ESTOI and SDR for the different methods using the WSJ + DNS dataset.

		ESTOI					SDR				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	0.26	0.41	0.57	0.72	0.49	−4.94	0.05	5.03	10.03	2.54
	MMSE-LSA	0.26	0.41	0.57	0.73	0.49	0.58	5.53	9.44	13.15	**7.17**
	MMSE-STSA (1)	0.26	0.42	0.58	0.73	0.50	0.54	5.50	9.45	13.19	**7.17**
	MMSE-STSA (0.5)	0.26	0.41	0.57	0.72	0.49	0.62	5.53	9.40	13.07	7.16
	MSS	0.27	0.42	0.59	0.74	0.50	−1.08	4.53	9.14	13.18	6.44
	PSS	0.26	0.42	0.58	0.73	0.50	−0.93	4.49	9.00	13.14	6.43
	SQ-MSS	0.27	0.42	0.59	0.74	**0.51**	−1.62	4.24	8.95	12.62	6.05
	DeepXi-LSA*	0.44	0.62	0.76	0.85	0.67	4.63	8.58	12.11	15.60	**10.23**
	DeepXi-STSA*	0.44	0.62	0.76	0.85	**0.67**	4.45	8.39	11.90	15.31	10.01
	LSTM	0.47	0.64	0.76	0.84	0.68	3.73	7.74	11.14	13.94	9.14
	FullSubNet*	0.48	0.66	0.78	0.86	**0.70**	5.62	9.47	13.03	16.44	**11.14**
	CRN	0.46	0.64	0.76	0.85	0.68	3.33	7.51	11.31	14.70	9.21
	GCRN	0.54	0.70	0.80	0.86	0.73	4.91	8.06	10.66	13.35	9.24
	DPCRN	0.48	0.66	0.78	0.86	0.70	3.93	8.38	12.10	15.46	9.97
	Uformer*	0.55	0.72	0.81	0.88	**0.74**	5.84	9.67	13.12	16.42	**11.26**
	DCCRN*	0.51	0.69	0.80	0.87	0.72	5.54	9.72	13.24	16.51	11.25
	DCCRN*(SNR)	0.52	0.70	0.81	0.89	0.73	4.92	9.39	13.24	16.81	11.09
	CTSNet	0.60	0.75	0.84	0.89	**0.77**	7.65	11.36	14.43	17.27	12.68
	G2Net	0.60	0.76	0.84	0.89	**0.77**	7.63	11.32	14.41	17.38	**12.69**
	TaylorSENet	0.59	0.74	0.83	0.89	0.76	7.42	11.14	14.26	17.22	12.51
cafe	noisy	0.31.	0.45	0.60	73.76	0.52	−4.92	0.04	5.03	10.03	2.55
	MMSE-LSA	0.30	0.44	0.59	0.74	0.52	−3.01	2.81	7.75	12.13	**4.92**
	MMSE-STSA (1)	0.31	0.45	0.60	0.74	0.52	−3.00	2.80	7.74	12.14	**4.92**
	MMSE-STSA (0.5)	0.30	0.44	0.59	0.73	0.52	−3.02	2.80	7.72	12.07	4.89
	MSS	0.31	0.46	0.61	0.75	**0.53**	−3.30	2.62	7.81	12.31	4.86
	PSS	0.31	0.45	0.60	0.75	**0.53**	−3.22	2.57	7.64	12.19	4.80
	SQ-MSS	0.31	0.46	0.61	0.75	**0.53**	−3.62	2.44	7.74	11.94	4.62
	DeepXi-LSA*	0.49	0.66	0.78	0.86	0.70	4.20	8.63	12.17	15.76	**10.19**
	DeepXi-STSA*	0.49	0.66	0.78	0.86	**0.70**	3.96	8.39	11.92	15.46	9.93
	LSTM	0.52	0.67	0.78	0.85	0.71	3.65	7.90	11.21	14.04	9.20
	FullSubNet*	0.52	0.69	0.80	0.88	**0.72**	4.98	9.59	13.25	16.74	**11.14**
	CRN	0.51	0.67	0.78	0.87	0.71	3.44	7.83	11.44	14.94	9.41
	GCRN	0.60	0.73	0.82	0.87	0.76	5.33	8.55	11.16	13.74	9.69
	DPCRN	0.52	0.70	0.80	0.88	0.72	3.96	8.66	12.29	15.69	10.15
	Uformer*	0.60	0.75	0.83	0.89	**0.77**	5.98	10.07	13.44	16.78	11.57
	DCCRN*	0.56	0.72	0.82	0.89	0.75	5.47	10.03	13.53	16.82	11.46
	DCCRN*(SNR)	0.58	0.74	0.84	0.90	0.76	6.17	10.69	14.30	17.62	**12.20**
	CTSNet	0.66	0.79	0.86	0.91	**0.80**	7.98	11.97	15.04	17.86	13.21
	G2Net	0.66	0.79	0.86	0.91	**0.80**	8.32	12.08	15.06	18.01	**13.37**
	TaylorSENet	0.65	0.78	0.85	0.90	0.79	8.02	11.84	14.89	17.78	13.13
Babble	noisy	0.27	0.40	0.55	0.69	0.48	−4.91	0.05	5.03	10.02	2.55
	MMSE-LSA	0.27	0.40	0.55	0.70	0.48	−0.25	4.80	8.81	12.61	**6.49**
	MMSE-STSA (1)	0.27	0.40	0.55	0.71	0.48	−0.34	4.74	8.81	12.63	6.46

(continued)

Table 2. Continued.

		ESTOI					SDR				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
	MMSE-STSA (0.5)	0.26	0.39	0.54	0.70	0.47	−0.20	4.81	8.78	12.53	6.48
	MSS	0.27	0.41	0.56	0.71	**0.49**	−1.44	4.07	8.61	12.73	5.99
	PSS	0.27	0.40	0.55	0.71	0.48	−1.52	2.88	8.40	12.63	5.85
	SQ-MSS	0.27	0.41	0.56	0.72	**0.49**	−1.61	4.01	8.52	12.23	5.79
	DeepXi-LSA*	0.43	0.61	0.74	0.84	**0.66**	2.62	7.48	11.39	15.12	**9.15**
	DeepXi-STSA*	0.43	0.61	0.74	0.84	**0.66**	2.33	7.22	11.14	14.80	8.87
	LSTM	0.46	0.63	0.75	0.83	0.67	1.58	6.41	10.39	13.52	7.98
	FullSubNet*	0.45	0.64	0.77	0.86	**0.68**	2.88	8.03	12.36	16.08	**9.84**
	CRN	0.46	0.64	0.76	0.84	**0.68**	1.47	6.56	10.63	14.27	8.23
	GCRN	0.54	0.71	0.80	0.86	0.73	3.26	7.36	10.37	13.08	8.52
	DPCRN	0.45	0.65	0.78	0.86	0.69	1.47	7.24	11.50	15.11	8.83
	Uformer*	0.53	0.71	0.81	0.88	**0.73**	3.81	8.76	12.66	16.10	**10.33**
	DCCRN*	0.50	0.68	0.79	0.87	0.71	3.30	8.55	12.54	16.10	10.12
	DCCRN*(SNR)	0.52	0.70	0.81	0.88	**0.73**	3.05	8.54	12.81	16.71	10.28
	CTSNet	0.61	0.77	0.85	0.90	**0.78**	6.12	11.03	14.46	17.24	12.21
	G2Net	0.61	0.76	0.85	0.89	**0.78**	6.17	10.97	14.53	17.51	**12.30**
	TaylorSENet	0.59	0.75	0.84	0.89	0.77	6.05	10.79	14.29	17.29	12.10

For all deep learning methods, the uncompressed spectrum was used. Bold font indicates the best average score in each group. SDR = signal-to-distortion ratio; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction; LSTM = long short-term memory; CRN = convolutional recurrent network.

Tables 9 and 10 present the HASPI scores for listeners with simulated mild and moderate hearing losses, denoted "Mild" and "Moderate," respectively. In contrast to the HASQI scores, the HASPI scores decreased with increasing hearing loss, probably because the simulated hearing loss itself leads to poorer speech intelligibility. When the input SNR was lower than 5 dB, deep learning methods improved the HASPI score in most cases, but this was not the case when the input SNR was 10 dB. Moreover, as was the case for the HASQI scores for the simulated hearing-impaired listeners, compression of the input features did not improve HASPI scores, perhaps because the reduction of speech distortion and residual noise gained by compression of the input features was not audible for the simulated hearing-impaired listeners.

Results for Voice Bank + DEMAND Dataset

Simulated Normal-hearing Listeners. Tables 11 and 12 give the results for the five groups of speech enhancement methods. Several observations can be made. First, all metric scores were relatively high and most were above the corresponding metric scores for the noisy speech signals at 10 dB SNR in the WSJ + DNS dataset. Second, when the input features were not compressed, NB-PESQ scores were sometimes lower for the deep-learning methods than for the conventional methods, while the other metric scores were much higher. Third, when the input features were compressed, the performance gap between conventional methods and deep-learning methods became larger. Fourth, estimating the phase implicitly did not improve all of the metric scores for the single-stage deep-learning methods, while performance improved markedly when decoupling-style deep-learning methods were used. Fifth, the performance of DCCRN was strongly influenced by the loss function for relatively high SNRs. When the MSE was used as the loss function, the PESQ, ESTOI, DNS-SIG scores for the enhanced speech signals were sometimes lower than those for the unprocessed noisy speech signals. However, when the SNR was used as the loss function for DCCRN, denoted DCCRN(SNR), scores markedly improved, except for the SDR. Finally, hybrid methods yielded the second-highest scores in terms of NB-PESQ, WB-PESQ, and ESTOI when the input features for the deep-learning methods were uncompressed. Because the SNR in the Voice Bank + DEMAND dataset is relatively high, the competitive performance of hybrid methods for this dataset further confirms that these methods may be more suitable for relatively high SNR scenarios.

The HASQI and HASPI scores for simulated normal-hearing listeners denoted by "Normal" are shown in Table 13. The deep learning methods only slightly increased the HASQI and HASPI scores when the input features were not compressed, and the HASQI scores were increased by only about 10% when the input features were compressed. The small benefit

Table 3. As Table 1 but using the compressed spectrum as input features for all deep learning methods.

		WB-PESQ					NB-PESQ				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	Noisy	1.09	1.22	1.47	1.84	1.41	1.42	1.74	2.10	2.47	1.93
	LSTM	1.24	1.51	1.88	2.36	1.75	1.96	2.45	2.82	3.17	2.60
	FullSubNet*	1.38	1.75	2.22	2.74	**2.02**	2.28	2.74	3.10	3.40	**2.88**
	CRN	1.24	1.49	1.87	2.32	1.73	1.97	2.44	2.84	3.17	2.60
	GCRN	1.27	1.65	2.14	2.65	1.93	1.97	2.59	3.01	3.33	2.72
	DPCRN	1.26	1.55	1.94	2.41	1.79	2.02	2.56	2.97	2.71	2.56
	Uformer*	1.54	2.03	2.53	2.99	**2.27**	2.48	2.96	3.30	3.55	**3.08**
	DCCRN*	1.46	1.97	2.48	2.93	2.21	2.35	2.91	3.26	3.52	3.01
	CTSNet	1.49	1.98	2.52	3.00	2.25	2.35	2.88	3.26	3.54	3.01
	G2Net	1.50	2.01	2.53	3.05	**2.27**	2.37	2.89	3.27	3.55	**3.02**
	TaylorSENet	1.50	1.99	2.55	3.04	**2.27**	2.37	2.89	3.27	3.55	**3.02**
Cafe	noisy	1.09	1.18	1.37	1.65	1.32	1.54	1.86	2.20	2.53	2.03
	LSTM	1.27	1.54	1.92	2.40	1.78	2.02	2.49	2.90	3.24	2.66
	FullSubNet*	1.38	1.79	2.29	2.80	**2.06**	2.26	2.75	3.12	3.44	**2.89**
	CRN	1.25	1.52	1.88	2.30	1.74	1.98	2.50	2.89	3.23	2.65
	GCRN	1.35	1.76	2.28	2.75	2.03	2.07	2.68	3.08	3.40	2.81
	DPCRN	1.29	1.61	1.99	2.45	1.83	2.06	2.62	3.01	3.35	2.76
	Uformer*	1.60	2.13	2.64	3.09	**2.36**	2.53	3.02	3.35	3.61	**3.13**
	DCCRN*	1.51	2.10	2.64	3.05	2.33	2.37	2.99	3.34	3.58	3.07
	CTSNet	1.58	2.15	2.67	3.10	2.38	2.45	3.00	3.34	3.60	3.10
	G2Net	1.63	2.20	2.72	3.15	**2.43**	2.51	3.00	3.34	3.59	3.11
	TaylorSENet	1.64	2.20	2.73	3.15	**2.43**	2.52	3.02	3.37	3.62	**3.13**
Babble	noisy	1.11	1.23	1.45	1.73	1.38	1.52	1.83	2.16	2.49	2.00
	LSTM	1.20	1.46	1.85	2.33	1.71	1.80	2.35	2.77	3.14	2.52
	FullSubNet*	1.28	1.66	2.22	2.81	**1.99**	2.01	2.58	3.02	3.36	**2.74**
	CRN	1.19	1.45	1.83	2.29	1.69	1.72	2.33	2.79	3.14	2.49
	GCRN	1.26	1.66	2.19	2.69	1.95	1.80	2.52	2.99	3.33	2.66
	DPCRN	1.22	1.52	1.97	2.44	1.79	1.78	2.42	2.92	3.27	2.60
	Uformer*	1.44	2.00	2.55	3.00	**2.25**	2.23	2.88	3.27	3.54	**2.98**
	DCCRN*	1.37	1.93	2.54	3.02	2.21	2.07	2.81	3.26	3.55	2.92
	CTSNet	1.45	2.04	2.62	3.07	2.30	2.21	2.87	3.28	3.55	2.98
	G2Net	1.43	1.98	2.60	3.09	2.27	2.21	2.81	3.25	3.53	2.95
	TaylorSENet	1.46	2.04	2.63	3.11	**2.31**	2.24	2.87	3.28	3.56	**2.99**

LSTM = long short-term memory; CRN = convolutional recurrent network.

for the HASQI scores partly reflects the fact that the HASPI scores for the unprocessed noisy speech were high, e.g. 97.1% for the simulated normal-hearing listeners. The results in Table 13 confirm that compression of the input features is important for deep learning methods when the enhanced speech is intended for normal-hearing listeners.

Simulated Hearing-impaired Listeners. The HASQI and HASPI scores for two types of simulated hearing-impaired listeners are also presented in Table 13. All of the traditional methods yielded lower (worse) HASQI and HASPI scores than for the unprocessed noisy speech. For all deep-learning methods, the HASQI and HASPI scores for the enhanced speech signals were similar to or only slightly higher than those for unprocessed signals. These negative findings for speech quality for the simulated hearing-impaired listeners could be a result of the relatively high SNR of the noisy speech in the Voice Bank + DEMAND dataset, which is probably, on average, above 10 dB, as inferred from the metric scores for the noisy

Table 4. As Table 2 but using the compressed spectrum as input features for all deep learning methods.

		ESTOI					SDR				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	0.26	0.41	0.57	0.72	0.49	−4.94	0.05	5.03	10.03	2.54
	LSTM	0.47	0.65	0.77	0.85	0.69	3.91	7.89	11.29	14.31	9.35
	FullSubNet*	0.49	0.68	0.79	0.87	**0.71**	6.22	9.94	13.43	16.78	**11.59**
	CRN	0.49	0.65	0.77	0.86	0.69	3.79	7.92	11.63	15.06	9.60
	GCRN	0.49	0.70	0.81	0.88	0.72	4.42	8.99	12.63	15.58	10.41
	DPCRN	0.50	0.68	0.80	0.87	0.71	4.59	8.85	12.39	15.55	10.34
	Uformer*	0.60	0.75	0.84	0.90	**0.77**	6.69	10.29	13.61	16.87	11.87
	DCCRN*	0.57	0.74	0.83	0.89	0.76	6.74	10.73	14.09	17.17	**12.18**
	CTSNet	0.60	0.76	0.84	0.90	0.78	7.67	11.35	14.44	17.27	12.68
	G2Net	0.61	0.76	0.85	0.90	0.78	7.70	11.34	14.43	17.36	12.71
	TaylorSENet	0.62	0.77	0.85	0.90	**0.79**	7.96	11.52	14.62	17.53	**12.91**
Cafe	noisy	0.31	0.45	0.60	0.74	0.52	−4.92	0.04	5.03	10.03	2.55
	LSTM	0.53	0.69	0.79	0.87	0.72	4.08	8.12	11.51	14.52	9.56
	FullSubNet*	0.54	0.71	0.81	0.89	**0.74**	5.90	10.11	13.61	17.05	**11.67**
	CRN	0.53	0.69	0.80	0.87	0.72	3.64	8.06	11.66	15.17	9.63
	GCRN	0.56	0.73	0.84	0.89	0.76	5.21	9.47	12.97	16.05	10.93
	DPCRN	0.55	0.71	0.82	0.89	0.74	4.82	9.18	12.54	15.80	10.59
	Uformer*	0.65	0.78	0.86	0.91	**0.80**	6.85	10.77	14.03	17.32	12.24
	DCCRN*	0.62	0.77	0.86	0.91	0.79	6.74	11.23	14.53	17.55	**12.51**
	CTSNet	0.66	0.79	0.86	0.91	0.81	8.02	11.96	14.96	17.79	13.18
	G2Net	0.68	0.80	0.87	0.91	0.81	8.47	12.10	15.02	17.89	13.37
	TaylorSENet	0.69	0.80	0.87	0.91	**0.82**	8.75	12.33	15.29	18.12	**13.62**
Babble	noisy	0.27	0.40	0.55	0.69	0.48	−4.91	0.05	5.03	10.02	2.55
	LSTM	0.47	0.65	0.77	0.85	0.69	2.11	6.88	10.73	14.04	8.44
	FullSubNet*	0.47	0.66	0.79	0.87	**0.70**	4.08	8.94	12.88	16.42	**10.58**
	CRN	0.46	0.66	0.77	0.85	0.69	1.52	6.82	10.89	14.52	8.44
	GCRN	0.51	0.72	0.82	0.88	0.73	3.50	8.74	12.51	15.45	10.05
	DPCRN	0.48	0.68	0.80	0.87	0.71	2.67	7.81	11.89	15.26	9.41
	Uformer*	0.59	0.75	0.84	0.89	**0.77**	4.74	9.47	13.15	16.56	10.98
	DCCRN*	0.55	0.74	0.84	0.90	0.76	4.62	9.86	13.72	16.98	**11.30**
	CTSNet	0.61	0.77	0.85	0.90	0.78	6.31	11.04	14.31	17.15	12.20
	G2Net	0.62	0.77	0.86	0.90	0.79	6.28	10.91	14.41	17.34	12.24
	TaylorSENet	0.63	0.78	0.86	0.91	**0.80**	6.82	11.32	14.68	17.66	**12.62**

LSTM = long short-term memory; CRN = convolutional recurrent network.

speech signals. The performance of DCCRN depended on the loss function, better performance being obtained when the SNR was used as the loss function.

With increasing hearing loss, speech quality increased, while speech intelligibility decreased, consistent with the results for the WSJ + DNS dataset. Although speech quality was somewhat improved by deep-learning methods, speech intelligibility was only marginally improved, partly because intelligibility was relatively high for the noisy speech in the Voiced Band + DEMAND dataset. However, for simulated hearing-impaired listeners with moderate loss, although the HASPI score was only 64.7% for the unprocessed noisy speech, deep learning methods still did not markedly improve intelligibility, indicating some limitation of deep learning methods. Because speech intelligibility was at least not decreased, removing noise may increase the acceptable noise level, and hearing-impaired listeners may then be more willing to wear hearing aids (Nabelek et al., 2006).

Table 5. As Table 1 but using DNS-MOS as the evaluation metric.

		DNS-OVL					DNS-SIG					DNS-BAK				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	1.24	1.69	2.23	2.57	1.93	1.53	2.40	3.04	3.28	2.56	1.27	1.77	2.54	3.13	2.18
	MMSE-LSA	1.58	1.96	2.26	2.53	2.08	2.04	2.51	2.79	2.99	2.58	2.24	2.966	3.39	3.68	3.06
	MMSE-STSA (1)	1.54	1.96	2.26	2.54	2.07	1.99	2.51	2.79	3.00	2.57	2.19	2.94	3.38	3.67	3.04
	MMSE-STSA (0.5)	1.63	1.98	2.25	2.53	**2.10**	2.13	2.53	2.78	2.99	**2.61**	2.32	2.97	3.39	3.67	**3.09**
	MSS	1.23	1.70	2.17	2.47	1.89	1.45	2.29	2.82	3.00	2.39	1.45	2.33	3.06	3.48	2.58
	PSS	1.25	1.72	2.17	2.48	1.90	1.44	2.28	2.81	3.04	2.39	1.56	2.46	3.04	3.41	2.62
	SQ-MSS	2.28	1.74	2.12	2.38	1.88	1.58	2.37	2.74	2.89	2.39	1.51	2.37	3.13	3.56	2.64
	DeepXi-LSA*	1.86	2.23	2.57	2.79	2.36	2.16	2.59	2.93	3.14	2.70	3.69	3.81	3.92	3.98	**3.85**
	DeepXi-STSA*	1.87	2.24	2.58	2.79	**2.37**	2.19	2.62	2.95	3.14	**2.73**	3.63	3.75	3.88	3.96	3.80
	LSTM	2.16	2.45	2.73	2.87	2.55	2.58	2.84	3.08	3.20	2.93	3.55	3.75	3.93	3.99	**3.80**
	FullSubNet*	2.23	2.50	2.74	2.89	**2.59**	2.70	2.91	3.11	3.23	**2.99**	3.54	3.77	3.92	3.98	**3.80**
	CRN	2.15	2.45	2.71	2.88	2.55	2.61	2.86	3.08	3.23	2.94	3.48	3.73	3.92	4.00	3.78
	GCRN	2.13	2.49	2.74	2.89	2.56	2.53	2.58	3.08	3.21	2.85	3.60	3.84	3.97	4.02	3.86
	DPCRN	2.16	2.49	2.78	2.93	2.59	2.62	2.89	3.12	3.26	2.97	3.50	3.77	3.96	4.02	3.81
	Uformer*	2.39	2.64	2.86	2.98	**2.72**	2.80	3.01	3.19	3.30	**3.07**	3.77	3.92	4.01	4.05	3.94
	DCCRN*	2.22	2.54	2.79	2.92	2.62	2.61	2.90	3.12	3.25	2.97	3.73	3.89	3.99	4.03	3.91
	DCCRN*(SNR)	2.31	2.64	2.86	2.97	2.69	2.67	2.98	3.17	3.28	3.03	3.79	3.96	4.05	4.08	**3.97**
	CTSNet	2.52	2.79	2.95	3.03	**2.82**	2.90	3.13	3.27	3.33	**3.16**	3.85	3.99	4.05	4.08	**3.99**
	G2Net	2.45	2.81	2.97	3.05	**2.82**	2.85	3.15	3.29	3.36	**3.16**	3.76	3.98	4.05	4.08	3.97
	TaylorSENet	2.30	2.67	2.89	2.99	2.71	2.76	3.04	3.22	3.31	3.08	3.60	3.88	4.00	4.04	3.88
Cafe	noisy	1.23	1.55	2.09	2.47	1.84	1.47	2.09	2.80	3.17	2.38	1.26	1.63	2.37	2.99	2.06
	MMSE-LSA	1.54	1.89	2.23	2.51	**2.04**	1.95	2.46	2.88	3.10	**2.60**	2.20	2.71	3.12	3.43	2.86
	MMSE-STSA (1)	1.54	1.89	2.24	2.51	**2.04**	1.95	2.47	2.89	3.10	**2.60**	2.20	2.71	3.11	3.42	2.86
	MMSE-STSA (0.5)	1.54	1.90	2.23	2.51	**2.04**	1.95	2.47	2.87	3.09	**2.60**	2.22	2.73	3.13	3.43	**2.88**
	MSS	1.42	1.77	2.19	2.47	1.96	1.76	2.36	2.88	3.10	2.52	1.90	2.44	2.97	3.32	2.66
	PSS	1.46	1.80	2.19	2.48	1.98	1.78	2.38	2.89	3.13	2.55	2.09	2.54	2.93	3.25	2.70
	SQ-MSS	1.38	1.74	2.14	2.41	1.92	1.74	2.33	2.79	2.99	2.46	1.71	2.34	3.01	3.42	2.62
	DeepXi-LSA*	1.92	2.28	2.61	2.84	2.41	2.27	2.67	2.99	3.19	2.78	3.58	3.72	3.86	3.96	**3.78**
	DeepXi-STSA*	1.94	2.29	2.61	2.83	**2.42**	2.31	2.70	3.00	3.20	**2.81**	3.51	3.67	3.82	3.92	3.73
	LSTM	2.25	2.53	2.78	2.91	2.62	2.70	2.93	3.14	3.25	3.01	3.511	3.75	3.92	3.99	**3.79**
	FullSubNet*	2.26	2.54	2.77	2.93	2.62	2.82	3.01	3.17	3.28	**3.07**	3.30	3.64	3.87	3.97	3.70
	CRN	2.29	2.55	2.79	2.94	**2.64**	2.79	2.98	3.15	3.28	3.05	3.46	3.72	3.91	4.00	3.77
	GCRN	2.24	2.55	2.81	2.92	2.63	2.66	2.93	3.14	3.25	3.00	3.61	3.84	3.98	4.03	3.87
	DPCRN	2.26	2.57	2.83	2.97	2.66	2.76	2.99	3.20	3.31	3.07	3.45	3.74	3.93	4.00	3.78
	Uformer*	2.50	2.73	2.91	3.02	**2.79**	2.93	3.10	3.25	3.34	**3.16**	3.73	3.91	4.01	4.05	3.92
	DCCRN*	2.28	2.60	2.84	2.97	2.67	2.69	2.97	3.17	3.29	3.03	3.68	3.88	4.00	4.04	3.90
	DCCRN*(SNR)	2.32	2.62	2.87	2.99	2.70	2.68	2.97	3.20	3.31	3.04	3.77	3.91	4.02	4.04	**3.93**
	CTSNet	2.62	2.85	3.00	3.06	2.88	3.00	3.19	3.31	3.37	**3.22**	3.84	3.99	4.06	4.08	**3.99**
	G2Net	2.55	2.83	3.00	3.07	2.86	2.97	3.18	3.32	3.38	3.21	3.75	3.95	4.05	4.07	3.95
	TaylorSENet	2.43	2.73	3.94	3.02	**3.03**	2.88	3.10	3.27	3.34	3.15	3.64	3.89	4.01	4.05	3.90
Babble	noisy	1.21	1.50	1.83	2.23	1.69	1.42	2.03	2.50	2.90	2.21	1.26	1.60	2.04	2.72	1.91
	MMSE-LSA	1.48	1.77	2.10	2.41	1.94	1.97	2.29	2.66	2.96	2.47	2.17	2.77	3.17	3.45	2.89
	MMSE-STSA (1)	1.48	1.76	2.10	2.41	1.94	1.95	2.28	2.66	2.97	2.47	2.15	2.76	3.16	3.44	2.88
	MMSE-STSA (0.5)	1.50	1.78	2.09	2.41	**1.95**	2.00	2.31	2.66	2.96	**2.48**	2.21	2.79	3.17	3.45	**2.91**
	MSS	1.34	1.62	2.01	2.36	1.83	1.68	2.14	2.59	2.94	2.34	1.72	2.43	2.98	3.36	2.62
	PSS	1.39	1.66	2.02	2.37	1.86	1.72	2.12	2.60	2.97	2.35	1.95	2.66	2.99	3.29	2.72
	SQ-MSS	1.30	1.59	1.97	2.29	1.79	1.65	2.16	2.55	2.83	2.30	1.56	2.24	2.97	3.42	2.55
	DeepXi-LSA*	1.76	2.14	2.49	2.77	**2.29**	2.06	2.51	2.87	3.12	2.64	3.49	3.68	3.82	3.95	**3.73**
	DeepXi-STSA*	1.77	2.15	2.49	2.75	**2.29**	2.09	2.54	2.88	2.12	**2.66**	3.30	3.61	3.78	3.91	3.65

(continued)

Table 5. Continued.

		DNS-OVL					DNS-SIG					DNS-BAK				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
	LSTM	2.15	2.40	2.67	2.86	2.52	2.63	2.84	3.04	3.19	2.93	3.35	3.61	3.88	4.00	3.71
	FullSubNet*	2.13	2.41	2.69	2.88	**2.53**	2.75	2.95	3.12	3.24	**3.01**	3.02	3.43	3.77	3.95	3.54
	CRN	2.12	2.43	2.69	2.87	**2.53**	2.57	2.83	3.05	3.20	2.91	3.45	3.73	3.92	4.02	**3.78**
	GCRN	2.09	2.45	2.71	2.87	2.53	2.53	2.84	3.06	3.20	2.91	3.47	3.75	3.94	4.01	3.79
	DPCRN	2.07	2.45	2.74	2.93	2.55	2.57	2.88	3.09	3.25	2.95	3.27	3.67	3.92	4.02	3.72
	Uformer*	2.37	2.65	2.86	2.99	**2.72**	2.81	3.03	3.20	3.32	**3.09**	3.64	3.87	3.99	4.04	**3.89**
	DCCRN*	2.12	3.48	2.78	2.93	2.58	2.55	2.88	3.12	3.26	2.95	3.50	3.78	3.96	4.03	3.82

DNS = deep noise suppression; MOS = deep noise suppression; MMSE = minimum mean-square error; LSA = log-spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction.

Discussion

Based on the metrics for simulated normal-hearing listeners, most deep-learning methods and hybrid methods gave better speech quality than traditional methods, even when the SNR was relatively high. Deep-learning methods gave markedly better performance with compressed input features than with uncompressed input features. The decoupling-style deep-learning methods performed best among the three groups of deep-learning methods, indicating that it is important to optimize the magnitude and phase separately. With decoupling, interaction effects between the magnitude and the phase can be avoided, thus improving estimation accuracy for both magnitude and phase. It might be expected that decoupling-style deep-learning methods would require more parameters and more computational resources than single-stage methods. Fortunately, this is not the case. To demonstrate this, Table 14 presents the model size and number of multiply-accumulate operations (MACs) in GMAC per second for each deep-learning method and the hybrid methods (Because DeepXi-LSA and DeepXi-STSA use the same network architecture, only a single model size and computational complexity are shown for this csae). For traditional methods, the model size and number of MACs are less than 5 K and 10 million MACs, respectively. It is clear that better performance does not necessarily require greater storage and computation. For example, the hybrid method has the second smallest model size, but it yielded good performance for relatively high SNRs.

Based on the metrics for the simulated hearing-impaired listeners, there was no benefit of using compressed input features; most of the deep-learning methods gave higher HASQI scores using the uncompressed features than using the compressed features. When the SNR was relatively low, e.g., $\leq$ 10 dB, the HASQI improvement relative to unprocessed speech was significant, while the improvement was minor when the SNR was high. However, hearing-impaired listeners have special difficulties at low SNRs (Moore, 2007), so effective noise reduction at low SNRs is likely to be beneficial. Both the traditional and the deep-learning methods led to positive effects on speech quality at low SNRs. Better performance was achieved using deep-learning methods and a benefit from deep-learning methods occurred over a wider range of SNRs. Hybrid methods achieved promising performance, and they even surpassed deep-learning-only methods in terms of several objective metrics, indicating that traditional speech enhancement can be improved by using deep-learning methods to estimate some key parameters.

Conclusions and Future Prospects

This paper has reviewed many representative frequency-domain monaural speech enhancement methods proposed over the last six decades. Traditional methods are often not data-driven and often rely on specific statistical models of the speech and noise and/or a deterministic model of the speech. Early researchers in this area often assumed a model, derived a method based on this model, and then evaluated the proposed method by simulation or experiment. Traditional methods usually did not estimate the complex spectrum of the clean speech in a single stage; even for the simplest traditional method, the noise PSD had to be estimated in an initial stage. In contrast, deep-learning methods are usually data-driven and their performance depends on the training dataset, the input features provided, the learning target, and the deep-learning network architecture. Deep learning methods do not require the assumption of a statistical signal model. Note that a given deep-learning architecture may yield significantly different performance when trained using different datasets and schemes. A disadvantage of deep-learning methods is their "black box" nature. It is hard to understand in detail how a DNN achieves its results, and it can be hard to interpret performance changes produced by changing the architecture of the DNN. From a theoretical viewpoint, it is important to open the black boxes to allow understanding of the mechanisms underlying better performance. This may in turn lead to better methods.

Table 6. As Table 3 but using DNS-MOS as the evaluation metric.

		DNS-OVL					DNS-SIG					DNS-BAK				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	1.24	1.69	2.23	2.57	1.93	1.53	2.40	3.04	3.28	2.56	1.27	1.77	2.54	3.13	2.18
	LSTM	2.19	2.51	2.76	2.91	2.59	2.53	2.86	3.09	3.24	2.93	3.76	3.88	3.98	4.03	3.91
	FullSubNet*	2.19	2.53	2.79	295	**2.62**	2.55	2.88	3.12	3.27	**2.95**	3.78	3.91	4.01	4.05	**3.94**
	CRN	2.18	2.52	2.78	2.94	2.60	2.54	2.88	3.12	3.26	**2.95**	3.72	3.87	3.99	4.04	3.90
	GCRN	1.96	2.51	2.83	2.98	2.57	2.22	2.83	3.14	3.29	2.87	3.84	3.97	4.04	4.08	3.98
	DPCRN	2.11	2.50	2.79	2.26	2.42	2.42	2.83	3.10	3.27	2.90	3.81	3.94	4.04	4.04	3.96
	Uformer*	2.45	2.70	2.90	3.02	2.77	2.78	3.03	3.22	3.22	3.06	3.96	4.00	4.05	4.08	4.02
	DCCRN*	2.46	2.72	2.91	3.03	**2.78**	2.79	3.04	3.22	3.32	**3.09**	3.97	4.03	4.07	4.10	**4.04**
	CTSNet	2.51	2.81	2.97	3.06	2.84	2.84	3.13	3.29	3.36	3.15	3.95	4.02	4.06	4.09	4.03
	G2Net	2.47	2.79	3.97	3.05	2.82	2.80	3.11	3.28	3.35	3.14	3.93	4.02	4.06	4.09	4.02
	TaylorSENet	2.53	2.82	3.00	3.07	**2.86**	2.86	3.14	3.31	3.37	**3.17**	3.96	4.04	4.08	4.10	**4.05**
Cafe	noisy	1.23	1.55	2.09	2.47	1.84	1.47	2.09	2.80	3.17	2.38	1.26	1.63	2.37	2.99	2.06
	LSTM	2.30	2.58	2.83	2.95	2.67	2.68	2.95	3.17	3.28	3.02	3.71	3.86	3.98	4.02	3.89
	FullSubNet*	2.26	2.59	2.86	3.00	2.68	2.68	2.96	3.19	3.31	3.04	3.67	3.88	4.01	4.07	**3.91**
	CRN	2.33	2.61	2.85	2.98	**2.69**	2.73	2.99	3.20	3.31	**3.06**	3.68	3.86	3.98	4.04	3.89
	GCRN	2.16	2.59	2.88	3.02	2.66	2.45	2.91	3.19	3.32	2.97	3.86	3.97	4.05	4.09	3.99
	DPCRN	2.26	2.62	2.88	3.04	2.70	2.61	2.96	3.20	3.34	3.03	3.78	3.94	4.04	4.10	3.96
	Uformer*	2.57	2.81	2.97	3.06	**2.85**	2.92	3.13	3.28	3.37	**3.18**	3.94	4.01	4.06	4.09	4.02
	DCCRN*	2.56	2.80	2.97	3.07	**2.85**	2.89	3.12	3.27	3.36	3.16	3.97	4.04	4.09	4.11	**4.05**
	CTSNet	2.62	2.88	3.03	3.10	2.91	2.96	3.19	3.344	3.39	3.22	3.95	4.04	4.09	4.10	4.04
	G2Net	2.62	2.84	3.01	3.08	2.89	2.96	3.16	3.31	3.38	3.20	3.96	4.03	4.08	4.09	4.04
	TaylorSENet	2.67	2.90	3.03	3.09	**2.92**	3.00	3.21	3.33	3.39	**3.23**	3.98	4.06	4.10	4.11	**4.06**
Babble	noisy	1.21	1.50	1.83	2.23	1.69	1.42	2.03	2.50	2.90	2.21	1.26	1.60	2.04	2.72	1.91
	LSTM	2.18	2.49	2.74	2.90	2.58	2.56	2.86	3.08	3.22	2.93	3.61	3.79	3.94	4.02	3.84
	FullSubNet*	2.13	2.49	2.80	2.95	2.59	2.56	2.88	3.14	3.27	**2.96**	3.53	3.77	3.96	4.03	3.82
	CRN	2.15	2.52	2.78	2.98	**2.61**	2.49	2.86	3.10	3.24	2.93	3.71	3.88	4.00	4.05	**3.91**
	GCRN	2.01	2.54	2.84	2.99	2.60	2.30	2.87	3.16	3.29	2.90	3.78	3.94	4.03	4.08	3.96
	DPCRN	2.06	2.48	2.79	2.98	2.58	2.37	2.81	3.10	3.28	2.89	3.73	3.90	4.03	4.10	3.94
	Uformer*	2.46	2.73	2.93	3.03	**2.79**	2.80	3.07	3.24	3.34	**3.11**	3.91	3.99	4.04	4.08	4.00
	DCCRN*	2.42	2.74	2.95	3.05	**2.79**	2.75	3.06	3.25	3.35	3.10	3.92	4.01	4.08	4.10	**4.03**
	CTSNet	2.54	2.87	3.01	3.07	2.87	2.88	3.19	3.31	3.37	3.19	3.91	4.03	4.07	408	**4.02**
	G2Net	2.49	2.81	2.99	3.06	2.83	2.85	3.14	3.30	3.36	3.16	3.86	4.00	4.05	4.07	4.00
	TaylorSENet	2.56	2.88	3.02	3.08	**2.88**	2.90	3.19	3.32	3.38	**3.20**	3.90	4.03	4.07	4.09	**4.02**

DNS = deep noise suppression; MOS = deep noise suppression; LSTM = long short-term memory; CRN = convolutional recurrent network.

As demonstrated in the evaluations presented in this paper, deep-learning speech enhancement methods[10] can yield much better performance than traditional methods. However their storage and computational requirements are much higher than for traditional methods, which limits their application in low-resource devices, such as hearing aids and cochlear implants. Moreover, a majority of deep-learning speech enhancement methods are noncausal and their algorithmic delay covers a wide range; some of them are even sentence based. Although many researchers have developed frame-wise deep-learning speech enhancement methods, the algorithmic delay is still larger than 30 ms in most cases. This delay is small enough for some applications, such as audio-visual conferences, smart home devices, and smartphones, but it is too long for hearing-assistance devices (Stone & Moore, 1999; Stone et al., 2008) and on-site sound reinforcement systems. For hearing aids it is generally believed that the delay should

Table 7. As Table 1 but using HASQI (%) as the evaluation metric. The HASQI scores for simulated normal-hearing and hearing-impaired listeners are included for completeness.

		Normal					Mild					Moderate				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	5.48	12.14	21.99	33.66	18.32	10.17	23.12	40.41	58.56	33.07	20.81	44.20	67.73	84.66	54.35
	MMSE-LSA	8.46	16.11	24.98	35.36	**21.23**	17.27	31.48	45.07	58.80	38.16	28.62	44.19	58.31	73.35	51.12
	MMSE-STSA (1)	8.53	16.16	25.03	35.40	21.28	17.35	31.66	45.41	59.25	**38.42**	28.94	44.82	59.09	74.09	51.74
	MMSE-STSA (0.5)	8.37	16.01	24.85	35.14	21.09	17.14	31.16	44.52	58.09	37.73	28.14	43.33	57.31	72.51	50.32
	MSS	7.73	15.12	23.95	34.29	20.27	16.16	30.86	45.23	59.24	37.87	29.48	47.10	61.26	74.57	53.10
	PSS	7.98	15.33	23.99	34.12	20.36	16.20	30.76	45.27	59.78	38.00	29.14	47.04	62.11	76.74	**53.76**
	SQ-MSS	7.17	14.55	23.49	33.91	19.78	15.89	30.24	42.95	54.53	35.90	27.92	43.19	54.07	64.38	47.39
	DeepXi-LSA*	21.49	32.39	41.27	49.24	**36.10**	30.82	47.58	60.40	71.55	52.59	36.27	58.92	75.48	86.85	64.38
	DeepXi-STSA*	21.61	32.40	41.12	49.03	36.04	31.72	48.66	61.35	72.32	**53.51**	38.10	60.60	76.55	87.42	**65.67**
	LSTM	20.94	30.36	37.59	44.13	33.26	34.32	49.65	60.68	69.50	53.54	43.30	62.97	75.87	84.44	66.65
	FullSubNet*	21.05	31.02	39.34	46.51	**34.48**	36.12	52.02	63.38	72.51	**56.01**	48.97	68.35	81.12	89.57	**72.00**
	CRN	20.29	29.65	36.94	43.60	32.62	34.50	50.54	62.99	73.07	55.28	42.81	63.54	77.63	87.13	67.78
	GCRN	23.30	33.47	41.11	47.44	36.33	41.06	56.59	66.71	74.93	59.82	51.00	71.42	82.80	89.85	73.77
	DPCRN	21.52	32.03	40.31	47.61	35.37	35.67	52.78	64.32	73.83	56.65	44.78	66.52	79.79	88.93	70.01
	Uformer*	27.11	36.95	44.28	50.39	**39.68**	43.03	56.91	66.76	75.18	**60.47**	53.89	70.76	82.03	89.99	**74.17**
	DCCRN*	24.43	34.86	42.73	49.42	37.86	37.44	53.35	64.56	73.76	57.28	45.52	66.68	80.08	88.81	70.27
	DCCRN*(SNR)	18.77	27.89	35.86	43.06	31.40	34.11	50.02	61.80	72.24	54.54	46.09	68.30	81.66	90.43	71.62
	CTSNet	29.53	38.82	45.74	52.01	**41.53**	46.33	60.80	70.41	77.86	**63.85**	58.27	76.44	86.50	92.43	**78.39**
	G2Net	26.30	35.47	42.04	47.79	37.90	44.08	58.94	68.76	76.82	62.15	57.06	75.94	86.31	92.45	77.94
	TaylorSENet	25.82	35.71	43.19	49.30	38.51	43.74	59.22	69.22	76.99	62.29	56.55	75.25	85.73	92.04	71.62
Cafe	noisy	7.41	14.48	23.67	34.69	20.06	12.84	25.47	41.98	59.30	34.90	22.25	44.43	67.73	84.59	54.75
	MMSE-LSA	8.46	16.11	24.98	35.36	**21.23**	15.42	29.26	43.79	57.30	36.44	23.97	41.68	59.45	75.05	50.04
	MMSE-STSA (1)	8.99	16.05	24.05	33.46	20.64	15.49	29.37	43.95	57.56	36.59	24.23	42.12	60.01	75.57	50.48
	MMSE-STSA (0.5)	8.98	16.03	23.96	33.30	20.57	15.28	29.02	43.36	56.69	36.09	23.55	40.93	58.63	74.36	49.37
	MSS	7.73	15.12	23.95	34.29	20.27	15.32	29.24	44.39	58.22	**36.79**	25.23	44.25	61.90	75.94	51.83
	PSS	7.98	15.33	23.99	34.12	20.36	15.17	28.83	43.83	58.03	36.47	24.72	43.75	62.36	77.66	**52.12**
	SQ-MSS	7.17	14.55	23.49	33.91	19.78	15.10	28.86	42.83	54.63	35.36	24.21	41.09	55.35	66.70	46.84
	DeepXi-LSA*	22.71	34.17	43.20	51.46	**37.89**	33.56	50.51	63.05	73.34	55.12	38.87	61.10	76.64	87.27	65.97
	DeepXi-STSA*	22.70	34.15	43.13	51.44	37.86	34.21	51.40	63.88	74.01	**55.88**	40.17	62.46	77.57	87.80	**67.00**
	LSTM	22.31	32.50	39.74	46.11	35.17	36.52	52.34	63.25	71.45	55.89	44.53	64.26	77.31	85.32	67.86
	FullSubNet*	21.12	31.81	40.37	48.45	**35.44**	36.66	53.63	65.63	74.89	57.70	48.83	69.42	82.42	90.37	**72.76**
	CRN	21.43	31.06	38.57	46.11	34.29	36.71	53.46	65.62	75.20	**57.75**	43.24	64.22	78.47	87.78	68.43

(continued)

Table 7. Continued.

Noise type	Model	Normal					Mild					Moderate				
		−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
	GCRN	26.11	36.34	43.38	49.77	38.90	44.35	59.55	69.24	76.60	62.44	53.49	72.24	83.33	89.68	74.69
	DPCRN	22.53	33.16	41.49	49.30	36.62	37.63	55.29	66.78	75.67	58.84	44.93	67.25	80.84	89.25	70.57
	Uformer*	27.65	38.36	45.77	52.22	**41.00**	45.14	59.97	69.63	77.58	**63.08**	53.92	72.47	83.56	90.77	**75.18**
	DCCRN*	25.82	36.84	44.54	51.70	39.73	39.79	56.44	67.26	76.02	59.88	46.39	67.99	81.08	89.30	71.19
	DCCRN*(SNR)	21.27	30.21	37.05	44.05	33.15	38.05	54.02	65.24	75.10	58.10	49.38	70.17	83.01	90.92	73.37
	CTSNet	32.32	41.63	48.43	54.63	**44.25**	50.38	64.16	73.05	80.02	**66.90**	61.34	77.92	87.07	92.80	79.78
	G2Net	29.79	38.54	44.49	50.31	40.78	49.20	63.02	71.92	78.99	65.78	61.07	77.95	87.41	92.81	**79.81**
	TaylorSENet	29.11	38.58	45.61	51.41	41.18	48.48	62.85	72.34	79.38	65.69	60.45	77.07	86.99	92.58	79.27
Babble	noisy	8.21	15.02	24.60	35.43	20.82	11.29	22.68	38.79	55.81	32.14	17.16	38.58	63.11	81.71	50.14
	MMSE-LSA	9.76	15.85	23.67	32.15	**20.36**	14.10	26.72	41.63	55.70	34.54	22.16	39.19	55.40	70.86	46.90
	MMSE-STSA (1)	9.70	15.80	23.63	32.10	20.31	14.06	26.71	41.71	55.90	**34.60**	22.24	39.55	55.98	71.49	47.32
	MMSE-STSA (0.5)	9.77	15.80	23.59	32.02	20.30	14.04	26.56	41.26	55.15	34.25	21.86	38.47	54.48	70.02	46.21
	MSS	9.20	15.35	23.11	31.90	19.89	13.49	26.23	41.50	56.16	34.35	22.28	41.12	58.13	72.39	**48.48**
	PSS	8.90	14.95	22.65	31.20	19.43	13.17	25.68	40.86	55.83	33.89	21.46	40.18	58.08	73.78	48.38
	SQ-MSS	9.55	15.65	12.22	32.20	20.18	13.69	26.37	40.60	53.25	33.48	22.16	39.12	52.73	63.66	44.42
	DeepXi-LSA*	19.73	31.19	41.74	50.86	35.88	28.27	45.98	60.41	71.56	51.56	30.28	54.33	72.29	84.63	60.38
	DeepXi-STSA*	19.77	31.28	41.81	50.92	**35.95**	28.68	46.63	61.10	72.13	**52.14**	31.39	55.66	73.41	85.29	**61.44**
	LSTM	17.95	28.67	38.03	44.86	32.38	30.40	48.30	61.13	69.84	52.42	35.41	58.88	74.11	83.36	62.94
	FullSubNet*	18.65	30.39	40.67	49.22	**34.73**	29.65	47.88	62.02	72.35	52.98	39.15	63.13	78.49	88.27	**67.26**
	CRN	17.83	28.80	37.49	45.06	32.30	31.26	49.98	63.48	73.55	**54.57**	34.37	58.51	75.09	85.64	63.40
	GCRN	21.81	32.79	41.62	48.24	36.12	37.48	55.67	67.46	75.11	58.93	43.81	67.66	80.89	88.29	70.16
	DPCRN	18.26	30.24	40.53	48.72	34.44	30.82	51.10	64.61	73.99	55.13	35.29	61.17	77.31	87.18	65.24
	Uformer*	22.52	35.42	44.94	52.07	**38.74**	38.09	56.38	67.55	75.74	**59.44**	43.91	66.88	80.26	88.57	**69.91**
	DCCRN*	21.37	33.31	42.89	50.54	37.03	33.59	52.14	64.81	73.99	56.13	37.58	61.88	77.17	86.91	65.89
	DCCRN*(SNR)	16.76	27.21	35.03	42.30	30.33	32.65	50.54	63.28	72.98	54.86	39.99	65.21	80.09	89.23	68.63
	CTSNet	28.49	40.40	48.67	55.28	**43.21**	43.88	61.39	71.64	78.69	**63.90**	51.56	74.16	85.36	91.28	75.59
	G2Net	25.55	35.93	44.00	50.51	39.00	43.04	59.91	70.19	77.48	62.66	52.43	74.00	85.03	91.48	**75.74**
	TaylorSENet	25.11	36.33	44.79	51.10	39.33	41.84	59.29	70.36	78.05	62.39	51.01	65.21	80.09	89.23	68.63

HASQI = hearing-aid speech quality index; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction; SQ-MSS = square root of magnitude spectral subtraction; CRN = convolutional recurrent network; SNR = speech-to-noise ratio.

Table 8. As Table 3 but using HASQI (%) as the evaluation metric.

		Normal					Mild					Moderate				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	5.48	12.14	21.99	33.66	18.32	10.17	23.12	40.41	58.56	33.07	20.81	44.20	67.73	84.66	54.35
	LSTM	22.70	33.51	41.95	49.44	36.90	32.92	49.07	60.72	70.31	53.26	39.86	60.98	74.41	83.72	64.74
	FullSubNet*	22.46	32.95	41.24	47.97	36.16	33.25	49.01	61.21	71.49	53.74	43.93	65.68	79.90	89.01	**69.63**
	CRN	23.35	33.67	42.28	49.97	**37.32**	35.22	51.08	63.30	73.11	**55.68**	41.41	63.04	77.77	87.30	67.38
	GCRN	22.49	34.80	42.90	49.09	37.32	29.59	48.46	61.83	71.75	52.91	34.80	61.90	78.55	87.54	65.70
	DPCRN	23.65	34.10	42.16	48.33	37.06	34.39	50.41	62.00	71.91	54.68	41.21	64.17	78.46	87.90	67.94
	Uformer*	31.90	41.06	47.58	52.95	**43.37**	42.54	56.81	67.06	75.71	**60.53**	52.38	71.57	83.28	90.80	**74.51**
	DCCRN*	29.01	38.50	44.65	49.74	40.48	37.99	53.60	64.34	73.39	57.33	47.06	69.43	82.37	90.25	72.28
	CTSNet	31.57	41.57	48.54	54.03	**43.93**	42.14	57.95	68.46	76.44	61.25	53.95	74.35	85.49	91.69	76.37
	G2Net	31.31	41.12	47.92	53.60	43.49	42.66	57.80	68.21	76.59	**61.32**	54.29	74.66	85.57	91.96	76.62
	TaylorSENet	31.43	41.25	48.00	53.56	43.56	42.45	57.82	68.39	76.45	61.28	54.68	74.85	86.09	92.19	**76.95**
Cafe	noisy	7.41	14.48	23.67	34.69	20.06	12.84	25.47	41.98	59.30	34.90	22.25	44.43	67.73	84.59	54.75
	LSTM	25.35	36.07	44.62	52.14	39.55	36.38	52.36	63.89	82.84	**58.87**	42.46	62.57	76.32	84.60	66.49
	FullSubNet*	23.29	33.96	42.20	49.57	37.26	35.93	52.03	64.19	72.23	56.60	45.25	66.29	80.59	89.53	**70.42**
	CRN	25.43	36.55	45.14	52.74	**39.97**	37.24	54.42	66.36	75.77	58.45	42.11	64.08	78.80	88.23	68.31
	GCRN	26.59	37.76	45.41	51.40	40.29	34.28	52.24	65.07	74.38	56.49	38.96	63.26	79.46	88.33	67.50
	DPCRN	25.08	35.52	43.22	50.06	38.47	37.70	53.98	65.19	74.59	57.87	44.01	66.09	79.63	88.87	69.65
	Uformer*	34.36	43.81	45.77	49.77	**43.43**	46.30	60.51	70.31	78.15	**63.77**	54.72	73.30	84.31	91.41	**75.94**
	DCCRN*	30.87	40.70	46.35	51.43	42.34	40.59	57.60	67.75	75.86	60.45	47.83	71.43	83.66	90.48	73.35
	CTSNet	34.33	44.64	50.93	56.45	46.59	46.56	61.89	71.24	78.71	64.60	56.82	75.91	85.92	92.06	77.68
	G2Net	35.35	44.61	50.66	56.05	46.67	48.23	62.11	71.49	78.71	65.14	58.44	76.50	86.46	92.15	78.39
	TaylorSENet	35.45	44.60	50.75	56.17	**46.74**	48.46	62.40	71.79	78.89	**65.39**	59.76	77.39	87.18	92.56	**79.22**
Babble	noisy	8.21	15.02	24.60	35.43	20.82	11.29	22.68	38.79	55.81	32.14	17.16	38.58	63.11	81.71	50.14
	LSTM	21.09	32.97	43.54	51.84	37.36	31.14	49.16	61.82	71.42	53.39	34.01	57.83	72.92	82.78	61.89
	FullSubNet*	16.83	27.59	37.95	46.87	32.31	29.48	47.65	61.21	71.93	52.57	34.32	59.92	76.41	87.12	**64.44**
	CRN	21.28	35.00	45.27	53.50	**38.76**	30.01	50.52	63.97	73.92	**54.61**	30.50	57.57	75.11	85.90	62.27
	GCRN	22.30	35.65	44.66	51.07	38.42	29.52	50.05	63.77	72.65	54.00	30.90	59.87	77.15	86.23	63.54
	DPCRN	20.72	32.72	42.96	50.16	36.64	31.45	49.93	63.52	72.86	54.44	33.67	59.19	76.62	86.82	64.04
	Uformer*	29.48	41.48	48.99	55.03	**43.75**	39.85	57.26	68.27	76.37	**60.44**	44.76	67.97	81.27	89.35	**70.84**
	DCCRN*	26.24	38.53	46.41	52.19	40.84	34.84	54.06	66.00	74.36	57.32	37.94	65.39	80.50	88.84	68.17
	CTSNet	30.81	43.43	50.72	56.52	**45.37**	41.45	59.38	69.69	76.97	61.87	48.52	72.04	83.79	90.51	73.72
	G2Net	29.62	41.57	50.02	55.84	44.26	42.38	59.02	69.80	77.97	62.29	49.35	71.82	83.94	90.59	73.93
	TaylorSENet	30.68	42.55	50.20	56.22	44.91	42.91	59.79	70.03	77.34	**62.52**	50.94	73.45	84.86	91.14	**75.10**

HASQI = hearing-aid speech quality index; LSTM = long short-term memory; CRN = convolutional recurrent network.

Table 9. As Table 1 but using HASPI (%) as the evaluation metric.

		Normal					Mild					Moderate				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	18.09	63.90	92.48	98.63	68.28	11.92	49.72	83.36	94.54	59.89	6.38	34.98	67.20	76.87	46.36
	MMSE-LSA	26.01	73.39	94.03	98.67	73.03	14.00	57.70	81.25	89.87	60.71	8.21	36.68	58.14	68.29	42.83
	MMSE-STSA (1)	26.74	73.91	94.08	98.69	**73.36**	14.28	58.03	81.52	90.15	**61.00**	8.55	37.13	58.59	68.65	43.23
	MMSE-STSA (0.5)	24.53	72.32	93.74	98.58	72.29	13.59	56.97	80.73	89.48	60.19	7.84	36.13	57.79	68.02	42.45
	MSS	25.40	72.05	93.22	98.59	72.32	13.60	55.06	81.04	90.79	60.12	8.33	37.40	60.34	67.90	43.49
	PSS	25.82	72.49	92.82	98.48	72.40	14.15	56.30	81.11	91.31	60.72	8.75	38.82	61.43	69.58	**44.65**
	SQ-MSS	23.11	70.43	93.27	98.65	71.37	11.52	52.53	78.86	87.93	57.71	7.20	34.11	56.41	62.95	40.17
	DeepXi-LSA*	76.96	97.37	99.39	99.81	**93.38**	50.33	79.64	88.93	93.10	**78.00**	28.42	55.17	67..66	73.63	56.22
	DeepXi-STSA*	76.68	97.33	99.37	99.81	93.30	52.28	81.57	90.13	93.75	79.43	31.21	58.50	70.04	74.93	**58.67**
	LSTM	76.82	95.74	98.55	99.47	92.65	57.59	82.70	88.41	91.30	80.00	41.22	62.80	69.66	72.46	61.54
	FullSubNet*	76.68	96.67	99.26	99.70	**93.08**	58.05	85.23	92.38	94.73	**82.60**	43.18	63.41	72.18	75.98	**63.69**
	CRN	76.50	95.50	98.38	99.35	92.43	61.96	84.47	90.10	92.08	82.15	41.89	63.94	70.80	73.28	62.48
	GCRN	81.42	96.49	98.65	99.47	94.01	61.99	85.10	90.69	92.93	82.68	40.08	64.62	72.03	74.32	**62.76**
	DPCRN	79.39	96.65	98.79	99.59	93.61	61.97	84.66	89.49	91.83	81.99	40.86	62.72	70.22	73.37	61.79
	Uformer*	89.18	98.29	99.40	99.73	**96.65**	73.95	87.51	90.69	92.46	**86.15**	53.39	65.28	71.39	74.04	**66.03**
	DCCRN*	86.09	97.82	99.39	99.76	95.77	63.54	85.63	91.64	93.55	83.59	38.65	61.34	70.15	73.81	60.99
	DCCRN*(SNR)	69.02	93.00	98.00	99.45	89.87	55.88	81.99	89.72	92.69	80.07	38.68	60.62	69.87	73.82	60.75
	CTSNet	90.76	97.67	98.99	99.62	**96.76**	70.03	86.77	91.26	93.45	85.38	49.45	68.01	75.24	74.99	**66.92**
	G2Net	89.21	97.54	98.81	99.35	96.23	73.07	87.46	90.65	92.76	**85.99**	50.49	68.27	73.48	75.19	66.86
	TaylorSENet	87.89	97.63	99.07	99.65	96.06	70.67	87.68	91.37	93.28	85.75	49.43	68.50	73.67	75.33	66.73
Cafe	noisy	31.74	72.72	93.88	98.52	74.22	14.01	45.74	78.37	92.01	57.53	6.53	29.99	61.88	74.74	43.29
	MMSE-LSA	38.61	76.04	93.35	98.15	**76.54**	12.82	46.83	73.25	86.64	54.89	6.22	25.42	50.25	65.36	36.81
	MMSE-STSA (1)	38.67	75.89	93.28	98.15	76.50	13.03	46.88	73.40	86.84	55.04	6.39	25.77	50.56	65.69	37.10
	MMSE-STSA (0.5)	37.85	75.59	93.16	98.06	76.17	12.45	46.36	72.67	86.21	54.42	5.96	25.00	50.02	65.10	36.52
	MSS	38.59	73.87	92.90	97.99	75.84	13.30	46.12	74.58	87.78	**55.45**	6.50	28.08	52.37	65.16	38.03
	PSS	37.85	73.51	92.06	97.58	75.25	13.53	45.99	73.47	87.59	55.15	6.60	27.94	52.27	66.37	**38.30**
	SQ-MSS	38.72	74.92	93.71	98.33	76.42	11.91	44.89	73.98	85.94	54.18	5.75	26.66	49.79	61.06	35.82
	DeepXi-LSA*	85.18	97.97	99.54	99.86	**95.64**	49.27	80.02	89.87	93.96	78.28	27.62	556.87	68.16	74.20	56.71
	DeepXi-STSA*	84.97	97.93	99.54	99.86	95.58	51.02	81.43	90.85	94.45	**79.44**	29.76	59.22	69.94	75.36	**58.57**
	LSTM	85.30	97.19	99.18	99.60	**95.32**	60.02	83.64	89.94	92.12	81.43	40.80	62.66	69.64	73.14	61.56
	FullSubNet*	84.37	97.35	99.18	99.76	95.17	54.93	83.34	92.92	95.56	81.69	37.17	63.08	72.54	76.55	**62.34**
	CRN	83.47	96.44	98.97	99.57	94.61	63.74	84.52	90.79	92.93	**83.00**	42.45	62.48	70.17	73.39	62.12

(continued)

Table 9. Continued.

		Normal					Mild					Moderate				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
	GCRN	87.81	97.37	99.20	99.70	96.02	66.36	86.55	91.66	93.46	84.51	43.93	65.91	72.04	74.32	61.05
	DPCRN	85.37	97.34	99.26	99.74	95.43	62.55	84.21	90.37	92.66	82.45	40.77	62.87	70.18	73.66	61.87
	Uformer*	93.61	98.97	99.62	99.84	**98.01**	75.94	88.07	91.56	93.42	**87.25**	53.74	66.32	71.97	74.31	**66.59**
	DCCRN*	91.17	98.599	99.57	99.86	97.30	65.02	86.04	92.05	94.35	84.37	40.38	63.15	70.71	74.05	62.07
	DCCRN*(SNR)	76.98	94.54	98.25	99.46	92.31	57.80	82.79	90.69	93.43	81.18	39.99	61.74	70.53	74.17	61.61
	CTSNet	94.06	98.59	99.47	99.81	97.63	75.33	89.08	92.15	94.07	87.66	54.15	69.56	73.70	75.07	68.12
	G2Net	93.36	98.35	99.12	99.68	97.63	78.28	89.47	911.92	93.45	**88.28**	56.28	69.79	73.80	75.31	**68.80**
	TaylorSENet	93.20	98.42	99.46	99.79	**97.72**	75.83	89.30	92.65	94.15	87.98	55.18	70.03	73.96	75.58	68.69
Babble	noisy	34.63	74.85	93.80	98.07	75.34	10.29	35.14	72.04	90.83	52.08	2.68	15.86	54.24	75.68	37.12
	MMSE-LSA	37.27	74.39	91.97	97.25	**75.55**	7.78	33.13	69.37	86.18	**49.12**	3.20	17.96	49.10	65.74	34.00
	MMSE-STSA (1)	36.87	73.86	91.79	97.15	74.92	7.80	33.05	69.17	86.21	49.06	3.23	18.12	49.20	65.93	34.12
	MMSE-STSA (0.5)	37.41	74.14	91.74	97.26	75.14	7.67	32.84	69.06	85.79	48.84	3.18	17.63	48.93	65.53	33.82
	MSS	33.67	71.28	91.18	96.93	73.27	7.28	31.44	68.83	87.47	48.76	2.92	18.45	51.04	66.27	**34.67**
	PSS	32.17	68.95	89.72	96.32	71.79	7.26	31.07	66.76	86.65	47.94	3.03	18.05	50.05	66.74	34.47
	SQ-MSS	36.87	74.19	92.43	97.56	75.26	6.67	31.12	69.93	86.69	48.60	2.76	18.22	49.69	62.54	33.30
	DeepXi-LSA*	78.54	97.25	99.24	99.67	**93.68**	33.95	72.67	88.97	94.08	72.42	17.26	47.56	66.30	73.59	51.18
	DeepXi-STSA*	78.28	97.20	99.23	99.66	93.59	35.19	73.92	89.70	94.50	**73.33**	18.41	49.86	68.52	74.86	**52.91**
	LSTM	75.31	96.22	98.98	99.50	**92.50**	38.42	75.89	89.03	91.90	73.81	25.48	54.28	68.47	72.40	55.16
	FullSubNet*	74.51	95.57	98.63	99.25	91.99	32.43	71.14	90.55	94.80	72.23	21.18	52.29	70.25	76.21	54.98
	CRN	72.29	95.87	98.94	99.60	91.68	47.02	81.89	90.59	93.02	**78.13**	27.81	57.82	69.38	73.20	**57.05**
	GCRN	79.67	96.36	98.97	99.56	93.64	49.72	80.54	90.97	93.32	78.64	27.61	58.08	70.97	74.19	57.71
	DPCRN	76.26	96.63	99.39	99.74	93.01	43.79	78.79	89.71	92.32	76.15	23.92	53.96	68.59	72.87	54.84
	Uformer*	87.31	98.76	99.68	99.85	**96.40**	61.43	85.91	91.63	93.42	**83.10**	37.85	61.49	70.69	74.04	**61.02**
	DCCRN*	83.29	97.80	99.38	99.70	95.04	46.60	80.80	91.21	94.27	78.22	24.87	55.29	69.01	73.59	55.69
	DCCRN*(SNR)	58.13	91.04	97.16	98.76	86.27	38.79	75.74	89.66	93.25	74.36	22.88	55.20	69.35	73.92	55.34
	CTSNet	91.43	98.52	99.64	99.87	**97.37**	63.53	87.58	92.21	93.95	**84.32**	40.17	66.57	73.33	75.29	**63.84**
	G2Net	89.19	97.37	99.25	99.54	96.34	63.59	86.40	91.61	93.43	83.76	39.61	65.12	73.00	75.33	63.27
	TaylorSENet	89.08	98.06	99.47	99.79	96.60	60.62	85.73	91.96	94.16	83.12	38.43	65.19	73.24	75.57	63.11

HASPI = hearing-aid speech perception index; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction; CRN = convolutional recurrent network; SNR = speech-to-noise ratio.

Table 10. As Table 3 but using HASPI (%) as the evaluation metric.

		Normal					Mild					Moderate				
Noise type	Model	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.	−5 dB	0 dB	5 dB	10 dB	Ave.
Factory1	noisy	18.09	63.90	92.48	98.63	68.28	11.92	49.72	83.36	94.54	59.89	6.38	34.98	67.20	76.87	46.36
	LSTM	81.22	97.55	99.13	99.68	94.40	58.71	82.96	88.85	91.70	80.56	39.50	61.99	69.53	72.84	60.97
	FullSubNet*	81.05	97.64	99.36	99.72	94.44	57.13	83.39	91.42	94.35	81.57	39.24	59.72	69.64	74.68	60.82
	CRN	82.86	97.33	99.07	99.63	**94.72**	64.95	84.31	89.09	91.46	**82.45**	39.92	62.09	69.68	72.91	**61.15**
	GCRN	82.85	96.99	98.92	99.65	94.60	49.72	80.60	89.50	92.91	78.18	24.92	57.83	70.01	73.96	56.68
	DPCRN	85.31	97.21	98.81	99.54	95.22	63.96	83.92	89.48	91.89	82.31	34.41	59.02	68.93	73.00	58.84
	Uformer*	94.49	98.83	99.38	99.72	**98.10**	76.45	87.52	91.14	93.28	**87.10**	50.00	65.37	72.23	74.76	**65.59**
	DCCRN*	93.55	98.61	99.35	99.72	97.81	68.15	85.14	90.39	93.11	84.20	42.61	62.54	71.02	74.21	62.60
	CTSNet	92.95	98.32	99.23	99.67	97.54	63.49	85.61	90.93	93.31	83.34	42.76	65.53	72.86	75.17	64.08
	G2Net	94.83	98.58	99.32	99.73	**98.12**	70.48	85.86	90.65	93.13	85.03	45.16	65.61	73.15	75.01	64.73
	TaylorSENet	94.77	98.48	99.21	99.69	98.04	70.84	86.93	90.82	93.07	**85.42**	45.07	66.89	73.14	75.18	**65.07**
Cafe	noisy	31.74	72.72	93.88	98.52	74.22	14.01	45.74	78.37	92.01	57.53	6.53	29.99	61.88	74.74	43.29
	LSTM	90.34	98.30	99.56	99.81	**97.00**	62.32	84.38	90.20	92.70	82.40	41.09	62.11	69.98	73.29	**61.62**
	FullSubNet*	87.56	97.82	99.30	99.77	96.11	55.94	83.04	92.45	95.27	81.68	37.47	60.98	70.59	75.44	61.12
	CRN	89.23	98.32	99.55	99.80	96.73	65.62	84.88	90.35	92.62	**83.37**	41.00	61.55	69.86	73.11	61.38
	GCRN	88.71	97.99	99.49	99.82	96.51	58.78	84.36	90.66	93.64	81.86	33.58	61.12	71.03	73.98	59.93
	DPCRN	89.33	97.78	99.23	99.71	96.51	65.28	85.12	90.32	93.04	83.44	37.91	61.17	69.63	73.69	60.60
	Uformer*	97.40	99.37	99.68	99.86	**99.08**	78.65	88.99	92.08	94.09	**88.45**	54.50	67.22	72.84	74.83	**67.35**
	DCCRN*	95.38	99.18	99.65	99.85	98.52	71.78	87.57	91.79	94.09	86.31	46.28	65.47	71.76	74.60	64.53
	CTSNet	95.21	99.07	99.63	99.86	98.44	72.77	88.68	91.99	94.02	86.87	49.06	68.08	73.55	75.37	66.52
	G2Net	97.10	99.23	99.70	99.87	**98.98**	78.12	88.77	92.06	93.79	88.19	53.33	68.13	73.61	75.36	67.61
	TaylorSENet	96.85	99.07	99.62	99.85	98.85	78.24	88.89	92.24	93.89	**88.32**	54.15	69.34	74.08	75.52	**68.27**
Babble	noisy	34.63	74.85	93.80	98.07	75.34	10.29	35.14	72.04	90.83	52.08	2.68	15.86	54.24	75.68	37.12
	LSTM	84.28	98.24	99.65	99.87	**95.51**	43.88	78.45	90.03	92.97	76.33	27.09	56.19	68.83	72.70	**56.20**
	FullSubNet*	79.33	96.67	99.05	99.64	93.67	38.73	77.32	91.41	94.79	75.56	23.72	53.56	68.21	74.45	54.99
	CRN	82.84	98.64	99.05	99.64	93.67	46.95	82.02	90.03	92.55	**77.89**	25.09	56.60	68.07	72.45	55.55
	GCRN	79.69	97.56	99.48	99.82	94.14	43.51	79.03	90.32	93.43	76.57	19.84	53.83	69.58	73.82	54.27
	DPCRN	81.57	97.19	99.45	99.72	94.48	48.51	80.67	90.31	92.70	78.05	21.79	52.81	68.14	72.85	53.90
	Uformer*	95.33	99.45	99.81	99.91	**98.63**	69.04	87.74	92.18	93.99	**85.74**	40.39	64.63	72.48	74.99	**63.12**
	DCCRN*	92.74	99.15	99.70	99.85	97.86	60.52	85.20	91.62	93.93	82.82	31.42	59.67	70.51	74.31	58.98
	CTSNet	93.42	99.24	99.80	99.92	98.10	63.60	87.30	91.98	93.81	84.17	36.03	64.00	72.80	75.58	62.10
	G2Net	93.60	99.03	99.79	99.91	98.08	64.25	86.21	91.81	93.72	84.00	36.22	62.98	72.56	75.16	61.73
	TaylorSENet	94.01	99.09	99.75	99.91	**98.19**	65.01	87.23	92.20	93.53	**84.49**	39.06	65.61	73.36	75.60	**63.41**

HASPI = hearing-aid speech perception index; LSTM = long short-term memory; CRN = convolutional recurrent network.

Table 11. Objective test results using the Voice Bank + DEMAND dataset when the input feature was uncompressed.

Model	WB-PESQ	NB-PESQ	ESTOI	SDR	DNS-OVL	DNS-SIG	DNS-BAK
Noisy	1.97	3.02	0.79	8.54	2.69	3.34	3.12
MMSE-LSA	2.39	3.27	**0.79**	15.32	2.79	3.29	3.48
MMSE-STSA (1)	2.38	3.27	**0.79**	15.20	2.79	3.29	3.47
MMSE-STSA (0.5)	2.36	3.26	**0.79**	**15.38**	**2.80**	**3.30**	**3.49**
MSS	2.37	3.27	**0.79**	13.99	2.71	3.24	3.37
PSS	2.28	2.92	0.74	12.32	2.68	3.23	3.32
SQ-MSS	**2.48**	**3.28**	**0.79**	13.90	2.76	3.25	3.47
DeepXi-LSA*	**2.67**	**3.41**	**0.84**	**18.54**	**3.05**	3.37	**3.95**
DeepXi-STSA*	**2.67**	3.37	**0.84**	17.73	**3.05**	**3.38**	3.92
LSTM	2.27	3.10	0.78	15.59	3.08	3.40	**3.95**
FullSubNet*	**2.58**	**3.33**	**0.83**	**19.90**	3.07	3.41	3.92
CRN	2.50	3.24	**0.83**	17.91	**3.09**	**3.43**	3.92
GCRN	2.33	3.18	0.81	18.66	**3.09**	**3.43**	3.92
DPCRN	2.49	3.29	0.83	18.73	3.06	3.39	3.92
Uformer*	**2.72**	3.45	**0.85**	**20.55**	**3.09**	3.40	**3.97**
DCCRN*	2.12	2.93	0.76	16.19	2.77	3.17	3.68
DCCRN (SNR)	**2.72**	**3.47**	**0.85**	14.89	3.08	3.41	3.95
CTSNet	2.48	3.25	0.82	19.38	**3.12**	**3.44**	**3.95**
G2Net	2.50	3.29	**0.83**	19.70	3.07	3.43	3.87
TaylorSENet	**2.52**	**3.38**	**0.83**	**19.76**	3.08	**3.44**	3.88

Best scores are highlighted in bold. SDR = signal-to-distortion ratio; DNS = deep noise suppression; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; MSS = magnitude spectral subtraction; PSS = power spectral subtraction; CRN = convolutional recurrent network.

Table 12. As Table 11 but for the compressed input feature.

Model	WB-PESQ	NB-PESQ	ESTOI	SDR	DNS-OVL	DNS-SIG	DNS-BAK
Noisy	1.97	3.02	0.79	8.54	2.69	3.34	3.12
LSTM	**2.75**	3.43	0.84	17.29	**3.12**	**3.43**	3.98
FullSubNet*	**2.75**	**3.51**	**0.85**	**20.62**	**3.12**	3.42	**4.01**
CRN	2.70	3.42	0.84	18.42	3.11	**3.43**	3.96
GCRN	2.68	3.42	0.84	19.17	3.12	**3.42**	3.98
DPCRN	2.79	3.46	0.85	19.46	3.11	3.40	4.01
Uformer*	**3.07**	**3.64**	**0.87**	**21.18**	**3.13**	3.41	**4.03**
DCCRN*	2.13	2.94	0.77	16.12	2.81	3.18	3.78
CTSNet	2.92	3.58	**0.86**	20.07	3.12	3.41	4.01
G2Net	2.94	3.58	**0.86**	19.33	3.14	**3.45**	3.99
TaylorSENet	**3.02**	**3.62**	**0.86**	**20.22**	**3.17**	**3.45**	**4.04**

SDR = signal-to-distortion ratio; DNS = deep noise suppression; LSTM = long short-term memory; CRN = convolutional recurrent network.

be as small as possible, and some hearing aids have delays as small as 0.5 ms. It is still challenging to reduce the delay to below 4 ms without affecting performance, although some researchers have tried to solve this problem (Vary, 2006; Schröter et al., 2022; Zheng et al., 2022). Tammen & Doclo (2021) proposed a deep multiframe approach to reduce the delay for hearing aids, and this approach was further extended for binaural noise reduction (Tammen & Doclo, 2022). From the application perspective, future work should concentrate on reducing the complexity, storage, and latency of deep-learning methods to facilitate their application in hearing aids and cochlear implants.

Table 13. Values of the HASQI (%)/HASPI (%) for the different methods using the Voice Bank + DEMAND dataset.

	Uncompressed			Compressed		
Model	Normal	Mild	Moderate	Normal	Mild	Moderate
Noisy	49.54/97.06	80.50/83.10	92.74/64.70	49.54/97.06	80.50/83.10	92.74/64.70
MMSE-LSA	46.37/96.18	76.68/78.41	89.41/61.95	-	-	-
MMSE-STSA (1)	46.37/96.20	76.89/78.55	89.63/62.13	-	-	-
MMSE-STSA (0.5)	45.95/96.09	76.15/78.26	89.06/61.79	-	-	-
MSS	**46.86**/96.22	77.02/78.11	89.03/61.15	-	-	-
PSS	46.00/95.79	**78.02/78.91**	**90.48/62.21**	-	-	-
SQ-MSS	46.15/**96.65**	70.38/75.10	81.37/57.51	-	-	-
DeepXi-LSA*	**57.15/99.06**	84.04/84.44	**94.26**/65.09	-	-	-
DeepXi-STSA*	57.04/99.02	**84.15/84.80**	**94.26/65.51**	-	-	-
LSTM	43.10/97.95	72.94/80.06	87.12/63.16	55.95/98.68	80.88/82.99	89.88/64.93
FullSubNet*	50.86/98.14	82.95/**83.53**	**94.87/65.20**	54.32/**98.92**	84.11/**84.79**	**94.87/65.19**
CRN	**54.22/98.53**	**83.49**/83.12	93.38/64.96	**59.31**/98.80	**84.32**/83.44	93.44/64.42
GCRN	48.60/97.71	81.96/82.31	93.71/64.89	55.37/98.78	83.91/84.06	93.91/64.99
DPCRN	52.46/98.37	82.39/83.22	93.80/65.22	57.36/98.76	83.67/83.62	93.49/64.64
Uformer*	**57.65/99.00**	**85.51**/84.64	**94.89/65.58**	**60.40/99.28**	**86.37/85.07**	**95.18/65.35**
DCCRN*	45.63/89.17	73.12/69.50	83.68/55.96	47.21/89.29	73.31/69.51	83.01/55.42
DCCRN (SNR)	52.94/98.69	82.98/**84.94**	91.40/65.50	-	-	-
CTSNet	50.46/98.11	82.97/**85.28**	94.35/**65.86**	55.91/99.04	86.03/84.21	95.56/65.38
G2Net	50.60/**98.23**	85.12/84.67	95.12/65.72	60.04/**99.10**	**86.89/85.12**	95.56/**65.71**
TaylorSENet	**52.81**/98.19	**86.09**/84.61	**95.42**/65.67	**60.07**/98.97	86.49/84.85	**95.58**/65.56

For all deep-learning methods, both the uncompressed spectrum and the compressed spectrum were used. Bold font indicates the best average score in each group. HASQI = hearing-aid speech quality index; HASPI = hearing-aid speech perception index; MMSE = minimum mean-square error; LSA = log-spectral amplitude; STSA = short-time spectral amplitude; LSTM = long short-term memory; CRN = convolutional recurrent network; SNR = speech-to-noise ratio.

Table 14. Comparisons in terms of model size (in million) and MACs (in GMAC per second).

Model	Model size (M)	MACs (G/s)
DeepXi	1.95	0.12
LSTM	21.82	2.19
FullSubNet	5.64	29.83
CRN	17.58	2.57
GCRN	9.77	2.42
DPCRN	0.72	0.77
Uformer	3.34	5.29
DCCRN	3.67	11.13
CTSNet	4.35	5.57
G2Net	7.39	2.83
TaylorSENet	5.45	6.43

LSTM = long short-term memory; CRN = convolutional recurrent network.

For multitalker scenarios, all deep-learning methods show reduced performance. Although training deep-learning models with multitalker speech signals can reduce the performance degradation, this problem is far from being solved.

Finally, subjective listening tests using both normal-hearing and hearing-impaired listeners are needed to fully evaluate the benefits of deep-learning methods for speech intelligibility and speech quality, and to evaluate the disturbing effects of the remaining background noise.

Acknowledgments

We thank three reviewers and the associate editor Torsten Dau for very helpful and insightful comments on an earlier version of this paper.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key Research & Development (R&D) Program of China (2021YFB3201702).

ORCID iDs

Chengshi Zheng https://orcid.org/0000-0001-5656-994X
Brian C. J. Moore https://orcid.org/0000-0001-7071-0671

Notes

1. Loizou & Kim (2011) divided the distortions relative to clean speech into three regions: In the first region, the magnitude of the estimated clean speech is smaller than that of the true clean speech, while in the second and third regions, the reverse is true. The first region corresponds to speech-attenuation distortion while the other two regions correspond to speech-amplification distortion. This speech-amplification distortion happens when the estimated spectral gain is larger than the ideal spectral gain.
2. In statistical signal processing, an MMSE estimator is an estimation method that minimizes the mean square error of a specific physical quantity between the true and estimated values.
3. The log kurtosis ratio is defined as the ratio of the kurtosis of the enhanced speech and that of the noisy speech, on a logarithmic scale (Inoue et al., 2010a). The kurtosis was introduced because it reflects the percentage of tonal components among all components. A larger kurtosis means more "musical noise" components.
4. This conclusion about the unstructured nature of speech phase only applies when the Hamming/Hanning window is used as the analysis window before performing the STFT analysis. Paliwal et al. (2011) have discussed the impact of the shape and duration of the analysis window on the short-time phase spectrum and showed that speaker-dependent information is contained in the phase spectrum when the analysis window is replaced by a Chebyshev window. Moreover, the phase difference between adjacent frames of voiced speech segments is not random, and thus speech phase can be recovered frame by frame with proper initialization (Gerkmann et al., 2015).
5. Mohammadiha et al. (2013) defined the speech and noise basis matrices as $\mathbf{b}^{(s)}$ and $\mathbf{b}^{(n)}$, respectively. These can be obtained respectively by decomposing the speech and noise matrices $\mathbf{s}$ with $\mathbf{s}_{k,l} = |S(k, l)|$ and $\mathbf{n}$ with $\mathbf{n}_{k,l} = |N(k, l)|$ into a linear combination of their corresponding basis matrices, i.e., $\mathbf{s} = \mathbf{b}^{(\mathbf{s})}\mathbf{v}^{(\mathbf{s})}$ and $\mathbf{n} = \mathbf{b}^{(n)}\mathbf{v}^{(n)}$, with $\mathbf{v}^{(s)}$ and $\mathbf{v}^{(n)}$ referred to as the speech and noise NMF coefficients, respectively.
6. Many speech enhancement approaches only handle noisy speech sampled at 8 or 16 kHz. Such speech is referred to as narrowband or wideband speech. When higher sampling rates are used, the speech is referred to as super-wideband or fullband speech.
7. A GRU is a variation of an RNN proposed by Cho et al. (2014). Like an LSTM model, a variant of GRU with minimal gate unit has only one forget gate, which decides how much of the previous data will be forgotten and how much will be retained for use in the next steps. However, a GRU does not contain an output gate and has many fewer parameters than an LSTM.
8. An autoencoder is often used to learn a representation (encoding) of a specific dataset. An autoencoder creates a dimension-reduced representation of the dataset. If the autoencoder is effective, the dataset can be reconstructed by the autoencoder.
9. Validation loss between the mapped target and the true target is an objective metric that is often used to evaluate the performance of a deep learning model for the validation dataset. In contrast, training loss is a metric for evaluating the performance of a deep learning model for the training dataset. The training loss is also used to adjust the model parameters.
10. All source codes are available at: https://github.com/cszheng-ioa/Sixty-years-of-frequency-domain-monaural-speech-enhancement.

References

Abdel-Hamid, O., Mohamed, Ar., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, *54*(11), 4311–4322. https://doi.org/10.1109/TSP.2006.881199

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*(3), 235–238.

Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T. (2018). Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 716–720). IEEE.

Benesty, J., & Chen, J. (2011). *Optimal time-domain noise reduction filters: A theoretical study*. Springer. https://doi.org/10.1007/978-3-642-19601-0.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). *Noise reduction in speech processing*. 2. Springer Science & Business Media.https://doi.org/10.1007/978-3-642-00296-0

Benesty, J., & Huang, Y. (2011). A single-channel noise reduction MVDR filter. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 273–276). https://doi.org/10.1109/ICASSP.2011.5946393.

Benesty, J., Makino, S., & & Chen, J. (2006). *Speech enhancement*. Springer Science & Business Media. https://doi.org/10.1007/3-540-27489-8.

Berouti, M., Schwartz, R., & Makhoul, J. (1979) Enhancement of speech corrupted by acoustic noise. In *1979 IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 4, pp. 208–211). Washington, DC, USA. https://doi.org/10.1109/TENCON.1993.320066.

Bie, X., Leglaive, S., Alameda-Pineda, X., & Girin, L. (2022). Unsupervised speech enhancement using dynamical variational autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2993–3007.

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in Amplification*, *14*(2), 113–120.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

Bradley, J. S., Sato, H., & Picard, M. (2003). On the importance of early reflections for speech in rooms. *Journal of the Acoustical Society of America*, *113*(6), 3233–3244. https://doi.org/10.1121/1.1570439

Braun, S., Kuklasiński, A., Schwartz, O., Thiergart, O., Habets, E. A., Gannot, S., Doclo, S., & Jensen, J. (2018). Evaluation and comparison of late reverberation power spectral density estimators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(6), 1056–1071.

Braun, S., & Tashev, I. (2021). A consolidated view of loss functions for supervised deep learning-based speech enhancement. In *2021 44th International Conference on Telecommunications* *and Signal Processing (TSP)* (pp. 72–76). https://doi.org/10.1109/TSP52935.2021.9522648.

Breithaupt, C., Gerkmann, T., & Martin, R. (2007). Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *IEEE Signal Processing Letters*, *14*(12), 1036–1039. https://doi.org/10.1109/LSP.2007.906208

Breithaupt, C., Gerkmann, T., & Martin, R. (2008). A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4897–4900). https://doi.org/10.1109/ICASSP.2008.4518755.

Breithaupt, C., Krawczyk, M., & Martin, R. (2008). Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4037–4040). https://doi.org/10.1109/ICASSP.2008.4518540.

Breithaupt, C., & Martin, R. (2003). MMSE estimation of magnitude-squared DFT coefficients with superGaussian priors. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 896–899). https://doi.org/10.1109/ICASSP.2003.1198926.

Byrne, D., & Dillon, H. (1986). The National Acoustic Laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing*, *7*(4), 257–265.

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 345–349. https://doi.org/10.1109/89.279283

Chang, O., Tran, D. N., & Koishida, K. (2021). Single-channel speech enhancement using learnable loss mixup. In *Interspeech* (pp. 2696–2700).

Chen, B., & Loizou, P. C. (2007). A Laplacian-based MMSE estimator for speech enhancement. *Speech Communication*, *49*(2), 134–143. https://doi.org/10.1016/j.specom.2006.12.005

Chen, J., & Wang, D. (2017). Long short-term memory for speaker generalization in supervised speech separation. *Journal of the Acoustical Society of America*, *141*(6), 4705–4714. https://doi.org/10.1121/1.4986931

Chen, J., Wang, Y., & Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1993–2002. https://doi.org/10.1109/TASLP.2014.2359159

Chen, J., Wang, Y., Yoho, S. E., Wang, D., & Healy, E. W. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *Journal of the Acoustical Society of America*, *139*(5), 2604–2612. https://doi.org/10.1121/1.4948445

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). https://doi.org/10.3115/v1/W14-4012.

Choi, H. S., Kim, J. H., Huh, J., Kim, A., Ha, J. W., & Lee, K. (2019). Phase-aware speech enhancement with deep complex U-net. In *2019 International Conference on Learning Representations* (pp. 1–20).

Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 466–475. https://doi.org/10.1109/TSA.2003.811544

Cohen, I. (2005). Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 870–881. https://doi.org/10.1109/TSA.2005.851940

Cohen, I. (2005). Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation. *Speech Communication*, *47*(3), 336–350. https://doi.org/10.1016/j.specom.2005.02.011

Cohen, I., & Berdugo, B. (2001). Speech enhancement for non-stationary noise environments. *Signal Processing*, *81*(11), 2403–2418. https://doi.org/10.1016/S0165-1684(01)00128-1

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 30–42. https://doi.org/10.1109/TASL.2011.2134090

Darwin, C. (2009). Listening to speech in the presence of other sounds. In B. C. J. Moore, L. K. Tyler, & W. D. Marslen-Wilsen (Eds.), *The Perception of Speech: From Sound to Meaning* (pp. 151–169). Oxford University Press. https://doi.org/10.1098/rstb.2007.2156.

Davis, A., Nordholm, S., & Togneri, R. (2006). Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 412–424. https://doi.org/10.1109/TSA.2005.855842

Delfarah, M., & Wang, D. (2017). Features for masking-based monaural speech separation in reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(5), 1085–1094. https://doi.org/10.1109/TASLP.2017.2687829

Dillon, H. (2012). *Hearing aids* (2nd ed.). Boomerang Press.

Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proceedings of EUROSPEECH*, *2*, 1–4.

Doire, C. S., Brookes, M., Naylor, P. A., Hicks, C. M., Betts, D., Dmour, M. A., & Jensen, S. H. (2016). Single-channel online enhancement of speech corrupted by reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 572–587. https://doi.org/10.1109/TASLP.2016.2641904

Du, Z., Zhang, X., & Han, J. (2020). A joint framework of denoising autoencoder and generative vocoder for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1493–1505. https://doi.org/10.1109/TASLP.2020.2991537

Elko, G. W., Diethorn, E., & Gänsler, T. (2003). Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation. In *Proc. International Workshop on Acoustic Echo and Noise Control* (pp. 67–70). Citeseer.

Ephraim, Y, & Malah, D (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109–1121. https://doi.org/10.1109/TASSP.1984.1164453

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator.

IEEE Transactions on Acoustics, Speech, and Signal Processing, *33*(2), 443–445. https://doi.org/10.1109/TASSP.1985.1164550

Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 708–712). https://doi.org/10.1109/ICASSP.2015.7178061.

Erkelens, J., & Heusdens, R. (2008). Tracking of nonstationary noise based on data-driven recursive noise power estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(6), 1112–1123. https://doi.org/10.1109/TASL.2008.2001108

Erkelens, J. S., Hendriks, R. C., Heusdens, R., & Jensen, J. (2007). Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(6), 1741–1752. https://doi.org/10.1109/TASL.2007.899233

Fang, H., Becker, D., Wermter, S., & Gerkmann, T. (2023). Integrating uncertainty into neural network-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 1587-1600. https://doi.org/10.1109/TASLP.2023.3265202.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and models*. 3rd edition Springer Science & Business Media, Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68888-4

Fu, S. W., Hu, T.y., Tsao, Y., & Lu, X. (2017). Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In *2017 IEEE 27th international workshop on machine learning for signal processing* (pp. 1–6). https://doi.org/10.48550/arXiv.1704.08504.

Fu, S. W., Liao, C. F., & Tsao, Y. (2020). Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality. *IEEE Signal Processing Letters*, *27*, 26–30. https://doi.org/10.1109/LSP.2019.2953810

Fu, Y., Liu, Y., Li, J., Luo, D., Lv, S., Jv, Y., & Xie, L. (2022). Uformer: A Unet based dilated complex and real dual-path conformer network for simultaneous speech enhancement and dereverberation. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7417–7421). https://doi.org/10.1109/ICASSP43922.2022.9746020.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gao, T., Du, J., Dai, L. R., & Lee, C. H. (2016). SNR-based progressive learning of deep neural network for speech enhancement. In *Proc. Interspeech 2016* (pp. 3713–3717). https://doi.org/10.21437/Interspeech.2016-224.

Gerkmann, T. (2014). Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase. *IEEE Transactions on Signal Processing*, *62*(16), 4199–4208. https://doi.org/10.1109/TSP.2014.2336615

Gerkmann, T., Breithaupt, C., & Martin, R. (2008). Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(5), 910–919. https://doi.org/10.1109/TASL.2008.921764

Gerkmann, T., & Hendriks, R. C. (2012). Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1383–1393. https://doi.org/10.1109/TASL.2011.2180896

Gerkmann, T., Krawczyk-Becker, M., & Le Roux, J. (2015). Phase processing for single-channel speech enhancement: History and recent advances. *IEEE Signal Processing Magazine*, *32*(2), 55–66. https://doi.org/10.1109/MSP.2014.2369251

Glasberg, B. R., & Moore, B. C. J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *Journal of the Acoustical Society of America*, *79*(4), 1020–1033.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236–243. https://doi.org/10.1109/ICASSP.1983.1172092

Gülzow, T., Ludwig, T., & Heute, U. (2003). Spectral-subtraction speech enhancement in multirate systems with and without non-uniform and adaptive bandwidths. *Signal Processing*, *83*(8), 1613–1631. https://doi.org/10.1016/S0165-1684(03)00080-X

Gustafsson, H., Nordholm, S., & Claesson, I. (2001). Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Transactions on Speech and Audio Processing*, *9*(8), 799–807. https://doi.org/10.1109/89.966083

Gustafsson, S., Jax, P., & Vary, P. (1998). A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)* (vol. 1. pp. 397–400 vol. 1). https://doi.org/10.1109/ICASSP.1998.674451.

Han, K., Wang, Y., & Wang, D. (2014). Learning spectral mapping for speech dereverberation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4628–4632). https://doi.org/10.1109/TASLP.2015.2416653.

Han, K., Wang, Y., Wang, D., Woods, W. S., Merks, I., & Zhang, T. (2015). Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(6), 982–992. https://doi.org/10.1109/TASLP.2015.2416653

Han, X., Huang, X., Liang, H., Ma, S., & Gong, J. (2018). Analysis of the relationships between environmental noise and urban morphology. *Environmental Pollution*, *233*, 755–763. https://doi.org/10.1016/j.envpol.2017.10.126

Hansen, J.H. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, *20*(1-2), 151–173. https://doi.org/10.1016/S0167-6393(96)00050-7

Hao, X., Su, X., Horaud, R., & Li, X. (2021). Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6633–6637). https://doi.org/10.1109/ICASSP39728.2021.9414177.

Hao, X., Su, X., Wang, Z., & Zhang, H.Batushiren (2019). UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition. In *Proc. Interspeech 2019* (pp. 1786–1790). https://doi.org/10.1109/10.21437/Interspeech.2019-1567.

Hao, X., Su, X., Wen, S., Wang, Z., Pan, Y., Bao, F., & Chen, W. (2020). Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6959–6963). https://doi.org/10.1109/10.1109/ICASSP40776.2020.9053188.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90.

He, Z., Xie, S., Ding, S., & Cichocki, A. (2007). Convolutive blind source separation in the frequency domain based on sparse representation. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1551–1563. https://doi.org/10.1109/TASL.2007.898457

Healy, E. W., Tan, K., Johnson, E. M., & Wang, D. (2021). An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners. *Journal of the Acoustical Society of America*, *149*(6), 3943–3953. https://doi.org/10.1121/10.0005089

Hendriks, R. C., Heusdens, R., & Jensen, J. (2007). An MMSE estimator for speech enhancement under a combined stochastic–deterministic speech model. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(2), 406–415. https://doi.org/10.1109/TASL.2006.881666

Hendriks, R.C., Heusdens, R., & Jensen, J. (2010). MMSE based noise PSD tracking with low complexity. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4266–4269). https://doi.org/10.1109/ICASSP.2010.5495680.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, *87*(4), 1738–1752.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE transactions on speech and audio processing*, *2*(4), 578–589.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hirsch, H. G., & Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing* (vol. 1, pp. 153–156). https://doi.org/10.1109/ICASSP.1995.479387.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hou, X., & Zhu, X. (2011). Speech enhancement using harmonic regeneration. In *2011 IEEE International Conference on Computer Science and Automation Engineering* (vol. 1, pp. 150–152. https://doi.org/10.1109/CSAE.2011.5953190.

Hu, G. (2006). *Monaural speech organization and segregation* [PhD Thesis], The Ohio State University.

Hu, Y., & Kokkinakis, K. (2014). Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. *Journal of the Acoustical Society of America*, *135*(1), EL22–EL28. https://doi.org/10.1121/1.4834455

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*.

Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio processing*, *12*(1), 59–67. https://doi.org/10.1109/TSA.2003.819949

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238. https://doi.org/10.1109/TASL.2007.911054

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2261–2269). https://doi.org/10.1109/CVPR.2017.243.

Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014). Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1562–1566). https://doi.org/10.1109/ICASSP.2014.6853860.

Huang, Y. A., & Benesty, J. (2012). A multi-frame approach to the frequency-domain single-channel noise reduction problem. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1256–1269. https://doi.org/10.1109/TASL.2011.2174226

Hussain, A., Chetouani, M., Squartini, S., Bastari, A., & Piazza, F. (2007). Nonlinear speech enhancement: An overview. *Progress in Nonlinear Speech Processing*, 217–248.

Inoue, T., Saruwatari, H., Shikano, K., & Kondo, K. (2011). Theoretical analysis of musical noise in Wiener filtering family via higher-order statistics. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5076–5079). https://doi.org/10.1109/ICASSP.2011.5947498.

Inoue, T., Saruwatari, H., Takahashi, Y., Shikano, K., & Kondo, K. (2010a). Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(6), 1770–1779. https://doi.org/10.1109/TASL.2010.2098871

Inoue, T., Takahashi, Y., Saruwatari, H., Shikano, K., & Rondo, K. (2010b). Theoretical analysis of musical noise in generalized spectral subtraction: Why should not use power/amplitude subtraction? In *2010 18th European Signal Processing Conference* (pp. 994–998).

Itakura, F., & Satio, S. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics*. IEEE, p. C-17–C-20.

Jensen, J., & Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2009–2022. https://doi.org/10.1109/TASLP.2016.2585878

Jin, W., Liu, X., Scordilis, M. S., & Han, L. (2010). Speech enhancement using harmonic emphasis and adaptive comb filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 356–368. https://doi.org/10.1109/TASL.2009.2028916

Johnson, M. T., Yuan, X., & Ren, Y. (2007). Speech signal enhancement through adaptive wavelet thresholding. *Speech Communication*, *49*(2), 123–133. https://doi.org/10.1016/j.specom.2006.12.002

Jukić, A., Waterschoot, T.v., Gerkmann, T., & Doclo, S. (2015). Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(9), 1509–1520. https://doi.org/10.1109/TASLP.2015.2438549

Kates, J. M. (2008). *Digital hearing aids*. Plural Publishing.

Kates, J. M., & Arehart, K. H. (2014). The hearing-aid speech quality index (HASQI) version 2. *Journal of the Audio Engineering Society*, *62*(3), 99–117. https://doi.org/10.17743/jaes.2014.0006

Kates, J. M., & Arehart, K. H. (2021). The hearing-aid speech perception index (HASPI) version 2. *Speech Communication*, *131*, 35–46.

Kates, J. M., & Arehart, K. H. (2022). An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids. *Hearing Research*, *426*, 108608. https://doi.org/10.1016/j.heares.2022.108608

Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(7), 1315–1329.

Kim, G., Lu, Y., Hu, Y., & Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, *126*(3), 1486–1494.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980.

Krawczyk, M., & Gerkmann, T. (2014). STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1931–1940. https://doi.org/10.1109/TASLP.2014.2354236

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Krueger, A., & Haeb-Umbach, R. (2010). Model-based feature enhancement for reverberant speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1692–1707. https://doi.org/10.1109/TASL.2010.2049684

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Le, X., Chen, H., Chen, K., & Lu, J. (2021). DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. *arXiv preprint arXiv:2107.05429*.

Lebart, K., Boucher, J. M., & Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, *87*, 359–366.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lee, D. L., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791. https://doi.org/10.1038/44565

Lee, K. Y., & Jung, S. (2000). Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, *8*(3), 282–291. https://doi.org/10.1109/89.841210

Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R. (2020). A recurrent variational autoencoder for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 371–375). IEEE.

Leglaive, S., Girin, L., & Horaud, R. (2018). A variance modeling framework based on variational autoencoders for speech enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE.

Li, A., Liu, W., Luo, X., Yu, G., Zheng, C., & Li, X. (2021a). A Simultaneous Denoising and Dereverberation Framework with Target Decoupling. In *Proc. Interspeech 2021* (pp. 2801–2805). https://doi.org/10.21437/Interspeech.2021-1137.

Li, A., Liu, W., Luo, X., Zheng, C., & Li, X. (2021b). ICASSP 2021 Deep Noise Suppression Challenge: Decoupling magnitude and phase optimization with a two-stage deep network. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6628–6632). https://doi.org/10.1109/ICASSP39728.2021.9414062.

Li, A., Liu, W., Zheng, C., Fan, C., & Li, X. (2021c). Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1829–1843. https://doi.org/10.1109/TASLP.2021.3079813

Li, A., You, S., Yu, G., Zheng, C., & Li, X. (2022a). Taylor, can you hear me now? a Taylor-unfolding framework for monaural speech enhancement. In L. D. Raedt (Ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (pp. 4193–4200). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/582. Main Track.

Li, A., Zheng, C., Peng, R., & Li, X. (2021d). On the importance of power compression and phase estimation in monaural speech dereverberation. *JASA Express Letters*, *1*(1), 14802. https://doi.org/10.1121/10.0003321

Li, A., Zheng, C., Zhang, L., & Li, X. (2022b). Glance and gaze: A collaborative learning framework for single-channel speech enhancement. *Applied Acoustics*, *187*, 108499. https://doi.org/10.1016/j.apacoust.2021.108499

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 Processing*, *22*(4), 745–777. https://doi.org/10.1109/TASLP.2014.2304637

Li, J., Luo, D., Liu, Y., Zhu, Y., Li, Z., Cui, G., Tang, W., & Chen, W. (2021e). Densely connected multi-stage model with channel wise subband feature for real-time speech enhancement. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6638–6642). https://doi.org/10.1109/ICASSP39728.2021.9413967.

Li, J., Sakamoto, S., Hongo, S., Akagi, M., & Suzuki, Y. (2008). Adaptive β-order generalized spectral subtraction for speech enhancement. *Signal Processing*, *88*(11), 2764–2776. https://doi.org/10.1016/j.sigpro.2008.06.005

Li, X., Leglaive, S., Girin, L., & Horaud, R. (2019). Audio-noise power spectral density estimation using long short-term memory. *IEEE Signal Processing Letters*, *26*(6), 918–922. https://doi.org/10.1109/LSP.2019.2911879

Lim, J., Oppenheim, A., & Braida, L. (1978). Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(4), 354–358. https://doi.org/10.1109/TASSP.1978.1163117

Lim, J. S. (1983). *Speech enhancement*. Prentice-Hall Englewood Cliffs.

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, *67*(12), 1586–1604. https://doi.org/10.1109/PROC.1979.11540

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

Liu, W., Li, A., Ke, Y., Zheng, C., & Li, X. (2021). Know Your Enemy, Know Yourself: A Unified Two-Stage Framework for Speech Enhancement. In *Proc. Interspeech 2021* (pp. 186–190). https://doi.org/10.21437/Interspeech.2021-238.

Liu, W., Li, A., Zheng, C., & Li, X. (2022). A separation and interaction framework for causal multi-channel speech enhancement. *Digital Signal Processing*, *126*, 103519. https://doi.org/10.1016/j.dsp.2022.103519

Loizou, P. (2005). Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 857–869. https://doi.org/10.1109/TSA.2005.851929

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. CRC Press. https://doi.org/10.1201/b14529

Loizou, P. C., & Kim, G (2011). Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 47–56. https://doi.org/10.1109/TASL.2010.2045180

Löllmann, H. W., & Vary, P. (2007). Uniform and warped low delay filter-banks for speech enhancement. *Speech Communication*, *49*(7-8), 574–587. https://doi.org/10.1016/j.specom.2007.04.009

Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Proc. Interspeech 2013* (pp. 436–440). https://doi.org/10.21437/Interspeech.2013-130.

Lu, Y., & Loizou, P. C. (2008). A geometric approach to spectral subtraction. *Speech Communication*, *50*(6), 453–466. https://doi.org/10.1016/j.specom.2008.01.003

Luo, X., Zheng, C., Li, A., Ke, Y., & Li, X. (2022). Analysis of trade-offs between magnitude and phase estimation in loss functions for speech denoising and dereverberation. *Speech Communication*, *145*, 71–87. https://doi.org/10.1016/j.specom.2022.10.003

Luo, Y., & Mesgarani, N. (2018). TaSNet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 696–700). https://doi.org/10.1109/ICASSP.2018.8462116.

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *27*(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Macartney, C., & Weyde, T. (2018). Improved speech enhancement with the Wave-U-Net. *arXiv:1811.11307 [cs, eess]* https://doi.org/10.48550/arXiv.1811.11307

Malah, D., Cox, R., & Accardi, A. (1999). Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2.

Martin, R. (1994). Spectral subtraction based on minimum statistics. In: *1994 European Signal Processing Conference* (pp. 1182–1185).

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, *9*(5), 504–512. https://doi.org/10.1109/89.928915

Martin, R. (2002). Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. I-253–I-256). https://doi.org/10.1109/ICASSP.2002.5743702.

Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and superGaussian priors. *IEEE transactions on speech and audio processing*, *13*(5), 845–856.

Martin, R. (2006). Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, *86*(6), 1215–1229. https://doi.org/10.1016/j.sigpro.2005.07.037

Martin, R., Heute, U., & Antweiler, C. (2008). *Advances in digital speech transmission*. John Wiley & Sons.

Masuyama, Y., Yatabe, K., Koizumi, Y., Oikawa, Y., & Harada, N. (2019). Deep Griffin-Lim iteration. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 61–65). IEEE.

Masuyama, Y., Yatabe, K., Koizumi, Y., Oikawa, Y., & Harada, N. (2021). Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network. *IEEE Journal of Selected Topics in Signal Processing*, *15*(1), 37–50. https://doi.org/10.1109/JSTSP.2020.3034486

McAulay, R., & Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(2), 137–145. https://doi.org/10.1109/TASSP.1980.1163394

McCallum, M., & Guillemin, B. (2013). Stochastic-deterministic MMSE STFT speech enhancement with general a priori information. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(7), 1445–1457. https://doi.org/10.1109/TASL.2013.2253100

Mohammadiha, N., Smaragdis, P., & Leijon, A. (2013). Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(10), 2140–2151. https://doi.org/10.1109/TASL.2013.2270369

Moore, B. C. J. (2013). *An introduction to the psychology of hearing* (6th Ed.), Brill.

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues* (2nd Ed.), Wiley.

Moore, B. C. J., Wojtczak, M., & Vickers, D. A. (1996). Effect of loudness recruitment on the perception of amplitude modulation. *Journal of the Acoustical Society of America*, *100*(1), 481–489.

Mowlaee, P., & Saeidi, R. (2013). Iterative closed-loop phase-aware single-channel speech enhancement. *IEEE Signal Processing Letters*, *20*(12), 1235–1239. https://doi.org/10.1109/LSP.2013.2286748

Nabelek, A. K., Freyaldenhoven, M. C., Tampas, J. W., Burchfield, S. B., & Muenchen, R. A. (2006). Acceptable noise level as a predictor of hearing aid use. *Journal of the American Academy of Audiology*, *17*(09), 626–639.

Naithani, G., Barker, T., Parascandolo, G., Bramslw, L., Pontoppidan, N. H., & Virtanen, T. (2017). Low latency sound source separation using convolutional recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 71–75). https://doi.org/10.1109/WASPAA.2017.8169997.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., & Juang, B. H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1717–1731. https://doi.org/10.1109/TASL.2010.2052251

Narayanan, A., & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7092–7096). https://doi.org/10.1109/ICASSP.2013.6639038.

Naylor, P. A., & Gaubitch, N. D. (2010). *Speech dereverberation*. Springer. https://doi.org/10.1007/978-1-84996-056-4

Nian, Z., Du, J., Yeung, Y. T., & Wang, R. (2022). A time domain progressive learning approach with SNR constriction for single-channel speech enhancement and recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6277–6281). IEEE.

Nicolson, A., & Paliwal, K. K. (2019). Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, *111*, 44–55. https://doi.org/10.1016/j.specom.2019.06.002

Nicolson, A., & Paliwal, K. K. (2020). Deep Xi as a front-end for robust automatic speech recognition. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering* (pp. 1–6).

Oyamada, K., Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N., & Ando, H. (2018). Generative adversarial network-based approach to signal reconstruction from magnitude spectrogram. In *2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 2514–2518). IEEE.

Paliwal, K., Wójcicki, K., & Shannon, B. (2011). The importance of phase in speech enhancement. *Speech Communication*, *53*(4), 465–494. https://doi.org/10.1016/j.specom.2010.12.003

Pandey, A., & Wang, D. (2019b). 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 (ICASSP)* (pp. 6875–6879). https://doi.org/10.1109/ICASSP.2019.8683634.

Pandey, A., & Wang, D. (2018). A new framework for supervised speech enhancement in the time domain. In *Proc. Interspeech 2018* (pp. 1136–1140). https://doi.org/10.21437/Interspeech.2018-1223.

Pandey, A., & Wang, D. (2022). Self-attending RNN for speech enhancement to improve cross-corpus generalization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1374–1385. https://doi.org/10.1109/TASLP.2022.3161143

Pandey, A., & Wang, D. (2019a). A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *27*(7), 1179–1188. https://doi.org/10.1109/TASLP.2019.2913512

Parchami, M., Zhu, W. P., Champagne, B., & Plourde, E. (2016). Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits and Systems Magazine*, *16*(3), 45–77.

Park, S. R., & Lee, J. W. (2017). A fully convolutional neural network for speech enhancement. In *Proc. Interspeech 2017* (pp. 1993–1997). https://doi.org/10.21437/Interspeech.2017-1465.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Human Language Technology* (pp. 357–362). https://doi.org/10.3115/1075527.1075614.

Peer, T., & Gerkmann, T. (2022). Phase-aware deep speech enhancement: It's all about the frame length. *JASA Express Letters*, *2*(10), 104802.

Peer, T., Welker, S., & Gerkmann, T. (2022). DiffPhase: Generative diffusion-based STFT phase retrieval. *arXiv preprint arXiv:2211.04332*.

Pinheiro, P., & Collobert, R. (2014). Recurrent convolutional neural networks for scene labeling. In E. P. Xing, & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, *Proceedings of Machine Learning Research* (vol. 32, pp. 82–90). PMLR.

Plapous, C., Marro, C., Mauuary, L., & Scalart, P. (2004). A two-step noise reduction technique. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 1, pp. I–289–92). https://doi.org/10.1109/ICASSP.2004.1325979.

Plapous, C., Marro, C., & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(6), 2098–2108. https://doi.org/10.1109/TASL.2006.872621

Plomp, R. (1964). The ear as a frequency analyzer. *Journal of the Acoustical Society of America*, *36*(9), 1628–1636. https://doi.org/10.1121/1.1919256

Popelka, G. R., Moore, B. C. J, Fay, R. R., & Popper, A. N. (2016). *Hearing aids*. Springer. :https://doi.org/10.1007/978-3-319-33036-5

Portnoff, M. (1979). Magnitude-phase relationships for short-time Fourier transforms based on Gaussian analysis windows. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 4, pp. 186–189). IEEE.

Průša, Z., Balazs, P., & Søndergaard, P. L. (2017). A noniterative method for reconstruction of phase from STFT magnitude. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(5), 1154–1164.

Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C. H. (2020). On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, *27*, 1485–1489. https://doi.org/10.1109/LSP.2020.3016837

Quatieri, T. F. (2006). *Discrete-time speech signal processing: Principles and practice*. Pearson Education India.

Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, *48*(2), 220–231. https://doi.org/10.1016/j.specom.2005.08.005

Reddy, C. K., Dubey, H., Koishida, K., Nair, A., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., & Srinivasan, S. (2021). *INTERSPEECH 2021 Deep Noise Suppression Challenge* 2796–2800. https://doi.org/10.21437/Interspeech.2021-1609.

Reddy, C. K., Gopal, V., & Cutler, R. (2021). DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6493–6497). https://doi.org/10.1109/ICASSP39728.2021.9414878.

Reddy, C. K., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., Matusevych, S., Aichner, R., Aazami, A., Braun, S., Rana, P., Srinivasan, S., & Gehrke, J. (2020). The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results. https://doi.org/10.21437/Interspeech.2020-3038.

Rethage, D., Pons, J., & Serra, X. (2018). A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5069–5073). https://doi.org/10.1109/ICASSP.2018.8462417.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 2, pp. 749–752). https://doi.org/10.1109/ICASSP.2001.941023.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N Navab, J Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4˙28.

Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M. E., & Bleeck, S. (2014). Evaluation of the sparse coding shrinkage noise reduction algorithm in normal hearing and hearing impaired listeners. *Hearing Research*, *310*, 36–47. https://doi.org/10.1016/j.heares.2014.01.006

Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M. E., & Bleeck, S. (2015). Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners. *Hearing Research*, *327*, 175–185. https://doi.org/10.1016/j.heares.2015.07.019

Schasse, A., & Martin, R. (2014). Estimation of subband speech correlations for noise reduction via MVDR processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(9), 1355–1365. https://doi.org/10.1109/TASLP.2014.2329633

Schroeder, M. R. (1965). Apparatus for suppressing noise and distortion in communication signals. US Patent 3, 180, 936.

Schroeder, M. R. (1968). Processing of communications signals to reduce effects of noise. US Patent 3, 403, 224.

Schröter, H., Escalante-B, A. N., & Rosenkranz, T., & (2022). Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7407–7411). IEEE.

Schröter, H., Rosenkranz, T., Escalante-B, A. N., & Maier, A. (2022). Low latency speech enhancement for hearing aids using deep filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2716–2728. https://doi.org/10.1109/TASLP.2022.3198548

Schuller, B., Wöllmer, M., Moosmayr, T., & Rigoll, G. (2009). Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, *2009*, 1–17. https://doi.org/10.1155/2009/942617

Schwerin, B., & Paliwal, K. (2014). Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement. *Speech Communication*, *58*, 49–68.

Sherratt, R., Townsend, D., & Guy, C. (1999). Cancellation of siren noise from two way voice communications inside emergency vehicles. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)* (vol. 4, pp. 2395–2398). https://doi.org/10.1109/ICASSP.1999.758421.

Sim, B. L., Tong, Y. C., Chang, J., & Tan, C. T. (1998). A parametric formulation of the generalized spectral subtraction method. *IEEE Transactions on Speech and Audio Processing*, *6*(4), 328–337. https://doi.org/10.1109/89.701361

Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1–3. https://doi.org/10.1109/97.736233

Soni, M. H., Shah, N., & Patil, H. A. (2018). Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5039–5043). https://doi.org/10.1109/ICASSP.2018.8462068.

Stone, M. A., & Moore, B. C. J. (1999). Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses. *Ear and Hearing*, *20*(3), 182–192. https://doi.org/10.1097/00003446-199906000-00002

Stone, M. A., Moore, B. C. J., Meisenbacher, K., & Derleth, R. P. (2008). Tolerable hearing aid delays. V. estimation of limits for open canal fittings. *Ear and Hearing*, *29*(4), 601–617. https://doi.org/10.1097/AUD.0b013e3181734ef2

Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2019). Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 239–243). https://doi.org/10.1109/WASPAA.2019.8937222.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, *9*(1), 21–29. https://doi.org/10.1109/89.890068

Suhadi, S., Last, C., & Fingscheidt, T. (2011). A data-driven approach to a priori SNR estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 186–195. https://doi.org/10.1109/TASL.2010.2045799

Sun, L., Du, J., Dai, L. R., & Lee, C. H. (2017). Multiple-target deep learning for LSTM-RNN based speech enhancement. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)* (pp. 136–140). https://doi.org/10.1109/HSCMA.2017.7895577.

Sun, M., Li, Y., Gemmeke, J. F., & Zhang, X. (2015). Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(7), 1233–1242. https://doi.org/10.1109/TASLP.2015.2427520

Taghia, J., Taghia, J., Mohammadiha, N., Sang, J., Bouse, V., & Martin, R. (2011). An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In *2011 IEEE International Conference on Acoustics, Speech* *and Signal Processing (ICASSP)* (pp. 4640–4643). https://doi.org/10.1109/ICASSP.2011.5947389.

Takahashi, N., Goswami, N., & Mitsufuji, Y. (2018). MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *2018 International Workshop on Acoustic Echo and Noise Control* (pp. 106–110). https://doi.org/10.1109/IWAENC.2018.8521383.

Takamichi, S., Saito, Y., Takamune, N., Kitamura, D., & Saruwatari, H. (2020). Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks. *Signal Processing*, *169*, 107368.

Talmon, R., Cohen, I., & Gannot, S. (2011). Transient noise reduction using nonlocal diffusion filters. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(6), 1584–1599. https://doi.org/10.1109/TASL.2010.2093651

Tammen, M., & Doclo, S. (2021). Deep multi-frame MVDR filtering for single-microphone speech enhancement. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8443–8447). https://doi.org/10.1109/ICASSP39728.2021.9413775.

Tammen, M., & Doclo, S. (2022). Deep multi-frame MVDR filtering for binaural noise reduction. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)* (pp. 1–5). https://doi.org/10.1109/IWAENC53105.2022.9914742.

Tamura, S. (1989). An analysis of a noise reduction neural network. In *1989 International Conference on Acoustics, Speech, and Signal Processing* (vol. 3, pp. 2001–2004). https://doi.org/10.1109/ICASSP.1989.266851.

Tan, K., & Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In *Proc. Interspeech 2018* (pp. 3229–3233). https://doi.org/10.21437/Interspeech.2018-1405.

Tan, K., & Wang, D. (2019). Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6865–6869). https://doi.org/10.1109/ICASSP.2019.8682834.

Tan, K., & Wang, D. (2020). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 380–390. https://doi.org/10.1109/TASLP.2019.2955276

Tan, K., & Wang, D. (2021). Towards model compression for deep learning based speech enhancement. *IEEE/ACM Transactions Audio, Speech Language Processing*, *29*(5), 1785–1794. https://doi.org/10.1109/TASLP.2021.3082282

Tan, Z. H., Dehak, N. et al. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech & Language*, *59*, 1–21. https://doi.org/10.1016/j.csl.2019.06.005

Thiemann, J., Ito, N., & Vincent, E. (2013). The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. *Journal of the Acoustical Society of America*, *133*, 3591–3591. https://doi.org/10.1121/1.4799597

Tu, Y., Tashev, I., Zarar, S., & Lee, C. H. (2018). 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). https://doi.org/10.1109/ICASSP.2018.8461944.

Udrea, R. M., Vizireanu, N., Ciochina, S., & Halunga, S. (2008). Nonlinear spectral subtraction method for colored noise reduction using multi-band bark scale. *Signal Processing*, *88*(5), 1299–1303.

Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. (2016). Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *2016 9th ISCA Speech Synthesis Workshop*. https://doi.org/10.21437/SSW.2016-24.

Valin, J. M. (2018). 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). https://doi.org/10.1109/MMSP.2018.8547084.

Valin, J. M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., & Krishnaswamy, A. (2020). A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech. In *Proc. Interspeech 2020* (pp. 2482–2486). https://doi.org/10.21437/Interspeech.2020-2730.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*, 247–251. https://doi.org/10.1016/0167-6393(93)90095-3

Vary, P. (2006). An adaptive filter-bank equalizer for speech enhancement. *Signal Processing*, *86*(6), 1206–1214. https://doi.org/10.1016/j.sigpro.2005.06.020

Veaux, C., Yamagishi, J., & King, S. (2013). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation* (pp. 1–4). https://doi.org/10.1109/ICSDA.2013.6709856.

Vincent, E., Sawada, H., Bofill, P., Makino, S., & Rosca, J. P. (2007). First stereo audio source separation evaluation campaign: Data, algorithms and results. In *2007 International Conference on Independent Component Analysis and Signal Separation* (pp. 552–559). https://doi.org/10.1007/978-3-540-74494-8˙69.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1096–1103). https://doi.org/10.1145/1390156.1390294.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, *7*(2), 126–137. https://doi.org/10.1109/89.748118

Von Neumann, T., Kinoshita, K., Drude, L., Boeddeker, C., Delcroix, M., Nakatani, T., & Haeb-Umbach, R. (2020). End-to-end training of time domain audio separation and recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7004–7008). https://doi.org/10.1109/ICASSP40776.2020.9053461.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726. https://doi.org/10.1109/TASLP.2018.2842159

Wang, D., & Lim, J. (1982). The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *30*(4), 679–681. https://doi.org/10.1109/TASSP.1982.1163920

Wang, T., Zhu, W., Gao, Y., Feng, J., & Zhang, S. (2022). HGCN: Harmonic gated compensation network for speech enhancement. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 371–375). https://doi.org/10.1109/ICASSP43922.2022.9747521.

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849–1858. https://doi.org/10.1109/TASLP.2014.2352935

Wang, Y., & Wang, D. (2012). Cocktail party processing via structured prediction. *Advances in Neural Information Processing Systems*, *25*.

Wang, Y., & Wang, D. (2013). Towards scaling up classification-based speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *21*(7), 1381–1390. https://doi.org/10.1109/TASL.2013.2250961

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

Wang, Z. Q., Wichern, G., & Roux, J. L. (2021). On the compensation between magnitude and phase in speech separation. *IEEE Signal Processing Letters*, *28*, 2018–2022. https://doi.org/10.1109/LSP.2021.3116502

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust asr. In: *Latent Variable Analysis and Signal Separation* (pp. 91–99). Cham. https://doi.org/10.1007/978-3-319-22482-4˙11.

Weninger, F., Hershey, J. R., Le Roux, J., & Schuller, B: (2014). Discriminatively trained recurrent neural networks for single-channel speech separation. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 577–581). https://doi.org/10.1109/GlobalSIP.2014.7032183.

Westhausen, N. L., & Meyer, B. T. (2020). Dual-signal transformation LSTM network for real-time noise suppression. In *Proc. Interspeech 2020* (pp. 2477–2481). https://doi.org/10.21437/Interspeech.2020-2631.

Williamson, D. S., & Wang, D. (2017). Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(7), 1492–1501. https://doi.org/10.1109/TASLP.2017.2696307

Williamson, D. S., Wang, Y., & Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(3), 483–492. https://doi.org/10.1109/TASLP.2015.2512042

Wilson, K. W., Raj, B., Smaragdis, P., & Divakaran, A. (2008). Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4029–4032). https://doi.org/10.1109/ICASSP.2008.4518538.

Wojcicki, K., Milacic, M., Stark, A., Lyons, J., & Paliwal, K. (2008). Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement. *IEEE Signal Processing Letters*, *15*, 461–464. https://doi.org/10.1109/LSP.2008.923579

Wong, L. L. N., Chen, Y., Wang, Q., & Kuehnel, V. (2018). Efficacy of a hearing aid noise reduction function. *Trends in Hearing*, *22*, 2331216518782839. https://doi.org/10.1177/2331216518782839

Xia, Y., Braun, S., Reddy, C. K. A., Dubey, H., Cutler, R., & Tashev, I. (2020). Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 871–875). https://doi.org/10.1109/ICASSP40776.2020.9054254.

Xiang, Y., & Bao, C. (2020). A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1826–1838.

Xiang, Y., Bao, C., & Yuan, J. (2020). A weekly supervised speech enhancement strategy using cycle-gan. In *2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 1–5). IEEE.

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014a). Dynamic noise aware training for speech enhancement based on deep neural networks. In *Proc. Interspeech 2014*. pp. 2670-2674. DOI: 10.21437/Interspeech.2014-571.

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014b). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, *21*(1), 65–68. https://doi.org/10.1109/LSP.2013.2291240

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7–19. https://doi.org/10.1109/TASLP.2014.2364452

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

Yin, D., Luo, C., Xiong, Z., & Zeng, W. (2020). PHASEN: A phase-and-harmonics-aware speech enhancement network. In *2020 34th AAAI Conference on Artificial Intelligence* (pp. 9458–9465). https://doi.org/10.1609/aaai.v34i05.6489.

Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., & Kellermann, W. (2012). Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 114–126. https://doi.org/10.1109/MSP.2012.2205029

You, C. H., Koh, S. N., & Rahardja, S. (2005). β-order MMSE spectral amplitude estimation for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, *13*(4), 475–486. https://doi.org/10.1109/TSA.2005.848883

Yu, C., Zezario, R. E., Wang, S. S., Sherman, J., Hsieh, Y. Y., Lu, X., Wang, H. M., & Tsao, Y. (2020). Speech enhancement based on denoising autoencoder with multi-branched encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2756–2769.

Yu, D., & Deng, L. (2011). Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine*, *28*(1), 145–154. https://doi.org/10.1109/MSP.2010.939038

Yu, G., Li, A., Wang, Y., Guo, Y., Wang, H., & Zheng, C. (2022). Joint magnitude estimation and phase recovery using cycle-in-cycle gan for non-parallel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6967–6971). IEEE.

Zavarehei, E., Vaseghi, S., & Yan, Q. (2007). Noisy speech enhancement using harmonic-noise model and codebook-based post-processing. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1194–1203. https://doi.org/10.1109/TASL.2007.894516

Zhang, K., He, S., Li, H., & Zhang, X. (2021). DBNet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement. In: *2021 Interspeech*. https://doi.org/10.21437/Interspeech.2021-1042.

Zhang, Q., Nicolson, A., Wang, M., Paliwal, K. K., & Wang, C. (2020). DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *28*, 1404–1415. https://doi.org/10.1109/TASLP.2020.2987441

Zhang, Q., Wang, M., Lu, Y., Zhang, L., & Idrees, M. (2019). A novel fast nonstationary noise tracking approach based on MMSE spectral power estimator. *Digital Signal Processing*, *88*, 41–52. https://doi.org/10.1016/j.dsp.2019.01.019

Zhang, Y., & Zhao, Y. (2013). Real and imaginary modulation spectral subtraction for speech enhancement. *Speech Communication*, *55*(4), 509–522.

Zhao, Y., Wang, D., Johnson, E. M., & Healy, E. W. (2018). A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions. *Journal of the Acoustical Society of America*, *144*(3), 1627–1637. https://doi.org/10.1121/1.5055562

Zhao, Y., Wang, D., Xu, B., & Zhang, T. (2020). Monaural speech dereverberation using temporal convolutional networks with self attention. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1598–1607. https://doi.org/10.1109/TASLP.2020.2995273

Zhao, Z., Elshamy, S., & Fingscheidt, T. (2019). A perceptual weighting filter loss for DNN training in speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 229–233). https://doi.org/10.1109/WASPAA.2019.8937189.

Zheng, C., Ke, Y., Luo, X., & Li, X. (2023). Convolutional neural network-based models for speech denoising and dereverberation: Algorithms and applications. In M. Naved, V. A. Devi, L. Gaur, & A. A. Elngar (Eds.), *IoT-enabled convolutional neural networks: Techniques and applications* (pp. 65–95). River Publishers.

Zheng, C., Liu, W., Li, A., Ke, Y., & Li, X. (2022). Low-latency monaural speech enhancement with deep filter-bank equalizer. *Journal of the Acoustial Society of America*, *151*(5), 3291–3304. https://doi.org/10.1121/10.0011396

Zheng, C., Peng, X., Zhang, Y., Srinivasan, S., & Lu, Y. (2021). Interactive speech and noise modeling for speech enhancement. *2021 Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(16), 14549–14557.

Zucatelli, G., & Coelho, R. (2021). Harmonic and non-harmonic based noisy reverberant speech enhancement in time domain. *arXiv:2112.04949 [cs, eess]* https://doi.org/10.48550/arXiv.2112.04949.